MODELING OF THE BIOLOGICAL TREATMENT PROCESS OF DOMESTIC WASTEWATER USING ARTIFICIAL NEURAL NETWORK

By

MOHAMMAD SAIFUL ISLAM

18MCE009P

A thesis submitted in partial fulfilment of the requirements for the degree of

MASTER of SCIENCE in CIVIL ENGINEERING

Department of Civil Engineering

CHITTAGONG UNIVERSITY OF ENGINEERING & TECHNOLOGY

DECEMBER 2023

Declaration

I hereby declare that the work contained in this Thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the Thesis contains no material previously published or written by another person except where due reference is cited. Furthermore, the Thesis complies with PLAGIARISM and ACADEMIC INTEGRITY regulation of CUET.

MOHAMMAD SAIFUL ISLAM
18MCE009P
Department of Civil Engineering
Chittagong University of Engineering & Technology (CUET)

Dedication

To my beloved and truly amazing **Parents**,

respected **Teachers**,

and

To my incredibly supportive **Supervisor**.

Approval/Declaration by the Supervisor

This is to certify that **MOHAMMAD SAIFUL ISLAM** has carried out this research work under my supervision, and that he has fulfilled the relevant Academic Ordinance of the Chittagong University of Engineering & Technology, so that he is qualified to submit the following Thesis in the application for the degree of **MASTER of SCIENCE in CIVIL ENGINEERING**. Furthermore, the Thesis complies with the PLAGIARISM and ACADEMIC INTEGRITY regulation of CUET.

Dr. Sudip Kumar Pal

Professor

Department of Civil Engineering

Chittagong University of Engineering & Technology (CUET)

Approval Statement

The thesis titled **Modeling of the Biological Treatment Process of Domestic Wastewater Using Artificial Neural Network** Submitted by **Mohammad Saiful Islam,** Roll No: **18MCE009P**, Session: **2018-2019** has been accepted as satisfactory in partial fulfilment of the requirement for the degree of **MASTER OF SCIENCE IN CIVIL ENGINEERING** on **09 December 2023**.

BOARD OF EXAMINATION

1.	**Dr. Sudip Kumar Pal** Professor Department of Civil Engineering Chittagong University of Engineering and Technology	Chairman (Supervisor)
2.	**Dr. Aysha Akter** Head & Professor Department of Civil Engineering Chittagong University of Engineering and Technology	Member (Ex-Officio)
3.	**Dr. Asiful Hoque** Professor Department of Civil Engineering Chittagong University of Engineering and Technology	Member
4.	**Dr. Mst. Farzana Rahman Zuthi** Professor Department of Civil Engineering Chittagong University of Engineering and Technology	Member
5.	**Dr. Md. Delwar Hossain** Professor Department of Civil Engineering Bangladesh University of Engineering and Technology	Member (External)

Acknowledgement

All praise and profound gratitude to ALMIGHTY, the Creator of the heavens and the earth, who guides everything in the best possible way. I would like to pay tribute to the remarkable spirits of Prophets and devoted believers who, across time, have sacrificed their lives to safeguard humanity, the environment, and moral principles, providing us the chance to live, learn, and flourish.

Words are not enough to express my sincere gratitude to my supervisor, **Prof. Dr. Sudip Kumar Pal**, Department of Civil Engineering, Chittagong University of Engineering & Technology (CUET), for his unwavering support and guidance throughout my studies. Without his constant encouragement and insightful discussions, this study would have been significantly more challenging. His constructive feedback, critical analysis, and meticulous review have greatly enhanced the quality of the thesis.

I thank Prof. Dr. Asiful Hoque, Prof. Dr. Aysha Akter, and Prof. Dr. Mst. Farzana Rahman Zuthi for their help and encouragement in completing this work. Additionally, I appreciate the cooperation and support provided by Md. Anwarul Islam, Assistant Technical Officer, and Md. Rabiul Hossain, Lab Attendant, during my activities at the Environmental Engineering Laboratory, CUET.

I acknowledge financial support from the Directorate of Research and Extension (DRE), CUET, through Prof. Dr. Sudip Kumar Pal's project, which facilitated the execution of a part of this study. I am sincerely thankful to Md. Mehedi Hassan Masum, Research Assistant Professor, Center for River, Harbor, and Landslide Research (CRHLSR), CUET, as well as Ratul Barua and Isfak Ahmed from Port City International University, Chattogram, for their assistance during this work.

Finally, I am endlessly grateful to my parents for their constant support, incredible encouragement, and patience during my research, particularly for cheering me up in tough moments.

ABSTRACT

Predicting effluent wastewater characteristics in advance can aid plant operators in making informed decisions to tackle environmental risks related to the treatment process and receiving water bodies. Artificial intelligence (AI)-based modeling is a promising tool for handling wastewater composition variability, and its application to the modeling of the Activated Sludge System (ASS) for treating domestic wastewater with varied compositions in due consideration for the local inputs and missing real data demands more studies. This study aims to develop a reliable AI-based artificial neural network (ANN) model that accurately reflects typical ASS parameters and to predict the effluent quality of domestic wastewater derived from that treatment system.

The study employed seven machine learning (ML) algorithms and three different ANN architectures to address various seasonal fluctuations, considering 18 parameters for the first time in this kind of study. Synthetic data and real samples collected from domestic areas in Chattogram city during dry and wet periods are used to validate the model performance. Three ANNs have shown strong predictive capabilities for synthetic wastewater data, with R^2 values over 0.9 and low RMSE (0.066–0.073) and MAE (0.051-0.059) for the 5-day Biochemical Oxygen Demand (BOD_5), Chemical Oxygen Demand (COD), and Total Suspended Solids (TSS). The standing of ANN models based on various architectures and their performance is ANN-III>ANN-I>ANN-II. Furthermore, the ANN models consistently outperformed other ML models with R^2 values exceeding 0.9 and maintaining low RMSE (0.009–0.019) and MAE (0.007–0.011) in predicting BOD_5, COD, and TSS for synthetic seasonal wastewater data. Random Forest, Gradient Boosting Regressor, and three ANNs performed well against real wastewater data, achieving high R^2 values of 0.9 and above while maintaining low RMSE (0.066–0.084) and MAE (0.052-0.067) in predicting BOD_5, COD, and TSS as model outputs. The study found that Multivariate Linear Regression and Extra Trees can even perform satisfactorily over other ML models with limited real wastewater data from the field and seasonal fluctuations.

AI-based ANN-III model has proven to be an effective predictive tool in modeling the complex process of ASS, which exhibits sound performance across various conditions, including seasonal variations, consistently achieves R^2 exceeding 0.9, and maintains low RMSE (0.009–0.084) and MAE (0.007–0.067). This study could help wastewater professionals monitor WWTPs effectively, identify issues, take remedial action, and make decisions related to wastewater treatment, quality control, process optimization, and environmental pollution control and management.

বিমূর্ত

ক্ষতিকারক দূষণের অন্যতম উৎস হিসেবে বিবেচিত গৃহস্থালী বর্জ্য পানি পরিবেশে সরাসরি নিঃসরণের কারণে মারাত্মক স্বাস্থ্য ঝুঁকি তৈরি করে। জনস্বাস্থ্য এবং পরিবেশ সুরক্ষার জন্য BOD_5, COD এবং TSS এর মতো গুরুত্বপূর্ণ দূষণকারী উপাদানের ঘনত্ব হ্রাস করা অতীব জরুরি। বর্জ্য পানির উপাদানগত বহুমাত্রিক সংমিশ্রণ এবং দ্রুত পরিবর্তনশীলতার প্যার্টান নির্ণয়ে সম্প্রতি কৃত্রিম বুদ্ধিমত্তা (AI) এর সক্ষমতা পরিলক্ষিত হচ্ছে। AI ভিত্তিক মডেলসমূহ বর্জ্য পানির বিভিন্ন উপাদানের পরিমাণ ও ঘনত্বের প্রকৃত ডেটা ব্যবহার করে সামষ্টিক আগাম ধারণা প্রদানের মাধ্যমে প্ল্যান্ট অপারেটরদের ট্রিটমেন্ট পদ্ধতিতে পরিবেশগত ঝুঁকি প্রতিরোধে এবং জল পরিবেশের সুরক্ষায় সঠিক সিদ্ধান্ত গ্রহণে সহায়তা করে। AI টুল ব্যবহার করে গৃহস্থালী বর্জ্য পানির অ্যাক্টিভেটেড স্লাজ সিস্টেম (ASS) ভিত্তিক কার্যকর ট্রিটমেন্ট সংক্রান্ত কিছু গবেষণা বিভিন্ন দেশে পাওয়া গেলেও খাদ্যাভ্যাস, পরিবেশ এবং জলবায়ুগত পরিবর্তনের ফলে স্থানীয় বর্জ্য পানির গঠনগত ভিন্নতায় বাংলাদেশে তেমন গবেষণা নেই। অতএব, ASS এর মাধ্যমে স্থানীয় গৃহস্থালী বর্জ্য পানি ট্রিটমেন্ট পরবর্তী বর্জ্যের গুণগত মান সংক্রান্ত সামষ্টিক পূর্বাভাস লাভের জন্য AI ভিত্তিক কৃত্রিম নিউরাল নেটওয়ার্ক (ANN) মডেল তৈরি পূর্বক এর কার্যকারিতা নির্ণয়ন এই গবেষণার মূল উদ্দেশ্য।

এই গবেষণায় প্রথমবারের মতো ঋতু পরিবর্তন সহ 18 টি প্যারামিটার বিবেচনা করে সাতটি মেশিন লার্নিং (ML) অ্যালগরিদম এবং তিনটি ভিন্ন ANN আর্কিটেকচার ব্যবহার করা হয়েছে। সিন্থেটিক ডেটার পাশাপাশি মডেল পারফরম্যান্স যাচাই করার জন্য চট্টগ্রাম শহরের বিভিন্ন ঘরবাড়ি থেকে শুষ্ক ও আর্দ্র উভয় সময়ই বেশ কিছু বর্জ্য পানির নমুনা সংগ্রহ করা হয়। মৌলিক পরিসংখ্যান এবং পারস্পরিক সম্পর্ক দ্বারা পরীক্ষিত ML টুল বাস্তব ডেটার অনুপস্থিতিতে স্থানীয় প্রেক্ষাপটে বর্জ্য পানি শোধনাগারের বৈশিষ্ট্য এবং ক্রিয়াকলাপের সাথে সাদৃশ্যপূর্ণ কৃত্রিম বর্জ্য পানির ডেটা তৈরি এবং বায়োলজিক্যাল ট্রিটমেন্ট ডিজাইনে তাৎপর্যপূর্ণ ফলাফল প্রদর্শন করেছে। এই গবেষণায় তিনটি ANN এর কৃত্রিম বর্জ্য পানির ডেটার সাহায্যে ০.৯ এর বেশি R^2, RMSE (০.০৬৬–০.০৭৩) এবং MAE (০.০৫১-০.০৫৯) দ্বারা BOD_5, COD এবং TSS এর দৃঢ় পূর্বাভাস প্রদর্শনের সক্ষমতা পরিলক্ষিত হয়। র‍্যান্ডম ফরেস্ট, গ্রেডিয়েন্ট বুস্টিং রিগ্রেসর এবং তিনটি ANN শুধুমাত্র BOD_5, COD এবং TSS এর প্রকৃত বর্জ্য পানির ডেটার সাহায্যে ০.৯ এর বেশি R^2, RMSE (০.০৬৬–০.০৮৪) এবং MAE (০.০৫২-০.০৬৭) প্রদর্শন করলেও প্রকৃত বর্জ্য পানির বিভিন্ন প্যারামিটারের অল্প ডেটাতে মাল্টিভেরিয়েট লিনিয়ার রিগ্রেশন এবং এক্সট্রা ট্রিস ভাল পারফর্ম করতে সক্ষম হয়েছে। ঋতু-অনুযায়ী কৃত্রিম বর্জ্য পানির ডেটা বিশ্লেষণে তিনটি ANN মডেল ধারাবাহিকভাবে ০.৯ এর বেশি R^2, নিম্ন RMSE (০.০০৯–০.০১৯) এবং MAE (০.০০৭-০.০১১) বজায় রেখে BOD_5, COD এবং TSS এর আগাম ধারণা প্রদানে তাৎপর্যপূর্ণ সক্ষমতা প্রদর্শন করে। অর্থাৎ, ডেটাসেটের সাথে সম্পর্কিত সীমাবদ্ধতা দূরীকরণে এবং গৃহস্থালী বর্জ্য পানির গুণগত মানের সামষ্টিক পূর্বাভাস প্রদানে ANN মডেলের ভূমিকা অপরিসীম।

অতএব, AI ভিত্তিক ANN মডেল ঋতু পরিবর্তন সহ বিভিন্ন পরিস্থিতিতে ধারাবাহিকভাবে ০.৯ এর বেশি R^2, নিম্ন RMSE (০.০০৯–০.০৮৪) এবং MAE (০.০০৭-০.০৬৭) সমন্বিত ফলাফলের মাধ্যমে ASS এ স্থানীয় গৃহস্থালী বর্জ্য পানি ট্রিটমেন্ট পরবর্তী বর্জ্যের গুণগত মানের আগাম তথ্য প্রদানে কার্যকার টুল হিসেবে প্রমাণিত হয়েছে। এই গবেষণালব্ধ ফলাফল বর্জ্য জল শোধনাগার কার্যকরভাবে নিরীক্ষণ, সমস্যা সনাক্তকরণ, প্রতিকারমূলক পদক্ষেপ গ্রহণ, বর্জ্য জল ট্রিটমেন্ট, গুণগত মান নিয়ন্ত্রণ, প্রক্রিয়া অপ্টিমাইজেশান, পরিবেশ দূষণ নিয়ন্ত্রণ এবং ব্যবস্থাপনা সম্পর্কিত সিদ্ধান্তে বিশদ গবেষণার জন্য নির্দেশিকা হিসাবে ব্যবহার করা যেতে পারে।

TABLE OF CONTENTS

ABSTRACT vi
বিমূর্ত vii
Table of Contents viii
List of Figures xi
List of Tables xiii
Nomenclature xv
Chapter 1: INTRODUCTION 1
1.1 Background 1
1.2 Justification of the Research 4
1.3 Aims and Objectives 6
1.4 Scope of the Study 6
1.5 Thesis Outline 7
Chapter 2: LITERATURE REVIEW 9
2.1 General 9
2.2 Brief History of Wastewater Treatment 10
2.3 Types of Wastewater Treatment System 12
2.3.1 Preliminary Treatment 12
2.3.2 Primary Treatment 14
2.3.3 Secondary Treatment 16
2.3.3.1 Attached Growth Process 17
2.3.3.2 Suspended Growth Process 20
2.3.4 Tertiary Treatment 24
2.4 Modeling History of Biological Wastewater Treatment System 25
2.4.1 Modeling History of Trickling Filter System 27
2.4.2 Modeling History of Waste Stabilization Pond System 29
2.4.3 Modeling History of Activated Sludge System 31
2.4.3.1 Fundamental Models of Activated Sludge System 33
2.4.3.2 Empirical Models of Activated Sludge System 36
2.5 Modeling of Activated Sludge System using AI Technique 37
2.5.1 Modeling of Activated Sludge System using Machine Learning 38
2.5.2 Modeling of Activated Sludge System using Deep Learning 40

2.6 Summary 46

Chapter 3: METHODOLOGY 48

3.1 General 48

3.2 Wastewater Quality Parameter 48

3.3 Wastewater Treatment Operational Parameter 50

3.4 Synthetic Data Generation 52

3.4.1 Synthetic Data Generation with AI Based Model 53

3.4.2 Processing of Synthetic Data 56

3.4.3 Pre-processing and Normalization of the Synthetic Dataset 56

3.5 Model Development 58

3.5.1 Machine Learning Models 59

3.5.1.1 Decision Tree 59

3.5.1.2 Random Forest 61

3.5.1.3 Extra Trees 62

3.5.1.4 Multivariate Linear Regression 63

3.5.1.5 K Nearest Neighbors Regressor 64

3.5.1.6 Gradient Boosting Regressor 65

3.5.1.7 Adaboost Regressor 66

3.5.2 Deep Learning Models 68

3.6 Field Sample Collection, Preservation and Testing 71

3.7 Study Area 71

3.7.1 Sample Collection 73

3.7.2 Sample Preservation and Storage 74

3.7.3 Laboratory Test 75

3.7.4 Quality Assurance and Quality Control 79

3.8 System Architecture 82

3.9 Model Calibration and Validation 84

3.10 Model Performance Visualization 86

3.11 Model Performance Evaluation Criteria 86

Chapter 4: RESULTS AND DISCUSSIONS 89

4.1 General 89

4.2 Capability of AI Based Model to Generate Wastewater Synthetic Data 89

4.2.1 Basic Statistics of Synthetic Data 94

4.2.2 Correlation of Synthetic Data 98
4.3 Category 1: Synthetic Data Associated with Mathematical Equation and Assumptions 100
4.4 Category 2: Real Wastewater Data against Generated Synthetic Data 106
4.5 Category 3: Seasonal Variation Based Performance 110
4.6 Category 4: Synthetic Data Generation from Real Wastewater Data 117
4.7 Summary of the Findings 125
Chapter 5: CONCLUSIONS AND RECOMMENDATIONS 129
5.1 General 129
5.2 Conclusions of the Study 129
5.3 Implications of this Study 130
5.4 Recommendations for Future Study 131
Bibliography 132
Appendices 143

LIST OF FIGURES

Fig. No.	Fig. Caption	Page No.
Fig. 2.1	Schematic diagram of wastewater treatment system	16
Fig. 3.1	Decision tree structure	60
Fig. 3.2	Random forest structure	61
Fig. 3.3	Extra trees structure	62
Fig. 3.4	KNN structure	64
Fig. 3.5	Gradient Boosting structure	66
Fig. 3.6	Adaboost regressor structure	67
Fig. 3.7	Artificial neural network structure	68
Fig. 3.8	Geographic locations for sampling in Chattogram	72
Fig. 3.9	Sample collection and on-spot parameter determination	73
Fig. 3.10	Wastewater sample arrangement and testing in laboratory	76
Fig. 3.11	Digestion of wastewater sample in COD reactor	78
Fig. 3.12	Flow chart of the study	83
Fig. 4.1	Violin plot of the generated dataset	90
Fig. 4.2	Box plot of the generated dataset	92
Fig. 4.3	Heatmap showing correlation among wastewater parameters	98
Fig. 4.4	(a) BOD_5, (b) COD and (c) TSS prediction of ANN-III (Category 1: Synthetic data associated with mathematical equation and assumptions)	103
Fig. 4.5	(a) BOD_5, (b) COD and (c) TSS prediction of Random Forest (Category 1: Synthetic data associated with mathematical equation and assumptions)	105
Fig. 4.6	(a) BOD_5, (b) COD and (c) TSS prediction of Random Forest (Category 2: Real wastewater data against synthetic data)	108
Fig. 4.7	(a) BOD_5, (b) COD and (c) TSS prediction of ANN-I (Category 2: Real wastewater data against synthetic data)	109
Fig. 4.8	(a) BOD_5, (b) COD and (c) TSS prediction of Multivariate Linear Regression for different scenarios (Category 3: Seasonal variation)	115

Fig. 4.9 (a) BOD_5, (b) COD and (c) TSS prediction of ANN-II for different scenarios (Category 3: Seasonal variation)..........116

Fig. 4.10 (a) BOD_5, (b) COD and (c) TSS prediction of Multivariate Linear Regression for different scenarios (Category 4: Synthetic data from seasonal real wastewater data)..........122

Fig. 4.11 (a) BOD_5, (b) COD and (c) TSS prediction of ANN-III for different scenarios (Category 4: Synthetic data from seasonal real wastewater data)..........123

LIST OF TABLES

Table No. **Table Caption** **Page No.**

Table 2.1. Summary of the outcomes of other studies using ML......................39

Table 2.2. Summary of the outcomes of other studies using DL 46

Table 3.1. Details of the variables for synthetic dataset....................................56

Table 3.2. Scaled value after normalization 58

Table 3.3. ANN variation with optimized network architecture......................70

Table 3.4. Overview on sample preservation methods 75

Table 3.5. List of analysed parameters and analytical procedures 75

Table 3.6. Approaches and technical specifications regarding the tools employed in laboratory test 80

Table 3.7. Selection of categories to address variability in ASS modeling 85

Table 4.1. Statistical description of the generated synthetic dataset 95

Table 4.2. Category 1 (Synthetic data associated with mathematical equation and assumptions): Performance evaluation of ML and DL for wastewater effluent (BOD_5, COD and TSS) 101

Table 4.3. Category 2 (Real wastewater data against synthetic data): Performance evaluation of ML and DL for wastewater effluent (BOD_5, COD and TSS) 107

Table 4.4. Category 3 (Seasonal variation): Performance evaluation of ML and DL for wastewater effluent (BOD_5) 112

Table 4.5. Category 3 (Seasonal variation): Performance evaluation of ML and DL for wastewater effluent (COD) 113

Table 4.6. Category 3 (Seasonal variation): Performance evaluation of ML and DL for wastewater effluent (TSS) 114

Table 4.7. Category 4 (Synthetic data from seasonal real wastewater data): Performance evaluation of ML and DL for wastewater effluent (BOD_5)..119

Table 4.8. Category 4 (Synthetic data from seasonal real wastewater data): Performance evaluation of ML and DL for wastewater effluent (COD) 120

Table 4.9. Category 4 (Synthetic data from seasonal real wastewater data): Performance evaluation of ML and DL for wastewater effluent (TSS)...121

Table 4.10. Performance comparison between ML and DL (all categories and scenarios) for wastewater effluent (BOD_5, COD and TSS) 126

NOMENCLATURE

Acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Network
ASS	Activated Sludge System
BOD_5	Five-Day Biochemical Oxygen Demand
COD	Chemical Oxygen Demand
COD_e	Effluent Chemical Oxygen Demand
DL	Deep Learning
F/M	Food to Microorganism Ratio
HRT	Hydraulic Retention Time
ML	Machine Learning
MLSS	Mixed Liquor Suspended Solids
MLVSS	Mixed Liquid Volatile Suspended Solids
QA	Quality Assurance
QC	Quality Control
Q_o	Design Flow Rate
Q_r	Recycle Activated Sludge Flow Rate
Q_{ratio}	Q_r/Q_o
Q_w	Waste Activated Sludge Rate
R^2	Coefficient of Determination
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
S_o	Primary Effluent BOD_5
S_e	Secondary Effluent BOD_5
SRT	Sludge Retention Time
TSS	Total Suspended Solids

V	Aeration Tank Volume
VL	Volumetric Loading
VSS	Volatile Suspended Solids
WWTP	Wastewater Treatment Plant
X	Aeration Tank MLSS
X_o	Primary Effluent TSS
X_e	Secondary Effluent TSS
X_w	Sludge SS Concentration

Chapter 1: INTRODUCTION

1.1 Background

There are many sources that generate wastewater and pollute the water. A significant contributor to water pollution is untreated domestic wastewater as hazardous pollutants such as pathogens, nutrients, organic matter, heavy metals, and other contaminants are found to present in an alarming concentrations in it (Davis & Cornwell, 2013). These pollutants may have significant impacts on the environment. Pathogens in domestic wastewater can potentially lead to a range of waterborne diseases, including cholera, typhoid, and dysentery, which can be particularly life-threatening, especially among children. Algae growth in aquatic areas can be accelerated by nutrients found in domestic wastewater. When these algae die, the process of decomposition depletes the oxygen level, causing fish and other aquatic life to perish. Organic matter and other pollutants in domestic wastewater significantly increase the demand for oxygen and thus in turn hindered availability of oxygen in natural water bodies essential for aquatic lives.

The world's population has been growing rapidly, leading to a contrary trend of insufficient access to clean water. In the absence of domestic wastewater treatment, the scarcity of freshwater due to discharge of wastewater, sewage, industrial effluents is getting worsens, and hence, wastewater treatment & its performance for future uses of treated wastewater becomes essential.

Wastewater treatment plants (WWTPs) play a vital role for ensuring public health and environmental protection. They filter pollutants out of wastewater before it is released into the environment, ensuring public health and maintaining water quality. WWTPs typically employ a combination of physical, chemical, and biological processes to reduce contaminants from wastewater following discharge guidelines, as appropriate.

Domestic wastewater treatment now has been seen an emerging issue in developing countries that were often overlooked earlier and hence, subsequent pollution is evident for open water bodies over the years to date in the absence of such treatment plants. Wastewater, when discharged directly into the environment, comprises a wide range of pollutants with significant adverse impacts (Friha et al., 2014). The major concern in WWTPs is to remove the harmful pollutants for human and aquatic lives, such as biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), total suspended solids (TSS), volatile suspended solids (VSS), organic and inorganic matters, total nitrogen (TN), total phosphorus (TP), total carbon etc. Wastewater treatment aspects depend on influents, treatment methods and ambient environments either natural or mechanical on the basis of guidelines of effluents quality. However, the reduction of priority pollutants concentrations is key towards performance evaluation. These influential parameters with significant variability depending upon the sources of origins showed a direct effect on the effluent quality. These are rather complex, so the relationships are varied from linear to nonlinear degrees. A variety of influent and effluent parameters have been seen in many types of research (Abolpour et al., 2021; Lubensky et al., 2019). Domestic wastewater treatment often depends on a natural approach such as treating using biological means, and for this natural option, there exist waste stabilization ponds, activated sludge systems, and trickling filter systems, where wastewater is treated using bacteria for reduction of BOD_5 and COD contents in raw wastewater. An activated sludge system is one of many biological systems that is frequently used in sewage treatment plants due to its flexibility and controlled operations (Frigon et al., 2013).

Following on, among many other organizations, The IWA (International Water Association) developed the Activated Sludge Model (ASM) with the aim of making it easier to design and evaluate performance of biological wastewater treatment systems, thereby encouraging and advancing practical approaches in this field (Lizarralde et al., 2015). Although activated sludge system (ASS) with different configurations has been used widely, parameter assessment and calibration require expertise and significant effort, due to the complexity and variability in wastewater compositions. In this align, the calibration of the model is based on specific treatment systems with selected parameters for simplifications. With simplifications and scope limitations, ASM models reported to found problematic and clunky (Moral et al., 2008). However, in due course, with the development of the computing system and its advancement with the inclusion of artificial intelligence (AI), a new door opens up to handle complex problems like wastewater composition, as discussed. Following on, ASS has been seen to model using numerous AI based tools in past decades (Bagheri et al., 2015; Rustum, 2009).

The performance of WWTPs and various wastewater treatment parameter values are predicted primarily using AI based machine learning (ML) and deep learning (DL) models (El-Rawy et al., 2021; Safder et al., 2022). One of the primary goals is to understand the nature of effluent composition in advance, depending on the variations anticipated. This early prediction of wastewater characteristics can reduce the wastage from the WWTP and at the same time, can help plant operators or decision-making for avoiding threats not only to the treatment units but also to receiving water bodies. In this integration, many recent articles showed a growing integration between AI and wastewater treatment processes (Alani & Khudhair, 2019; Mannina et al., 2019; Safeer et al., 2022). AI has been used successfully in wastewater treatment to combat numerous contaminants like BOD_5, COD, TN, SS, etc. These applications take into account the fluctuations and variations in wastewater composition and aim to achieve a more comprehensive approach towards their treatment goals. Artificial Neural Network (ANN) is an algorithm that effectively can classify water quality by producing accurate results in this case (Zhao et al., 2020).

The increased use of activated sludge system as of biological wastewater treatment system due to its performance over other treatment approaches, it is quite clear that developing countries, like Bangladesh adopt this system for further wastewater treatment. In addition, the significant development of AI based system, traditional monitoring and assessment of performance of the treatment plants needs to be adopted too, as it would be both cost and time effectiveness approach with appropriate calibration and validation.

While a few numbers of studies are available elsewhere, a limited or no such study is available for Bangladesh context. Moreover, the composition of wastewater has seen to vary with dietary input, environment, climate and many other issues, that are solely localized phenomenon, and hence, adaption of outcomes from the previous studies elsewhere may prove false impression on effluent quality estimation and prediction. It clearly demands more studies where absent to better understand the modeling of ASS for domestic wastewater using machine learning tools.

1.2 Justification of the Research

Pathogens, nutrients, organic debris, and heavy metals are all found in domestic wastewater. As a result, rivers, lakes, and ground water are polluted, causing diseases including cholera, typhoid, and dysentery. As a result, having a WWTP to manage wastewater pollutants to a certain degree as per requirement has become necessary. The main objective of WWTPs is to remove pollutants that are detrimental to human and aquatic life, such as BOD_5, COD, TSS, VSS, organic and inorganic materials, TN, TP, and so on. These are rather complex, so relationships are varied from linear to nonlinear in nature.

Domestic wastewater treatment frequently relies on a natural approach, such as treating sewage with bacteria culture for a reduction in BOD_5 and COD contents in raw wastewater. Among the different biological systems, an activated sludge system is commonly found in sewage treatment plants. With the development of the computing system and its advancement with the inclusion of AI in it, a new door opens up to handle complex problems like wastewater composition. Many researchers in other countries are using AI to compute the outcomes. With the advancement of the technology, it is certain that the use of AI to train models using real data help us to predict the probable future outcomes from that analysing the inherited, diverse and complex intellectual behaviour of parameters. Due to the complexity in wastewater composition, and in absence of data availability, accurate prediction of effluent quality is challenging. The pollutants present in wastewater hardly shows the linear relationships, although deterministic models assumed the relationship linear. Since, AI has the capacity to understand intrinsic variabilities in composition and their behaviour overlong the mathematical relationship, it is proposed that the prediction of effluent based on a number of input variables would be better presented with accuracy using AI based models. Thus, the goal of this study is to develop a reliable AI model capable of generating synthetic wastewater data that accurately replicates the characteristics of a typical activated sludge system. This will help in assessing whether AI-based ANN approaches are appropriate for monitoring, identifying issues, and promptly taking remedial measures for newly developed or existing WWTPs.

1.3 Aims and Objectives

The aim of this study is to model activated sludge system (ASS) to predict effluent derived from the domestic wastewater using artificial neural network. The sub objectives of this study are as follows.

- To assess the adequacy of an artificial intelligence-based machine learning tools to generate synthetic data in absence of real influent and effluent data in local context.
- To evaluate and model ASS to address variability in basic wastewater parameters using different AI based machine learning techniques.
- To reveal the best machine learning tools towards best performance of ASS modeling in terms of effluent quality.

1.4 Scope of the Study

Domestic wastewater treatment, often overlooked earlier, is now recognized as an emerging issue in developing nations. Hence, subsequent pollution is evident for open water bodies over the years without treatment plants. Wastewater encompasses extensive polluting effects once disposed directly into the surroundings.

The primary concern in wastewater treatment plants is removing harmful pollutants for human and aquatic lives. However, the reduction of priority pollutant concentrations is key towards performance evaluation. These influential parameters with significant variability depending upon the sources of origin directly affect the effluent quality. Since AI can understand intrinsic variabilities in composition and their behavior over the mathematical relationship, the prediction of effluent based on several input variables would be better presented with accuracy using AI-based models.

This study is focused on the performance of AI to predict the effluent parameters primarily BOD_5, COD and TSS derived by the activated sludge system. The scope of this research is centered specifically:

- Biological treatment system means activated sludge system.
- Synthetic data refers to the data generated by AI based machine learning techniques with or without assistance of real data and/or derived by the mathematical model.
- Real data of wastewater are obtained by collecting samples from the selected residential area in Chattogram city and testing during both dry and wet periods.
- Hydraulic Retention Time (HRT), Sludge Retention Time (SRT), Food to Microorganism Ratio (F/M), Mixed Liquor Suspended Solids (MLSS) and similar operational parameters are within the ranges commonly found in literature.
- The performance of wastewater treatment is based on BOD_5, COD and TSS by figuring out which model best performs in predicting the effluent.

1.5 Thesis Outline

The study is structured into five chapters. The first chapter, 'Introduction,' presents the thesis and outlines the research gaps.

A brief and selective overview of the relevant literature is presented in Chapter 2, ‘Literature Review’. It outlines the significance of wastewater treatment, its background, and the design of various wastewater treatment systems. Techniques for modeling activated sludge are reviewed and discussed. Finally, an overview of artificial intelligence modeling and control strategies is provided. This chapter also summarizes the previous researches in this area, knowledge gains and gaps, and then to identify the opportunities to work further.

The method used to carry out the present investigation is described in Chapter 3, 'Methodology'. It covers the procedure for collecting samples, the design of experiments, and their associated protocols, as well as the sources of the datasets used for developing different machine learning models such as Decision Tree, Random Forest, Extra Trees etc. and artificial neural networks. The criteria used to evaluate the performance of these models are also covered in this chapter.

Chapter 4, labeled as 'Results and Discussions', assesses the effectiveness of AI based models in designing the activated sludge process addressing BOD_5, COD and TSS as key output variables for WWTP performance. A detailed comparative assessment of different models comparing other studies is also done to check the adequacy, appropriateness and consistency of the present study.

The last chapter, 'Conclusions and Recommendations', aims to summarize the study's key results in conclusion and implications of this study in real field. Considering the limitations of the present study, a few recommendations have been made to enhance further research endeavors.

Chapter 2: LITERATURE REVIEW

2.1 General

The water is unique. Most life on earth comprises water, including microorganisms, plants, animals, humans, and even our brains. In addition, water is used for various tasks, including cleaning, irrigation, material transportation, industrial production, residential consumption, and irrigation. Even though the earth's surface is covered in water to a greater or lesser extent (more than 70%), only a minute fraction, approximately 0.5%, is suitable for various human activities (Gleick, 2000). This limited portion is diminishing due to rising needs in domestic, industrial, and agricultural sectors, leading to a decrease in available usable water. Wastes produced as a result of these uses are contaminants that further deteriorate the quality of the water supply, making it unusable.

Discharging wastewater containing numerous organic elements into water bodies reduces dissolved oxygen levels and poses a significant risk to various environmental issues. If this continues, the ecosystem might eventually become uninhabitable for higher life forms like fish. Additionally, industrial components may have contributed to the presence of harmful compounds (Spellman & Drinan, 2003; Tchobanoglous et al., 2003). Therefore, wastewater must be adequately treated before discharge to preserve life and safeguard the environment. Particularly, biological wastewater treatment aids in lowering the organic composition of the wastewater, reducing its effects on dissolved oxygen in the receiving water body. The following are additional advantages of biological wastewater treatment systems (Tchobanoglous et al., 2003):

- prevention of disease and troublesome conditions;

- preventing contaminating potable water sources;
- keeping water clean for bathing, recreation, and fish survival;
- ensuring the overall preservation of soil, water, and air quality for future utilization;
- reducing the concentration of ammonia, thereby decreasing its harmful impact on aquatic life like fish;
- preventing complications from excessive nitrogen compounds, such as cancer, methemoglobinemia in infants, and increased chlorine needs for disinfection

Activated sludge biological treatment systems are the most used form of biological wastewater treatment (Spellman & Drinan, 2003). Further details regarding the history of treating wastewater in general, specifically focusing on the activated sludge system (ASS), are outlined in subsequent sections. The chapter concludes by addressing the latest modeling advancements in the ASS process.

2.2 Brief History of Wastewater Treatment

Treating wastewater has a long history, going back to when early civilizations realized how important it was to control and manage sewage and other types of wastewater to prevent environmental contamination and safeguard public health. Here is a synopsis of significant historical advancements in wastewater treatment:

Ancient civilizations (3000 BCE–500 CE) originated some of the first known wastewater treatment methods. The Indus Valley Civilization in present-day Pakistan had advanced sanitation systems, including well-designed drainage systems and sewage disposal methods. Similarly, the ancient Romans constructed elaborate aqueducts, sewers, and public toilets, demonstrating an early understanding of managing wastewater.

In the case of the Medieval and Renaissance Periods (500 - 1600 CE), wastewater treatment practices were limited, and urban areas often suffered from poor sanitation. In some cases, wastewater was channeled into rivers or other bodies of water without proper treatment. However, some European cities began to develop rudimentary wastewater treatment methods, such as settling tanks and land applications.

As a result of the industrial revolution's enormous urbanization and industrialization throughout the 18^{th} and 19^{th} centuries, water bodies were increasingly polluted. As concerns about public health and the environment grew, cities began constructing more organized sewer systems to transport wastewater away from populated areas. During this period, the primary focus was transporting wastewater rather than treating it.

In the late 19^{th} and Early 20^{th} Centuries, advances in scientific understanding of disease transmission and waterborne illnesses, notably the work of researchers like John Snow and Louis Pasteur, highlighted the importance of proper wastewater treatment (Spellman & Drinan, 2003). Cities started implementing basic treatment processes such as sedimentation and aeration to reduce the pollution in wastewater before discharge.

The mid-20^{th} Century saw significant advancements in wastewater treatment technology. The activated sludge process, which involves the microbial breakdown of organic matter in wastewater, was developed and became a widely used treatment method. Additionally, the construction of secondary treatment plants began to increase.

From the late 20^{th} Century to the Present, further improvements and refinements in wastewater treatment processes continued, including developing tertiary treatment methods to remove nutrients and additional pollutants. Modern wastewater treatment plants often employ a combination of physical, chemical, and biological processes to effectively treat before releasing wastewater into water bodies or reusing it for non-potable purposes.

In recent decades, there has been an increasing emphasis on sustainable wastewater management and water reuse, focusing on sustainability and reuse (Rustum, 2009). Resources from wastewater have been recovered through efforts, including recovering cleaned water for irrigation, industrial operations, and even potable water supplies, as well as energy recovery from biogas created during treatment.

Throughout history, the need to manage and treat wastewater has been driven by concerns about public health, environmental protection, and the responsible use of water resources. Advances in technology, scientific understanding, and regulatory frameworks have contributed to developing more effective and environmentally friendly wastewater treatment methods.

2.3 Types of Wastewater Treatment System

Before being released back into the environment or reused, wastewater treatment is a technique used to eliminate impurities and pollutants from water. To achieve adequate purification, it often takes many steps. Preliminary, primary, secondary, and tertiary treatments are the four main phases of wastewater treatment. An explanation of each is given below.

2.3.1 Preliminary Treatment

Preliminary treatment is the first stage of the wastewater treatment process. Its primary goal is to remove large objects, debris, and coarse materials from the incoming wastewater that usually cause maintenance or operational problems in primary and secondary wastewater treatments and to protect downstream equipment and operations (Spellman & Drinan, 2003; Tchobanoglous et al., 2003). It is also known as pre-treatment in conventional treatment systems. Preliminary treatment typically involves the following processes:

- **Screening:** Wastewater flows through screens that have various sizes of openings. These screens trap large objects such as sticks, leaves, plastics, and other debris. The screened wastewater then continues to the subsequent treatment stage.

- **Grit Removal:** After screening, the wastewater may pass through a grit chamber where heavier, inorganic materials like sand, gravel, and small rocks settle due to their greater density. These materials are removed to prevent abrasion and damage to downstream equipment.

- **Oil and Grease Removal:** Some wastewater streams may contain oils and greases, particularly those from industrial processes or food service establishments. These can be skimmed off the surface of the wastewater in specialized tanks or clarifiers.

- **Pre-aeration (Optional):** In some cases, aeration may be provided at the preliminary treatment stage to introduce oxygen into the wastewater. This can help reduce odors, promote biological activities, and aid in breaking certain organic matter.

- **Flow Equalization (Optional):** Flow equalization involves controlling and regulating the flow rate and variability of the incoming wastewater. This helps to prevent hydraulic overload on downstream treatment processes and can enhance their efficiency.

Overall, preliminary treatment is crucial because it prevents large debris and heavy materials from interfering with the subsequent treatment processes, such as primary and secondary treatment. These solids could clog pipes, pumps, and other equipment without effective preliminary treatment, leading to operational problems and reduced treatment efficiency.

It's important to note that the specific preliminary treatment processes and equipment used can vary based on the characteristics of the incoming wastewater, the treatment facility's design, and regulatory requirements. Additionally, not all wastewater treatment plants may include every aspect mentioned above; the level of preliminary treatment can vary from plant to plant.

2.3.2 Primary Treatment

Primary treatment involves physically removing suspended solids and organic matter from the wastewater. This stage focuses on settling and separating heavier solids not removed during preliminary treatment. It includes all the units of the preliminary treatment system, as shown in Fig. 2.1, and the Primary Sedimentation Tank (PST), also known as the primary clarifier (Mara, 2004). The primary treatment process typically involves the following steps:

- **Screening:** Large particles like sticks, leaves, plastic, and other debris are removed from the wastewater by passing it through screens. This step helps prevent damage to downstream equipment and ensures smoother processing.

- **Grit Removal:** Following the initial screening process, the wastewater is directed to a grit chamber, where substances with greater weight, such as sand, gravel, and other small, dense particles, settle out.

- **Sedimentation:** In sedimentation tanks or clarifiers, the wastewater is allowed to sit undisturbed for a certain period. Due to gravity, larger solid particles and suspended materials fall to the tank's bottom during this period, forming a sludge layer. The clearer water at the top is then discharged for further treatment.

- **Floatation:** In some cases, flotation units are used to encourage the separation of oils and grease from the wastewater. Air is introduced into the wastewater, causing small oil droplets and grease particles to rise to the surface, where they can be skimmed off.

The grit chamber separates most dense suspended solids (SS) during the primary treatment phase, whereas the screen chamber efficiently removes most large floating materials. PST thus contributes significantly to the reduction of roughly 60–70% of fine settleable SS, which includes approximately 30–32% of organic SS (Mara, 2004). However, this system does not remove the colloidal and dissolved organic content of wastewater.

There needs to be more than primary treatment to meet strict environmental standards, especially in areas with high population density or industrial activities. After primary treatment, the wastewater may undergo secondary treatment processes (biological treatment) and tertiary treatment (chemical treatment and advanced filtration) to further purify the water before discharge or reuse. It's important to note that the required treatment level can vary based on local regulations, the influent wastewater quality, and the treated water's ultimate goal (e.g., discharge into a water body or reuse for non-potable purposes).

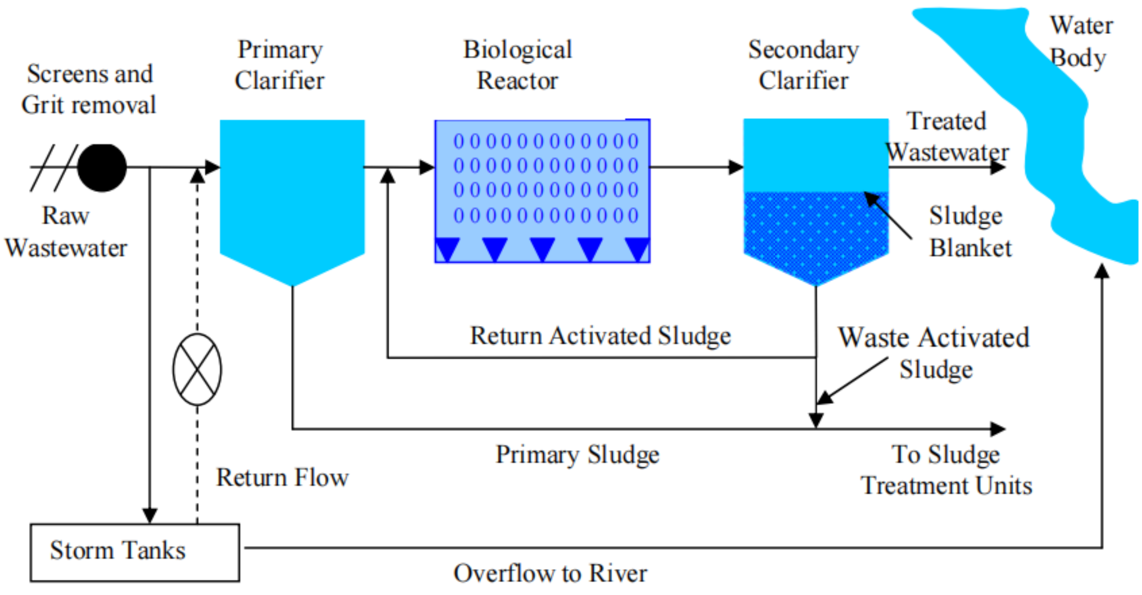

Fig. 2.1 Schematic diagram of wastewater treatment system (Rustum, 2009)

2.3.3 Secondary Treatment

Organic matter that is suspended and dissolved in the wastewater is taken out by a biological process known as secondary treatment. It involves using microorganisms to break down the remaining organic materials and convert them into stable substances like carbon dioxide, water, and more microorganisms (Davis & Cornwell, 2013).

Following primary treatment, wastewater undergoes secondary treatment to eliminate colloidal and soluble organic substances. Typically, biological processes are employed to remove the remaining colloidal and soluble organic content, as depicted in Fig. 2.1. The primary objectives of biological wastewater treatment include the coagulation and removal of both organic and inorganic non-settleable colloidal particles that persist after primary treatment. Additionally, this process aims to stabilize the dissolved organic matter, commonly measured as carbonaceous BOD_5, which remains in the primary treatment effluent. The treatment system typically involves either a suspended or attached growth (fixed film) process.

2.3.3.1 Attached Growth Process

A wastewater treatment method utilizing attached growth biological systems typically involves a reactor with a medium that facilitates biomass growth. This medium can either be inert material or biological sludge. In the case of inert media, the specific surface area and void space play crucial roles in enhancing biomass growth. Additionally, the system includes a secondary clarifier designed to separate excess suspended solids from the effluent. Based on the type of media used, the significant attached growth systems usually employed for domestic wastewater treatment include Trickling filters and Rotating biological contactors (Garg, 2014).

2.3.3.1.1 Trickling Filters

A trickling filter (TF) represents one of the earliest attached growth wastewater treatment systems. Typically, it features a circular tank containing a bed of coarse materials like large rocks, stones, ceramic pieces, or slag serving as filter media. The application of wastewater onto the support media is usually achieved through rotating distribution arms. Before the treated effluent is finally disposed of, the biomass solids that were washed away are separated in a secondary clarifier. Recirculation of effluent is commonly practiced, especially in high-rate trickling filters. The main advantages of recycling the filter effluent are as follows:

- ✓ It improves the flow distribution over the media, reducing the problem of clogging and filter flies.
- ✓ It helps in seeding the microorganisms, particularly in the case of industrial wastewater.
- ✓ It dilutes the incoming strong wastes and thereby decreases the organic loading to the filter.

- ✓ It maintains an average flow rate during low flow periods.
- ✓ It increases the contact efficiency of wastewater with media and biomass, thereby improving treatment efficiency.
- ✓ It raises the influent's dissolved oxygen content.

As the wastewater trickles through the filter media, microorganisms grow on the surface of the media or packing material in the form of a layer known as bio-film or slime layer (Karia & Christian, 2013). When wastewater passes over this stationary microbial film, contact between the substances (food) and microorganisms is established, and the substrate is decomposed or degraded aerobically by the attached biomass. Anaerobic conditions emerge close to the surface of the media, causing microorganisms to lose their attachment to the media. Consequently, the slime layer detaches and is carried away from the filter by the flow. This detachment process is referred to as sloughing.

TFs have been categorized as follows based on organic or hydraulic loadings:

- **Low rate or standard rate TF:** This type of TF is most dependable to obtain an effluent of consistent quality when the strength of influent wastewater varies. Usually, filter effluent is not recycled, and organic loading ranges from 0.08 to 0.30 BOD_5/d-m^3 of tank volume (volume of filter bed) (Karia & Christian, 2013).

- **High rate TF:** This type of TF is designed to take higher organic loadings with the recirculation of filter effluent and to prevent the filter's flooding or 'ponding'. Typically, excluding recirculation, organic loading ranges from 0.5 to 1.0 BOD_5/d-m^3 of tank volume (volume of filter bed) (Karia & Christian, 2013).

- **Intermediate rate TF:** Functionally, such filters are similar to high rate trickling filters and can be designed as single-stage or two-stage systems.

- **Super rate TF:** This type of TF normally uses plastic media with a large specific area to treat wastewater with high organic contents. The organic loading ranges from 1 to 2 kg BOD_5/m^3-d (Karia & Christian, 2013).

Additionally, TFs have been categorized as follows based on the number of units utilized in the series:

- **Single-stage trickling filter:** The system typically utilizes a single filter, and if multiple filters are necessary, they will be arranged in parallel. Recirculation of effluent flow is generally not employed.

- **Two-stage trickling filter:** This setup comprises two consecutive filters with effluent recirculation from each stage, and sometimes an intermediate clarifier is incorporated between the filters. The design of the system can involve either using equal volumes for both filters or assuming equal efficiencies for the two filters. This system is normally used for treating high strength wastewater or when removing nitrogenous organic matter is desired.

2.3.3.1.2 Rotating Biological Contactors

For domestic and industrial wastewater, a Rotating Biological Contactor (RBC) is a relatively simple and reliable biological treatment technology typically used for secondary treatment (Courtens et al., 2011). It is an attached growth technique in which the media, which typically take the shape of flat discs fixed on shafts and work similarly to trickling filters, are used. The flow of wastewater is perpendicular to the shaft, and the surface of wastewater is kept up to a level such that about 40% of the total surface area of discs is always submerged. The assembly of the discs and shaft rotates in the tank filled with wastewater. This shaft assembly with discs and equipment to rotate them is called one module. In field, many such modules may be required to treat wastewater and may be arranged in series or parallel.

The surfaces of revolving discs come into contact with the organic matter and microorganisms in wastewater and the atmosphere on alternate occasions as the shaft spins (Ivanciu et al., 2019). As the discs rotate, the microbes stick to them, transferring oxygen from the surrounding air into the wastewater to keep the system aerobic. In due course, the organic content of wastewater is consumed on the disc surface, and the microorganisms attached to the discs eventually develop into a biological film on the surfaces. An anaerobic environment eventually forms closer to the disc surface as the bio-film thickens, and the incoming wastewater flow washes off the thickened bio-film (Ivanciu et al., 2019).

2.3.3.2 Suspended Growth Process

When the microorganisms are maintained as suspension in the reactor by an appropriate mixing method, the process is known as suspended growth process, e.g., activated sludge process, waste stabilization pond etc.

2.3.3.2.1 Activated Sludge System

The most versatile biological process available to designers to treat almost all types of wastewater is the activated sludge system (ASS). This aerobic biological technique breaks down and stabilizes soluble and particulate organic matter in wastewater by using active microorganisms in a suspended condition within a reactor. The suspended biomass, called activated sludge, is quantified by Mixed Liquor Volatile Suspended Solids (MLVSS). The oxidation of organic matter and the development of new cells are the two fundamental functions of the process. A continuous oxygen supply facilitates the breakdown of organic materials and microbial growth. The mixture then goes to secondary clarifiers, where the microbial flocs gather to settle and can be added back to the aeration tank to continue the process (Garg, 2014).

The solid biomass produced in an aeration tank undergoes gravity settling in a clarifier to facilitate the process continuously. A large portion of the settled solids is typically reintroduced or recycled into the aeration tank. Any excess sludge is removed from the clarifier during the ongoing process for subsequent handling and disposal. The resulting clarified liquid is commonly discharged into a stream. However, this effluent can be reused or reclaimed. Therefore, ASS normally consists of:

- ✓ An aeration tank in which microorganisms are kept in suspension by aerating the wastewater
- ✓ A secondary clarifier in which biological flocs from the reactor are separated by gravity settling
- ✓ A recycle system to return the portion of activated sludge from the clarifier bottom to the reactor

Colloidal solids in suspension are eliminated through the physical and chemical adsorption on active biomass, as well as by enmeshment in the biological floc (Davis & Cornwell, 2013). Therefore, properly mixing wastewater with biomass in the reactor is essential. Soluble organic solids are removed by bio-sorption of matter by microorganisms and then by their biodegradation or decomposition and stabilization. During the biodegradation by oxidation of organic solids, a portion of organic matter is synthesized into new cells and another fraction is stabilized. A part of the synthesized cells will undergo self-oxidation (also known as auto-oxidation or endogenous respiration) in the reactor during the endogenous growth phase of microorganisms. Oxygen is required to support both synthesis and auto-oxidation. Normally, the oxygen needed is supplied through air by aerating the wastewater by surface aerators or diffuse aeration systems. The aeration system is so designed that it also supports proper mixing of the reactor content to generate the desired microbial floc in the reactor during the aeration (Karia & Christian, 2013; Tchobanoglous et al., 2003).

The major sub-processes involved in the removal of colloidal and soluble organic matter include:

- Dissolution of oxygen into liquid/wastewater (by aeration)
- Turbulent mixing of reactor wastewater and biomass (returned activated sludge)
- Adsorption of organic matter (substrate) by activated sludge (biomass)
- Molecular diffusion of dissolved oxygen and soluble substrate/nutrient into activated biomass (biological floc)
- Basic metabolism of microorganisms (cell synthesis)
- Bio-flocculation resulting from the production of cellular polymeric substances during the auto-oxidation phase
- Auto-oxidation of cells (endogenous respiration)
- Release of carbon dioxide (CO_2) from active cell mass
- Lysis or decomposition of dead cells

As seen in Fig. 2.1, the conventional ASS consists of an aeration tank and a clarifier (Najar & Engin, 2019). The settled effluent fed to the aeration tank is mixed with return sludge at the inlet end of the tank. Therefore, the oxygen demand by the microorganism is more in the initial length of the tank. This demand also increases with the shock load near the inlet end. So, synthesized biomass is also more near the inlet end but decreases with the length of the tank towards the outlet end. The microbial population and the system, therefore, hardly approach the relatively constant equilibrium condition similar to the complete mixed system.

The complete mix activated sludge process is so designed that effluent from the PST is mixed throughout the entire tank instantaneously. Because of complete mixing, the organic loading is considered uniform throughout the aeration tank, and the concentration of reactor biomass is not affected by the shock loadings. Therefore, oxygen demand and microbial growth are also assumed to be constant throughout the reactor.

2.3.3.2.2 Waste Stabilization Ponds

Waste stabilization ponds, commonly called oxidation ponds, represent a straightforward biological approach for wastewater treatment, especially in cases where a high-quality effluent is not essential and there is ample land for treatment. These ponds are utilized for treating both domestic and industrial wastewater that can undergo biological treatment (Garg, 2014).

Ponds are typically constructed using earthwork, featuring shallow depths about their large surface areas. Bunds are built around the ponds to prevent rainwater entry. In this setup, untreated wastewater is directly introduced to the ponds after removing floating materials via bar racks without primary treatment. Algae play a crucial role in supplying the necessary oxygen for the aerobic decomposition of organic solids and maintaining a symbiotic relationship with bacteria. This system boasts low construction costs and minimal operating expenses, as it requires minimal operational expertise and does not rely on mechanical equipment for aeration. The ponds can be designed as multi-celled structures, arranged in series or parallel configurations (Karia & Christian, 2013).

Following the screening of raw wastewater, suspended solids undergo gravitational settling at the pond bottom due to a long retention period. In the upper and intermediate layers, soluble organic matter is decomposed (oxidized) by microorganisms, mainly bacteria, in aerobic and facultative conditions, producing carbon dioxide (CO_2), nitrates, orthophosphate, and water. Oxygen essential for this process is provided by the photosynthetic metabolism of algae generated in the pond. Meanwhile, anaerobic bacteria decompose settled solids into stable end products. In practice, the facultative waste stabilization pond is mostly used to treat domestic wastewater.

When both primary and secondary treatment systems are provided, then it is generally known as complete treatment of wastewater, because the quality of final effluent of domestic wastewater obtained normally satisfies the prescribed limits set by the local authority or conforms to standards for disposal into the receiving streams.

Creating a suitable mixed culture of microorganisms in the bioreactor, maintaining ideal environmental conditions, and removing extra sludge produced are all necessary for a biological treatment unit to function well (Garg, 2014). High amounts of BOD_5 or COD in the final effluent may arise from wastewater with excess organic sludge that must be removed. This organic sludge can lower the dissolved oxygen (DO) of the receiving water body when released into a stream.

2.3.4 Tertiary Treatment

Tertiary treatment, the final step in wastewater treatment, is the process of further improving the effluent's quality before it is released into the environment or used once again. If the treated water is going to be used in irrigation, industrial activities, or even for drinking water production, this step is crucial. Different advanced methods can be used in tertiary treatment, including Filtration, Disinfection, and Nutrient Removal etc.

- **Filtration:** Wastewater is passed through sand, activated carbon, or other media to remove remaining suspended solids and fine particles.
- **Disinfection:** Pathogens like bacteria, viruses, and protozoa are destroyed or deactivated using disinfection methods such as chlorination, ultraviolet (UV) irradiation, or ozonation (Garg, 2014).
- **Nutrient Removal:** Phosphorus and nitrogen compounds are removed to prevent excessive nutrient enrichment (eutrophication) in receiving water bodies.

In summary, wastewater treatment involves a series of stages to remove contaminants and pollutants from wastewater. Each treatment stage addresses specific types of pollutants and plays a crucial role in producing clean and environmentally safe effluent. Overall, combining these treatment processes helps ensure that wastewater is treated to a level that meets environmental standards and minimizes its impact on receiving water bodies and ecosystems.

2.4 Modeling History of Biological Wastewater Treatment System

With the development of our knowledge of microbiology, chemistry, and engineering, biological treatment modeling for wastewater treatment has a long history. Here is a synopsis of how biological treatment modeling has evolved historically.

The idea of utilizing microorganisms to treat wastewater dates back to the late 19th century, when researchers observed that the action of bacteria and other microorganisms was engaged in natural processes like stream self-purification. The activated sludge process, a key biological wastewater treatment method, was developed by Edward Ardern and William Lockett in England in 1714. This process uses aeration and microbial activity to break down organic matter in wastewater (Rustum, 2009).

Researchers in the mid-20th century (1950s–1970s) began developing mathematical models to describe the kinetics of biological reactions in wastewater treatment. Researchers like C.P. Grady, G.P. Shida, and A.W. Parker made notable contributions.

The development of the ASM (Activated Sludge Model) series in the 1980s by the International Water Association (IWA) marked a significant advancement in biological treatment modeling. These models provided a framework for simulating the behavior of activated sludge systems (Rustum, 2009; Tchobanoglous et al., 2003).

In the 1990s and 2000s, modeling efforts expanded to include more complex processes and interactions, such as biological phosphorus removal and denitrification. These models allowed for a more comprehensive understanding of wastewater treatment systems.

With the advancement of robust computers and software (2000s-Present), modeling tools became more accessible and sophisticated. Computational fluid dynamics (CFD) and process modeling software are used for more accurate simulations of biological treatment processes. At the same time, biological treatment models have been integrated into control systems for wastewater treatment plants (2000s-Present). This allows for real-time monitoring and optimization of treatment processes, improving efficiency and reducing operational costs.

Integrating artificial intelligence (AI) and machine learning algorithms to optimize and forecast system performance is anticipated to be a part of current and future developments in biological treatment modeling. Additionally, there is a rising emphasis on environmentally friendly and energy-efficient treatment methods (Rustum, 2009).

Biological treatment modeling has played an important role in designing, operating, and optimizing wastewater treatment plants. As our understanding of microbiology and technology continues to advance, modeling approaches will evolve to meet environmental sustainability challenges and regulatory compliance in wastewater treatment.

In summary, the history of biological treatment modeling for wastewater treatment has evolved from simple conceptual models to complex, computer-based simulations that help design and operate wastewater treatment plants efficiently while meeting stringent environmental regulations. Ongoing research and development continue to refine and expand these modeling approaches to address new challenges in wastewater treatment.

2.4.1 Modeling History of Trickling Filter System

The design of trickling filters in wastewater treatment has evolved with the development of various mathematical and empirical models. These models have helped engineers and researchers optimize the design and operation of trickling filter systems.

Nitrifying bacteria may be kept alive under high hydraulic loadings using trickling filters, which provide a support medium for biofilm formation. The first continuous flow bioprocess used by sanitary engineers to treat wastewater was biofilms grown on support media nearly a century ago. Ammonia and dissolved oxygen permeate into the biofilm as the wastewater passes over the biological slime in a nitrifying trickling filter, where the bacteria use them for metabolism. Mass transportation and bioconversion are key operations. Modern trickling filters frequently use plastic support media, which can perform better but is significantly more expensive than filters made of rock (Siegrist & Gujer, 1987).

The key physicochemical and biological processes must be accurately described mathematically to design trustworthy trickling filters and evaluate nitrification efficacy. A reliable forecast of the anticipated bioconversion (for a filter with a specific packing material) is required regarding the critical operational variables, namely the hydraulic and nutrient loadings and the recirculation rate (where recirculation is utilized). Design engineers have had varying degrees of success using comparatively straightforward empirical design equations to simulate trickling filter performance for more than 30 years.

It is essential to address the findings of other studies because they all relate to BOD_5 elimination (Eckenfelder & Barnhart, 1963; Howell & Atkinson, 1976; Vayenas & Lyberatos, 1994). In other words, these empirical and semi-empirical equations are only applicable to the specific processes whose parameters have been established. Additionally, these equations are unable to predict the profiles of biofilm thickness and nutrient contents along the filter depth, nor can they describe how the biofilm separates from the support material. These profiles are required to determine how well a particular filter depth is utilized. More comprehensive models have been formulated in several studies (Alleman et al., 1984; Vaughan & Holder, 1984). These mathematical models are fairly sophisticated and time-consuming, and the values of the model parameters are highly unpredictable. Furthermore, the model equations lack an analytical solution, necessitating finite difference numerical techniques. A study was found to create a simple design concept for nitrifying trickling filters. A simplified empirical expression is used to enable easy integration of the overall material balance (Siegrist & Gujer, 1987).

The design of trickling filters has increasingly emphasized biofilm models from the 1980s to the present. These models contain variables including biofilm thickness, mass transfer, substrate consumption, and microbial population dynamics and consider the creation and evolution of biofilms on the filter media (Wik, 2003). Once more, analyzing flow patterns and wastewater distribution inside the filter bed is a key component of hydraulic models from the 1980s to the present. Using CFD (computational fluid dynamics) models, the hydraulic behavior of trickling filters has been predicted, and their design has been improved (Séguret et al., 2000).

Models have been incorporated into the control and optimization of trickling filter plants due to developments in automation and control systems (from the 2000s to the present). These models provide real-time monitoring and operational parameter alterations to increase treatment effectiveness (Tchobanoglous et al., 2003). With improvements in modeling methods, computational tools, and a rising focus on sustainability and environmental performance, the field is still developing.

2.4.2 Modeling History of Waste Stabilization Pond System

Waste stabilization ponds (WSPs), also known as lagoons, have been used for wastewater treatment for decades. The design of these ponds has evolved over time, and various models have been developed to aid in designing and optimizing WSP systems. Although WSPs have been a research topic for the past three decades and are very easy to develop and maintain, many processes are still not fully understood. This is due to WSPs' high level of internal complexity, in which each pond's hydraulic behavior and the various interactions between a large number of biochemical processes significantly impact how well the treatment works (Davis & Cornwell, 2013; Iturmendi et al., 2012).

Empirical guidelines and generalizations largely influenced the design of waste stabilization ponds in the early to mid-20th century. Instead of being dependent on mathematical models, these recommendations were developed using experience and field observations. In order to estimate the necessary pond volume based on the desired level of wastewater treatment, simple Hydraulic Retention Time (HRT) based models were developed at this time.

According to Moreno (1990), the hydraulic conditions of ponds have a considerable impact on the performance of waste stabilization ponds (WSP). The ponds are occasionally supposed to behave as either plug flow or fully mixed systems for design purposes. However, ponds typically have hydraulic conditions that fall between these two extremes, known as dispersed (non-ideal) flow. Dead zones and short-circuiting are other deviations from ideal conditions. Furthermore, according to Tchobanoglous et al. (2003), ponds seldom reach their theoretical retention time, primarily due to fluctuating flow rates and sludge build up that reduce a pond's active volume. Therefore, it is clear that a critical component of forecasting the performance of the pond is understanding and modeling hydraulics. Many models have been developed that solely take into account biological processes or only hydrodynamics (Iturmendi et al., 2012; Sah et al., 2012; Wood et al., 1998). Models that are solely concerned with water quality typically presumptively use plug flow or entirely mixed hydraulic conditions. Some models exclude other components and concentrate on oxygen dynamics or sedimentation processes. This is so because most of these models were created only to enhance understanding of the processes.

Several models have been developed that concentrate on various aspects of how HRAPs (high-rate algal ponds) function. For instance, some models focuse on sunlight disinfection in HRAPs. In contrast, the model by Buhr and Miller (1983) describes the crucial interaction between photosynthetic algae/cyanobacteria and heterotrophic bacteria. These models serve the same objective as WSP models, increasing the understanding of the system. Inadequate models that can accurately predict the effluent quality are also a problem for HRAPs.

In addition to hydraulics, biological processes, and environmental considerations, modeling of AFLs (aerated facultative lagoons) necessitates additional attention to mixing and DO fluctuation caused by the aerators. Many scientists have attempted to concentrate on this mixing element, including (Stropky et al., 2007). These models, like WSPs or HRAPs, are created to understand the system's operations; as a result, they are either centered on one element or the other. These models have aided in further understanding the processes in these systems.

However, no model currently can accurately anticipate how the pond will behave given various water quality criteria. A model like this could help designers during the building process and also aid them in determining how engineering modifications affect the system's performance. It is important to note that the design and modeling of WSPs can vary greatly depending on factors including climate, regional regulations, and particular treatment goals. Researchers and practitioners are developing and refining models to improve the efficiency of waste stabilization pond systems for wastewater treatment.

2.4.3 Modeling History of Activated Sludge System

Modeling the activated sludge process is vital to execute effective control actions for enhanced process performance. The fundamental reason why models are helpful is because they allow researchers to examine the impacts of changing the operating factors much faster than they could through experiments. Because of this, it is possible to explore a wide range of alternative designs and operating approaches without physically testing each scenario (Andrews, 2012; Barnett & Andrews, 1992; Orhon et al., 2005). It is feasible to promptly respond to any process alterations by employing simulations of models alongside possible corrective actions. This approach allows for the development of an operational strategy that enables the plant to adapt to new operating conditions efficiently, enhancing its stability, improving effluent quality, and concurrently reducing operational expenses. As a result, getting the best process designs at the lowest possible cost is possible while still meeting the required effluent quality criteria (Orhon et al., 2005; Rivas et al., 2008).

Modeling the wastewater treatment process is time-varying. It is made up of numerous sub processes with robust dynamics at different scales. Sludge dynamics (MLSS) and temperature are two factors that change slowly across timescales of days, weeks, and occasionally even months. The most notable change in the process comes from the variation of flow rate and pollutant concentrations. Some alterations, like dissolved oxygen (DO), occur even more quickly. Making decisions on what should be considered as inputs and outputs is challenging due to the high dimensionality. The biological cultures, recirculation, and control activities having non-linear properties all play an important role in the complex cause-and-effect relationship (Fu & Poch, 1995; Spellman & Drinan, 2003).

There are few trustworthy online measurement tools available for a biological process. For instance, quantifying certain process variables in real-time, like BOD_5, can be challenging. Furthermore, numerous sensors lack reliability due to their tendency for noise, slow responsiveness and frequent maintenance requirements (Harremoës et al., 1993; Olsson et al., 2005; Takeguchi et al., 1981). Consequently, the majority of data logs contain erroneous or incomplete values (Rustum & Adeloye, 2007a, 2007b). Most wastewater treatment facilities do not regularly monitor many aspects that affect the process. Largely unpredictable factors including the impact of poisonous industrial products and mechanical malfunctions are difficult to model mathematically

(Schuetze et al., 2002). Each wastewater treatment facility is unique. The amount and circumstances vary, and the industrial waste inputs nature changes depending on the location. Therefore, it is important to take into account the unique characteristics and conditions of any wastewater treatment facility that will be modeled mathematically. ASS, in particular, have shown advancements in modeling wastewater treatment processes, notwithstanding everything mentioned above. In the past 40 years, there has been a lot of interest in the mathematical modeling of the processes taking place inside the ASS (Schuetze et al., 2002).

There are two main aspects to consider when classifying activated sludge models: firstly, fundamental modeling involves identifying the specific area (primary clarifier, aeration reactor, secondary clarifier etc.) of the plant to model, and secondly, empirical modeling involves determining the modeling approach to be employed. Fundamental models are those that are derived using mechanistic equations. Empirical models, in contrast, are data-driven but do not consider the system's physical considerations. Statistical techniques are used to fit the model coefficients to the input-output map to identify empirical models from system input-output data. The modeling attempts made to simulate the activated sludge wastewater treatment facility using various modeling methodologies are described in the following subsections.

2.4.3.1 Fundamental Models of Activated Sludge System

Considerable attention has been given to developing models for primary clarifiers (PST) due to their significant impact on subsequent units and sludge treatment (Lessard & Beck, 1993). Simpler models, like the steady-state approach, are often used to depict the dynamic behavior of primary clarification because they are deemed less sensitive (Otterpohl & Freund, 1992). The majority of primary clarifier models solely simulate the behavior of suspended solids (SS) and do not consider any biological interactions within the reactor. A few studies consider certain biological events occasionally occurring in the primary clarifier (Lessard & Beck, 1988).

Modeling primary clarifiers accurately faces challenges due to the intricate dynamics of sedimentation processes. These challenges arise from factors such as fluctuating influent characteristics, diverse particle sizes and velocities, intricate flow patterns with density currents within the tank, scouring phenomena, and the impact of temperature. However, the majority of proposed models rely on variables that are not assessed during actual treatment work operations (Lessard & Beck, 1993; Otterpohl & Freund, 1992; Paraskevas et al., 1999).

Deterministic mathematical models use kinetic parameters, coefficients, and differential equations to explain the process. Changes in influent wastewater flow rate, composition, and concentration are all taken into consideration by these models. As a prominent deterministic model, the Activated Sludge Model Number 1 (ASM1) encompasses various variables and processes such as different organic matter fractions, biomass, nitrogen components, particulates, and alkalinity (Henze et al., 1987). The transformation of Portugal's Parada conventional WWTP into a biologically nitrogen-removing facility using ASM1 was investigated. Using the parameter settings suggested by the developers of ASM1, very high correlations between observed data and simulation results were attained. They assessed the wastewater treatment plant's capability for biological nitrogen removal through computer simulations. Their findings suggested that the plant could generate excellent effluent quality with the first tank dedicated to denitrification and the second and third tanks allocated for BOD_5 removal and nitrification, along with a recirculation flow ratio of 1.5 and a sludge age of 15 days.

Further development in modeling activated sludge WWTPs resulted in the creation of ASM2, ASM2D, and ASM3 models. ASM2 provided detailed biological kinetics and understanding of nitrification, denitrification, and biological phosphorus removal. ASM2D extended the ASM1 and ASM2 to include a model for biological phosphorus removal alongside simultaneous nitrification-denitrification in ASS. ASM3 is capable of predicting oxygen consumption, sludge production, as well as nitrification and denitrification processes within ASS (Gujer et al., 1995, 1999).

However, the complexity of models like ASM1 and others necessitates the assessment of numerous variables and substantial data on microorganism growth and decay. These factors are often not routinely evaluated by significant WWTPs, thus diminishing their priority to designers and operators (Lessard & Beck, 1993; Weijers & Vanrolleghem, 1997). A model of a WWTP system must also include a certain amount of assumptions and simplifications in which some relate to the mathematical model, and the remaining are associated with the physical system itself to be realistically usable (Jeppsson, 2012). When used in full-scale wastewater treatment plants, the majority of these models have an incorrect calculation of the return sludge concentration and inaccurate analysis of the sludge concentration profile near the effluent weirs. Since these models depend on particular site parameters, applying them to another site would necessitate significant calibration work (Lessard & Beck, 1993).

Deterministic models are favored for their ability to forecast beyond the scope of the available operating data but with limitations. These models are typically developed using controlled laboratory data, making them suitable for treatment plant design but potentially unsuitable for day-to-day operations (Lessard & Beck, 1993; Rustum & Adeloye, 2007a). Mechanistic models require calibration and might be challenging to adjust for any changes in the physical system because these models were created based on specific physical systems.

They rely on numerous parameters and coefficients, limiting their accuracy and applicability. Furthermore, many of these variables and indicators are often excluded from regular plant performance monitoring. Consequently, empirical models are increasingly preferred for modeling activated sludge WWTPs.

2.4.3.2 Empirical Models of Activated Sludge System

WWTPs require advanced control systems capable of achieving improved and adaptable performance due to their complex nature. These systems need to exhibit strong dynamical performance and resilience, effectively managing intricate, uncertain, and highly non-linear process connections. However, due to limitations in mechanistic models, researchers are exploring new techniques to handle these challenges. One such technique is the use of intelligent models like neural networks, fuzzy logic, data mining, etc. that handle uncertainty and system complexity in a manner analogous to human reasoning without the challenges associated with deterministic non-linear mathematics.

Stochastic modeling is a methodology that studies sequential observations generated over time. This approach has been implemented to predict the efficiency of treatment processes, demonstrating successful outcomes in various investigations. In the study conducted by Capodaglio (1994), both univariate and multivariate process models were employed to predict water flow and SS levels, based on rainfall measurements. An study incorporated stochastic models into a prototype real-time control setup for forecasting flow and employing these predictions for automated online regulation of an ASS in Denmark (Kristensen et al., 2004). The input data was precipitation, measured at the WWTP, while flow data from the final pumping station preceding the treatment facility served as the output. Expert systems (ES) models, which have been used for controlling WWTPs since the 1980s, are implemented using expert knowledge and databases. However, they may not apply to every system due to the challenge of collecting expert knowledge. On the other hand, fuzzy logic models make a balance between uncertain human expressions and rigid expert systems as they do not require complex mathematical relationships and are conceptually easy to understand. Additionally, they can handle imprecise data, enabling the modeling of intricate non-linear functions. However, adjusting the parameters of fuzzy membership functions is difficult and time-intensive, and determining the required quantity of fuzzy rules poses a significant challenge.

2.5 Modeling of Activated Sludge System using AI Technique

Wastewater treatment plants (WWTPs) are complex, nonlinear, and highly changeable systems characterized by flow rate, pollutant load, chemical composition, and hydraulic conditions. These complexities and uncertainties make it more challenging to simulate WWTP processes (Borzooei et al., 2019; Rout et al., 2021). Mechanistic models that simulate WWTP processes have been used to forecast the behavior of several variables (Buaisha et al., 2020; Fenu et al., 2010; Nopens et al., 2009). Mechanistic models, however, have several shortcomings because they require a great deal of simplicities and assumptions in order to be tractable and quantifiable as well.

Additionally, combining mechanistic models that mimic processes in many units proves challenging due to variances in the methods employed to create state variables. Poor generalization capabilities, high costs, and insufficient handling of the time-varying and highly nonlinear behaviors of processes affected by various known and unknown factors are additional issues that mechanistic models have (Liu et al., 2021; Shi et al., 2018). Many of these limitations are eliminated by artificial intelligence-based models, which do not rely on pre-designed processes based on fundamental principles but instead only on finding correlations between output and input data that enable predictions and/or facilitate decisions (Müller & Guido, 2016).

AI (Artificial Intelligence) is a powerful tool in wastewater treatment plants (WWTPs) that enhances efficiency, reduces costs, and improves system performance. It allows real-time monitoring of parameters like water quality, flow rates, and equipment performance and can process vast amounts of data to optimize treatment processes. AI-powered predictive maintenance helps identify equipment failures and maintenance needs before they occur, reducing downtime and extending equipment lifespan. WWTPs are energy-intensive facilities, and AI algorithms can optimize energy usage by adjusting equipment operation based on real-time demand and conditions. AI can analyze complex data patterns and recommend process adjustments to optimize treatment efficiency, leading to improved pollutant removal rates and reduced chemical usage. AI can detect abnormal conditions and potential system failures early, preventing pollution events and ensuring regulatory compliance. It can also analyze historical data to identify trends and patterns for making better decisions and improving treatment processes continuously. AI enables remote monitoring and control, reducing the need for on-site personnel and improving operational flexibility. It can adapt real-time treatment processes to handle changing influent characteristics due to weather, industrial discharges, or population growth.

2.5.1 Modeling of Activated Sludge System using Machine Learning

It is challenging to simulate WWTP processes because of thecomplexities and uncertainties associated with WWTPs (Borzooei et al., 2019; Rout et al., 2021). The behavior of certain variables has been predicted using mechanistic models, such as Activated Sludge Models (ASMs), which have been used extensively to simulate WWTP processes (Buaisha et al., 2020; Fenu et al., 2010; Nopens et al., 2009). Mechanistic models have several drawbacks because they require many assumptions and simplifications to be tractable and computable. ASMs, for instance, are only appropriate subject to a range of temperatures, pH levels, and alkalinities (Gujer et al., 1995, 1999; Hauduc et al., 2011; Henze et al., 1987).

Additionally, linking different mechanistic models that replicate processes in multiple units is challenging because of differences in the techniques employed to calculate state variables. For instance, TSS is calculated and incorporated differently in ASMs and second clarifier models (Tchobanoglous et al., 2003).

Table 2.1. Summary of the outcomes of other studies using ML

Input	Output	Methods	Model Performance	Reference
pH, temperature, TP, flow, Cond., Turbidity	COD	Random Forest	$R^2 = 0.91$	(Cheng et al., 2023)
F/M, SVI, T, Cond., pH, COD, BOD_5, N, P, Q, Q_r, Q_w	Eff. COD, N, P	Ensemble Parallel Hybrid (bagging, 3x MLP)	Training RMSE COD= 224.38 N= 0.36 P= 0.10 Testing RMSE COD= 127.71 N= 0.23 P= 0.15	(Keskitalo & Leiviskä, 2015)
BOD_5, TSS, VSS, influent flow rate, pH, temperature, F/M, SRT, WAS, and RAS	BOD_5 and TSS	RVFL-MRFO	BOD_5 $R^2 = 0.97$ RMSE = 1.147 and TSS $R^2 = 0.97$ RMSE = 1.1927	(Elmaadawy et al., 2021)
TSS, TDS, COD, BOD_5	TSS, TDS, COD, BOD_5	Support Vector Regression (SVR) and Regression Trees (RT)	(RMSE) (mg/L) TSS= 1049, TDS=1549, COD= 1172, BOD_5= 104 (R^2) TSS = 0.97, TDS= 0.851, COD= 0.893, BOD_5= 0.871	(Granata et al., 2017)

Other inherent problems with mechanistic models include poor generalization performance, high costs, and insufficient ability to effectively manage the dynamic and complex behaviors of processes influenced by both known and unknown factors (Liu et al., 2021; Shi et al., 2021; Verma et al., 2017).

Machine Learning (ML) models are free from many of these restrictions, because they only depend on identifying correlations between output and input data that allow predictions and/or facilitate decisions (Müller & Guido, 2016; Nadiri et al., 2018). A review of the findings from earlier studies that used machine learning in the context of activated sludge systems is shown in Table 2.1.

The fact that ML models simulate actual reaction/process conditions rather than pre-designed processes based on fundamental principles is a key benefit. They are so strong and comprehensive, which is important because many mechanisms involved in wastewater treatment are still not fully understood (Erdirencelebi & Yalpir, 2011; Lee et al., 2002; Nadiri et al., 2018; Wang et al., 2021). To replace mechanistic modeling of WWTPs, ML modeling is therefore frequently utilized (Liu et al., 2021; Shi et al., 2021; Verma et al., 2017).

2.5.2 Modeling of Activated Sludge System using Deep Learning

It is possible to use ANNs to forecast how well WWTPs will perform. ANNs were developed to replicate the water treatment process because of their excellent prediction accuracy (Alver & Kazan, 2020; Manu & Thalla, 2017; Newhart et al., 2019). The reliability of the historical data is essential for ANN modeling. The performance of ANN models may be compromised by poor historical data quality. However, ANN modeling only needs a little amount of data to produce reliable prediction outcomes.

The prediction of odors from wastewater treatment plants (WWTP) was conducted by Kang et al. (2020) using artificial neural networks (ANN). They explored both conventional and alternative process structures, successfully predicting odor characteristics based on functional groups using ANNs. Compared with other methods, the ANN-based model demonstrated better accuracy and error rates in odor characteristic prediction. The study's findings highlight the potential of ANN in odor prediction. This research contributes to advancing gas sensing technologies, particularly in environmental monitoring and quality control applications.

Ten months performance of Touggourt WWTP was predicted in terms of effluent COD using a Feed Forward Neural Network (FFNN) for different architectures, learning algorithms, and activation functions (Bekkari & Zeddouri, 2018). The best learning algorithms among Levenberg–Marquardt (LM), Quasi–Newton back-propagation (BFG), Resilient back-propagation (RP), and Conjugate gradient back-propagation (CGF), LM depicted lower errors and higher correlation coefficients. They obtained the best result for Tansig as the hidden layer activation function (HLAF) and Purelin as the output layer activation function (OLAF) as model architecture. The best architecture for the ANN-FFNN was (6-50-1). The highest Correlation Coefficient (R-value) was obtained in terms of validation.

Data from the raw wastewater treatment plant (WWTP) were classified into four categories using a clustering analysis method in a study (Zhao et al., 2016). The study aimed to predict key parameters like TP, BOD_5, COD, SS, and NH_3-N in the treated effluent using various input parameters. The researchers used a Back Propagation Neural Network (BPNN) as an artificial neural network (ANN) component for training to address seasonal fluctuations in wastewater composition. Through experimentation, they determined that the optimal configuration for the ANN involved 16 neurons. This comprehensive approach offers valuable insights into water quality prediction and highlights the significance of considering seasonal variability in wastewater treatment processes.

Sharghi et al. (2019) proposed a comprehensive approach to predict effluent BOD_5 using various clustering techniques. The study described three distinct clustering methods, with the first model employing an unsupervised technique called Self-Organizing Map (SOM). Through their investigation, the researchers compared the performance of three ANN models, and interestingly, the unsupervised SOM technique yielded the most promising results. Using clustering techniques to identify influential input variables showcases the potential for enhancing the accuracy and effectiveness of BOD_5 prediction in wastewater treatment plants. This research contributes valuable insights into the field of effluent quality monitoring and highlights the significance of combining clustering methods with artificial neural networks for improved wastewater management.

Three different approaches were used in a study to predict COD. The first approach utilized Principal Components Analysis (PCA), the second employed Mutual Information (MI), and the third approach involved six artificial neural network (ANN) models using various influent input parameters, including COD, BOD_5, pH, conductivity, TN, TO, and TSS (Elkiran & Abba, 2017). This study sheds light on the effectiveness of ANN modeling for COD prediction, contributing valuable insights to the field of wastewater treatment plant effluent forecasting. The research aims to assess the efficiency of Artificial Neural Networks (ANNs) in predicting COD for the Nicosia wastewater treatment plant. Sensitivity analysis is used as a feature selection method. The best-performing model, ANN I, achieved an R-square value of 0.7034 and an RMSE of 0.0108 with a structure of 8-8-1 and 203 epochs.

Furthermore, it has been found that the model performance is better with an increased number of input parameters (Hassen & Asmare, 2019). They have built five networks consisting of three SISO (Single input, single output), two MIMO (Multiple input, multiple output), and one MISO (Multiple input, single output). In order to get a predicted pH value, the best-obtained architecture was 1-41-1 for the input parameter, COD, where the MSE value of the training data was 0.054, testing was 0.057, and the R-values were 0.805 for training and 0.747

for testing. The model also worked for SISO to find the value of TN with the architecture (1-68-1), and the obtained MSE result for training data was 0.024 and testing data was 0.131, and the R-value for SISO was 0.937 for training and 0.779 for testing. Overall, for COD, SISO worked fine, with the highest R-value and lowest MSE value.

Keskitalo and Leiviskä (2015) propose a parallel hybrid modeling approach to enhance the Activated Sludge Model (ASM) prediction for a pulp water treatment plant. The study employs cross-validation ensemble methods and bagging, with the latter proving the most effective among the applied methods. The research showcases the potential of ANN ensembles in improving modeling accuracy for wastewater treatment processes and offers valuable insights for optimizing treatment plant performance.

A framework for plant operation monitoring is provided by the use of neural network models simulating wastewater treatment plants (Mjalli et al., 2007). The reduction of operating expenses and assessment of the environmental stability are the results of this monitoring system. They collected data throughout a year by taking measurements every five days. It was determined that the inputs to their model were the secondary treatment effluent (STE) outputs. The entire collection of data was divided in the proportions of 4:1:2 as training, validation, and testing data sets, respectively. It should be mentioned that Mjalli et al. (2007) made use of a MATLAB-created Feedforward Neural Network (FFNN). In all single-input networks, the first three levels of the network architecture are composed of the input, hidden, and output layers. The three layers (input, hidden, and output) comprise all single-input networks' network architecture. The network arrangement of multi-input networks consists of four layers (two hidden layers). With the COD as the input variable, the 1-40-1 network structure was the best ($R^2 = 0.987$, MSE = 0.021). The MSE between the network prediction and actual values served as the basis for the training output function. Nasr et al.

(2012) used an ANN to predict how well Alexandria's wastewater treatment facilities would function. Over the course of a year, the plant's performance was examined in terms of BOD_5, COD and TSS. Four groups were created from the collected data, each representing three months of the year. Backpropagation was used to create and train the feedforward network. The predicted and measured values correlate strongly ($R^2 = 0.903$).

Machine learning tool was used to forecast the total nitrogen concentration in the effluent of a WWTP in Ulsan, Korea, using a feedforward ANN and support vector machine (SVM) (Guo et al., 2015). Total nitrogen, phosphorus, and TSS from the influent flow were the models' inputs. The results showed that both models were successful at producing predictions according to the coefficient of determination (R^2), Nash-Sutcliffe efficiency (NSE), and relative efficiency (drel) criteria ($R^2 = 0.55$, NSE = 0.56, and drel = 0.8). The ANN model with three layers—input, hidden, and output—was more successful at correlating the input values to the T-N concentrations, even though the SVM model had a greater prediction efficiency.

In order to estimate the effluent quality index (EQI) for wastewater treatment plants in Tehran, Nezhad et al. (2015) used the MATLAB ANN toolkit. Three hidden layers of a feedforward back-propagated neural network were created. The model took BOD_5, TDS, TSS, FC, PO_4, NH_4, and pH as inputs. The findings showed that the 8-7-1 network structure was the most effective one for EQI prediction (R = 0.96 and MSE = 0.1). Additionally, Elkiran and Abba (2017) used a feed-forward neural network (FFNN) to estimate the COD of the effluent from wastewater treatment plants. An important measure for evaluating the effectiveness of WWTPs is the COD of the effluent. In order to forecast the effluent's COD, BOD_5, pH, T-P, T-N, TSS, and conductivity at WWTP inlets, an FFNN was used. Though many structures and input combinations were taken into account, the FFNN model with an 8-8-1 network structure and all four input parameters produced the best results and the highest accuracy in effluent COD prediction (R^2 = 0.7 and RMSE = 0.0108).

A foundation for plant operation monitoring is provided by the simulation of wastewater facilities using neural network models (Mjalli et al., 2007). This monitoring system makes it possible to reduce operating expenses and assess the level of environmental stability. Furthermore, earlier publications did not disclose the process utilized to choose the ANN configuration. As a result, this study offers a straightforward method for determining the ideal ANN design for predicting WWTP plant performance. The five-day biological oxygen demand BOD_5 can also be predicted using a number of regression models and artificial neural networks (Baki et al., 2019a; Kasem et al., 2018; Najafzadeh & Ghaemi, 2019).

The lengthy laboratory processes needed to measure the BOD_5 concentration, which take around five days, highlight the significance of BOD_5 modeling. The relationship between the parameters of wastewater and BOD_5 served as the basis for developing a regression model, which demonstrated great accuracy (R^2 up to 0.797) (Baki et al., 2019a). Another study was found to create a software sensor employing a feedforward ANN and dissolved oxygen level to monitor the BOD_5 concentration in the Sefid-rud River in Iran (Kasem et al., 2018). The performance of the created ANN was satisfactory, and a high R^2 value (up to 0.89) was produced. A BOD_5 soft sensor was created in WWTPs using deep neural networks and genetic algorithms (Qiu et al., 2016). On the BSM1 simulation platform, the developed sensor was evaluated in three weather scenarios, including dry, wet, and stormy circumstances. The results showed that the sensor performed well in these scenarios. Table 2.2 summarizes the findings from past research that applied deep learning to ASS.

Table 2.2. Summary of the outcomes of other studies using DL

Input	Output	Methods	Model Performance	Reference
Influent pH, T, SS, KN, influent BOD_5, influent COD.	Effluent COD	ANN	$R^2 = 0.9675$ MSE = 0.0020	(Bekkari & Zeddouri, 2018)
TP, BOD_5, COD, SS, NH_3-N, pH, Electricity Consumption, Coagulant, and Flocculant	TP, BOD_5, COD, SS, NH_3-N	Back Propagation and Adaptive learning rate	Annual Training Error = 0.0513 Testing Error= 0.051705 Seasonal Training Error = 0.066 Testing Error= 0.0712	(Zhao et al., 2016)
MLSS, EC, TDS, COD	Effluent BOD_5	ANN using SOM	$R^2 = 0.74$ RMSE = 0.046	(Sharghi et al., 2019)
COD, BOD_5, pH, Conductivity, TN, TP, TSS, SS	Effluent COD	ANN	$R^2 = 0.7034$ RMSE = 0.0108	(Elkiran & Abba, 2017)

2.6 Summary

The history of activated sludge WWTP, the significance of wastewater treatment, and the design of ASS are all covered in this chapter. The latest advancements in modeling activated sludge facilities for treating wastewater are also addressed. All AI-based models use real data from existing wastewater treatment plants, and their outputs are used to modify decisions regarding management for such WWTPs. Due to the increasing popularity of activated sludge systems as biological wastewater treatment systems and their superior performance to other treatment methods, developing countries like Bangladesh will undoubtedly continue to adopt these systems for wastewater treatment. While a few studies are available elsewhere, Bangladesh has very few or no studies of this kind. Moreover, every investigation examines or considers limited parameters associated with the quality and functioning of activated sludge systems.

Additionally, the composition of wastewater has been observed to fluctuate with food habits, environment, climate, and many other factors that are only local context-based. As a result, adaptation of results from prior studies elsewhere may prove to be a false impression on effluent quality evaluation and prediction. Again, developing and monitoring activated sludge systems based on AI, particularly ANN due to its capability of capturing intrinsic variabilities, is a significant problem in the absence of WWTP and real data of various parameters regarding such WWTP. Therefore, this study aims to create a reliable AI model for producing synthetic data that accurately reflects the parameters of a typical activated sludge system. This will help assess the suitability of AI techniques for monitoring and assisting with the prompt identification and taking remedial action needed for existing or newly developed WWTPs using ANN tools.

Chapter 3: METHODOLOGY

3.1 General

The resources and procedures used to complete this thesis work are discussed in this chapter. Along with a brief review of the system architecture of various machine learning and deep learning models used in this work, the generation of synthetic data, pre-processing, feature selection, and dataset reassembly are all comprehensively explained. Collecting field samples from various sites in Chattogram City and conducting tests are also covered. The usefulness of several AI tools was evaluated, and the optimum machine learning tool for assessing the performance of activated sludge systems was determined through a series of tests. The whole operation of this study is classified into four categories: 1) synthetic data generation using AI based machine learning algorithm 2) collection of sample from domestic influent and/or water bodies/sources, 3) extensive laboratory analysis, and 4) model performance evaluation.

3.2 Wastewater Quality Parameter

In this study, all the values of different parameters related to wastewater quality and treatment facilities are either selected from the literature with appropriate references or following the effluent discharge parameters recommended in Bangladesh Environment Conservation Rules (BECR, 2023; Tchobanoglous et al., 2003).

- **BOD_5**

BOD_5, a significant parameter in stream pollution control and one of the regulatory standards for effluent discharge, gives an idea of any sample's biodegradability and the strength of the waste. Wastewater is considered weak, medium, strong, and very strong if the BOD_5 concentration of wastewater is 200, 350, 500, and more than 750 mg/L, respectively (Mara, 2004). Fecal sludge (FS) typically has a much higher BOD_5 than that of strong wastewater. If the FS is digested and old, then the BOD_5 value is lower than the fresh and undigested FS. The BOD_5 value ranges from 840 to 2,600 mg/L for septic tank sludge, which goes up to 7,600 mg/L for public toiled sludge (Koné & Strauss, 2004). In contrast, the highest BOD_5 values (14,000 to 33,500 mg/L) are found in feces (Rose et al., 2015). In this study, the influent BOD_5 is considered in the 150–400 mg/L range, and the effluent BOD_5 is 0-70 mg/L (BECR, 2023; Davis & Cornwell, 2013; Tchobanoglous et al., 2003).

- **COD**

COD serves as a crucial design parameter for treatment plants and regulatory standards for effluent discharge. It is employed to assess the concentration of organic and inorganic contaminants in wastewater. Analyzing COD in the laboratory is more convenient compared to BOD_5, as it typically takes only a few minutes to hours, depending on the method. Due to the high variability and concentrations of organic matter, COD is generally deemed more accurate than BOD_5, particularly in the analysis of fecal sludge. The ratio of BOD_5 and COD indicates the biodegradability of the wastewater/effluent. Environmental authorities and wastewater treatment facilities monitor the COD variation to manage the impact of wastewater discharges on water bodies. This monitoring aids in the development of optimal treatment procedures to minimize pollution and protect the environment. For this study, the influent COD is considered within the range of 250-1000 mg/L (Davis & Cornwell, 2013; Garg, 2014; Tchobanoglous et al., 2003).

3.3 Wastewater Treatment Operational Parameter

To achieve the desired level of wastewater treatment and meet regulatory discharge requirements, controlling and optimizing key parameters such as volumetric loading, MLSS, MLVSS, F/M, HRT, and SRT in the activated sludge system is essential. Volumetric loading, one crucial parameter in wastewater treatment, refers to the amount of organic or pollutant load that the activated sludge system can efficiently handle within a given capacity of an aeration tank or reactor.

- **MLSS and MLVSS**

All suspended particles in wastewater, including organic materials, inert particles, and microbes, are referred to as MLSS in the mixed liquid of an ASS. However, MLVSS focuses on the organic and biologically active portions of the suspended solids. Monitoring and controlling MLSS levels are essential for optimizing the efficiency of wastewater treatment plants by ensuring an adequate presence of microorganisms for effective contaminant removal. MLVSS is particularly crucial for assessing the health and activity of the biomass in the treatment process. For this investigation, as suggested in different guidelines, MLSS and MLVSS are considered as 1000–3000 mg/L and 800-2400 mg/L, respectively (Davis & Cornwell, 2013; Garg, 2014; Karia & Christian, 2013).

- **F/M Ratio**

The F/M ratio holds significant importance as it offers wastewater treatment operators and engineers valuable insights into and control over the biological processes occurring within the activated sludge system. It helps understand the equilibrium between the incoming organic pollutants (food) and the microbial population responsible for breaking down these contaminants. By adjusting the F/M ratio, operators can fine-tune the performance of the activated sludge system to achieve efficient wastewater treatment and effective removal of pollutants.

This study takes into account a range of 0.2-0.4 lb BOD_5/day for F/M, as recommended by different guidelines (Davis & Cornwell, 2013; Garg, 2014; Karia & Christian, 2013).

- **HRT**

HRT refers to the typical amount of time wastewater stays in the tank or reactor used for treatment. It establishes the time necessary for the activated sludge's microorganisms to efficiently break down and remove pollutants and organic materials from wastewater. A longer HRT can improve pollution removal since it gives more time for microbial activity. On the other hand, a shorter HRT can lead to less effective treatment. Controlling and adjusting the HRT is crucial to ensuring that the microorganisms have enough contact time with the wastewater to achieve the necessary degree of pollutant removal and treatment effectiveness. This investigation considers an HRT range of 4-8 hours, aligning with various recommendations provided in different guidelines (Davis & Cornwell, 2013; Garg, 2014; Karia & Christian, 2013).

- **SRT**

SRT, a crucial operational parameter in activated sludge wastewater treatment, represents the average duration that sludge or biomass particles remain within the treatment system before discharge. A longer SRT generally leads to more efficient biological treatment of wastewater because it gives microorganisms like bacteria and protozoa more time to break down organic matter and remove contaminants from the wastewater. This results in a more thorough reduction of pollutants and a higher-quality treated effluent. Conversely, a shorter SRT can reduce treatment efficiency and potentially lower water quality. This is because there might not be sufficient time for microorganisms to degrade the organic substances in the wastewater adequately.

In this study, SRT is considered 5–15 days, as suggested in different guidelines (Davis & Cornwell, 2013; Garg, 2014; Karia & Christian, 2013).

3.4 Synthetic Data Generation

The performance of WWTP modeling studies in the AI-based machine learning process depends heavily on the availability of time series data, as these are the main disturbances of a WWTP. In Bangladesh, no such real time series data is available. Again, obtaining sufficiently long and qualitatively adequate time series data has become increasingly challenging and difficult concerning security issues. Along this line of thinking, synthetic data generation is a promising tool for model training.

Dataset is the prerequisite to build, train, and test a model. However, in the absence of a wastewater treatment plant in Bangladesh, the prerequisite dataset for wastewater treatment plants is unavailable. Different wastewater parameters are needed to perform several tasks using Artificial Intelligence (AI), so a large data set with better dimensions is needed to feed the model. After rigorous searching and literature review, no proper dataset is available online that contains more samples with different parameters collected from WWTP. Though some datasets are available online, they have plenty of limitations, like, for instance, fewer wastewater parameters and fewer dependencies among parameters. In Bangladesh, no study exists that unveils the complex variable to model wastewater quality. Moreover, the existing dataset needs more complex terms and focuses on some basic WW parameters. Eventually, no WWTP is currently operating in Bangladesh either, and hence, no such database is available. However, in recent years, the Bangladesh government has taken the initiative to build WWTPs in Chattogram (The Business Standard, 2023).

3.4.1 Synthetic Data Generation with AI Based Model

Using AI models, a dataset containing various wastewater parameters was generated to address the limitations associated with estimating effluent wastewater characteristics. The dataset was created based on Volumetric loading (VL), Hydraulic retention time (HRT), and food-to-microorganism ratio (F/M) dependencies. Several values within fixed boundaries were used to generate a dataset.

Input parameters, including VL (Volumetric Loading), Aeration Tank Volume (V), F/M (Food to Microorganism Ratio), HRT (Hydraulic Retention Time), Design Flow Rate (Q_o), Aeration Tank MLSS (X), Percentile Volatile MLSS (P_{vol}), Mixed Liquor Suspended Solids (MLSS), Recycle Activated Sludge Flow Rate (Q_r), Sludge SS Concentration (X_w), Sludge Retention Time (SRT), Secondary Effluent TSS (X_e), Waste Activated Sludge Rate (Q_w), Primary Effluent TSS (X_o), Q_r/Q_o Ratio (Q_{ratio}), and Primary Effluent BOD_5 (S_o), Primary Influent COD (COD), and Mixed Liquid Volatile Suspended Solids (MLVSS), were provided as input to the model to generate output values representing Effluent BOD_5 (S_e), COD & TSS values.

The effectiveness of biological treatment systems, such as activated sludge processes, varies with temperature. The activity and growth of the microorganisms that break down substances in wastewater treatment are directly impacted by temperature. Elevated temperatures typically accelerate microbial activity, leading to rapid biological reactions and expediting organic matter breakdown. Additionally, temperature influences the rate at which biomass (microorganisms) reproduce and grow in wastewater treatment systems. Elevated temperatures can promote faster biomass growth, assisting in creating flocs that aid in removing pollutants. Bangladesh's historical climate maintains an average temperature of around 26^0C for most of the year. Because of this consistency, temperature is overlooked in data generation using AI-based models.

A total of 150,000 data points were generated initially. Total 3 variables (Volumetric loading, Hydraulic retention time, and Food to microorganism ratio) were considered primarily. Based on these 3 parameters, other parameters were calculated using the Eq. (3.1-3.9) as described below.

- **Based on Volumetric Loading:** Volumetric loading is a parameter that is needed to design WWTPs. The volumetric loading can be calculated by multiplying the BOD_5 with the influent inflow.

Based on 50,000 Volumetric Loading (VL) data, other variables and their corresponding values were generated.

At first, the volume of the Aeration tank is measured,

$$V = \frac{1000 * 8.34 * S_o * Q_o}{VL} \qquad \text{Eq. (3.1)}$$

where, V is the aeration tank volume, S_o is the primary effluent BOD_5, Q_o is the design flow rate and VL is the volumetric loading

$$V_{mg} = \frac{V * 7.48}{1{,}000{,}000} \qquad \text{Eq. (3.2)}$$

where V_{mg} is the aeration tank volume (Million Gallon)

Then, the Hydraulic Retention Time is measured,

$$HRT = \frac{24 * V_{mg}}{Q_o} \qquad \text{Eq. (3.3)}$$

where HRT is the hydraulic retention time, and Q_o is the design flow rate.

Moreover, Food to Microorganism Ratio (F/M) is calculated as follows:

$$F/M = \frac{8.34 * S_o * Q_o}{8.34 * \%Vol * X * V_{mg}} \quad \text{Eq. (3.4)}$$

where F/M is the Food-Microorganism ratio and S_o is the primary effluent BOD_5, Q_o is the design flow rate, X is the aeration tank MLSS

- **Based on HRT:** 50,000 data is generated based on Hydraulic Retention Time.

$$V_{mg} = \frac{HRT * Q_o}{24} \quad \text{Eq. (3.5)}$$

Then, the volumetric loading is measured,

$$VL = \frac{1000 * 8.34 * S_o * Q_o}{V} \quad \text{Eq. (3.6)}$$

Moreover, Food to Microorganism Ratio (F/M) is calculated as Eq. (3.4).

- **Based on F/M:** 50,000 data is generated based on Food to Microorganism Ratio (F/M).

$$V_{mg} = \frac{8.34 * S_o * Q_o}{8.34 * \%Vol * X * (\frac{F}{M})} \quad \text{Eq. (3.7)}$$

Then, the volumetric loading is measured using Eq. (3.6).

Furthermore, Recycle Activated Sludge Flow Rate (Q_r) and Waste Activated Sludge Rate (Q_w) are also calculated as follows. Considering $Q_o = Q_e$,

$$Q_r = \frac{Q_o(X - X_o)}{(X_w - X)} \quad \text{Eq. (3.8)}$$

$$Q_w = \frac{1}{X_w}(\frac{V_{mg} * X}{SRT} - Q_e * X_e) \quad \text{Eq. (3.9)}$$

3.4.2 Processing of Synthetic Data

The generated dataset was analyzed and assessed following the effluent discharge parameters recommended in Bangladesh Environment Conservation Rules and other standards shown in Table 3.1 (BECR, 2023). After undergoing filtering processes, 94,892 datasets were retained for 18 different parameters. This refined dataset was utilized to construct different AI based models designed to emulate real data. The data, with a large number of instances, was found to be more effective when fed to the model. Large data has a diversity that may be beneficial for model training with its linear and nonlinear relationship among variables, which makes the training robust and able to predict more precisely.

Table 3.1. Details of the variables for synthetic dataset

Variable Name	Unit	Range	Variable Name	Unit	Limit
VL	lb BOD/ day/1000ft^3	20-40	**SRT**	days	5-15
F/M	lb BOD/day/lb MLVSS	0.2-0.4	**Q_{ratio}**	---	0.25-0.75
HRT	hr	4-8	**S_o**	mg/L	150-400
Q_o	MGD	1.5-3.5	**S_e**	mg/L	0-70
X (MLSS)	mg/L	1000-3000	**COD**	mg/L	250-1000
MLVSS	mg/L	800-2400	**COD_e**	mg/L	0-180
Q_r	MGD	0.4-2.5	**X_o**	mg/L	180-240
X_w	mg/L	5000-10000	**X_e**	mg/L	0-40

3.4.3 Pre-processing and Normalization of the Synthetic Dataset

Preprocessing is the step of processing the data by manipulating it like cleaning, removing, adding, scaling, etc. Preprocessing data before using it for model training can reduce computational costs and increase the efficiency of the overall performance.

Min-max scaling, also known as min-max normalization or feature scaling, is a data preprocessing method used in machine learning and statistics to scale the values of a feature (variable) to a specific range. This technique rescales the values of every dataset of different features to a specific range without altering the inherent relationships among the original data points. The primary objective of normalization is to standardize all features within a dataset to a comparable scale. This aids in enhancing the performance of machine learning algorithms by facilitating the interpretation of relationships, mitigating the impact of outliers, and preventing certain features from disproportionately influencing the modeling process. Normalization is referred to as the Eq. (3.10), as presented below.

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad \text{Eq. (3.10)}$$

Where X_{scaled} is the rescaled value of the original data point X, X_{min} is the minimum value in the dataset of the feature, and X_{max} is the maximum value in the dataset for the same feature.

In this study, a min-max scaling approach, as outlined in Eq. (3.10), has been employed to normalize each parameter within a range of 0 to 1. The scaling values for a specific row, comprising 19 distinct parameters among 94,892 rows, are detailed in Table 3.2.

Table 3.2. Scaled value after normalization

Variable Name	Actual Value	Scaled Value	Variable Name	Actual Value	Scaled Value
VL	54	0.321	**SRT**	14	0.048
F/M	0.33	0.167	Q_{ratio}	0.37	0.218
HRT	5.81	0.032	S_o	209	0.237
Q_o	1.80	0.151	S_e	42	0.263
X (MLSS)	1906	0.453	**COD**	432	0.244
MLVSS	1144	0.261	COD_e	172	0.092
Q_r	2.53	0.073	X_o	191	0.177
X_w	5300	0.310	X_e	38.13	0.953

As an illustration, the SRT values, ranging from 5 to 15 days, undergo scaling, where the value 14 is transformed to 0.048. Similarly, for MLSS, with a range of 1000-3000 mg/L, the value 1906 is scaled to 0.453. Analogous procedures are applied to scale values for other parameters in the study.

3.5 Model Development

The development of machine learning and deep learning algorithms to forecast the BOD, COD, and TSS of wastewater treatment facilities operating in activated sludge systems is covered in this section. It describes the formulation of these methodologies, deep learning model architectures, model performance optimization through calibration, and accuracy evaluation through validation. It also emphasizes the use of artificial intelligence (AI)-generated and deterministic synthetic data sets for training and testing these predictive models in the context of pollution assessment and environmental monitoring. Python 3.11.5 was used to implement the models that were developed. Python is a well-known and flexible programming language with an easy-to-understand syntax that makes it straightforward for developers to create and maintain code. Because of its extensive ecosystem of libraries and frameworks, Python is a great choice for rapid prototyping and development as it offers time and effort savings. It is appropriate for many applications, including web development, data science, machine learning, and more, because of its simplicity, adaptability, extensive library, and supportive community. It is an increasingly popular choice for beginners and experienced developers because of its clarity and readability.

3.5.1 Machine Learning Models

Machine Learning (ML) models rely on finding relationships between output and input data that enable predictions and/or simplify decisions (Müller & Guido, 2016). A significant advantage is that ML models imitate actual reaction/process conditions rather than pre-designed processes based on fundamental concepts. They are highly robust and thorough, which is crucial because many mechanisms involved in wastewater treatment are still poorly understood (Chan & Huang, 2003; Girgin et al., 2010; Laskov & Lippmann, 2010; Nadiri et al., 2018; Sharafati et al., 2020). Therefore, ML modeling is commonly used in WWTPs (Cao & Yang, 2020; C. Guo et al., 2020; Hu et al., 2020; Lu & Ma, 2020; A. Verma et al., 2013; L. Wang et al., 2020) to forecast effluent wastewater characteristics based on data-driven decisions.

To find out predictive results for specific wastewater characteristics, this study uses seven distinct algorithms from among numerous machine learning models, including Decision Tree, Random Forest, Extra Trees, Multivariate Linear Regression, K-Neighbors Regression, Gradient Boosting Regressor, and Adaboost Regressor, as discussed in the following sections.

3.5.1.1 Decision Tree

A decision tree, a supervised learning tool, acts as a tree-structured classifier. Internal nodes represent dataset features, branches symbolize decision rules, and leaf nodes indicate outcomes. This approach has been effective in wastewater effluent modeling with limited data (Celik et al., 2013). However, this study evaluates its suitability with an increased number of input variables compared to prior research.

Using the default parameter setting, a decision tree has been trained using the training dataset after the necessary pre-processing. The trained decision tree consists of a root node, a set of intermediate nodes at different tree levels, and a set of leaf nodes at the lowest level of the tree. There are, in total, 17 input features with numeric values. The root and intermediate nodes learn to make decisions based on the inequality of an input parameter. As illustrated in Fig. 3.1, the root node splits into left and right branches based on the value of COD_e (COD_e<=0.405). Similarly, the intermediate nodes are split into left and right branches based on the inequality of input features. It uses the Gini impurity to reach a decision at each node. The Gini impurity measures how well a split separates the data into different categories. The better split is found for the lower Gini impurity. Finally, each leaf node decides the output results of the target values of BOD_5, COD, and TSS based on the average output of all the samples in the training dataset.

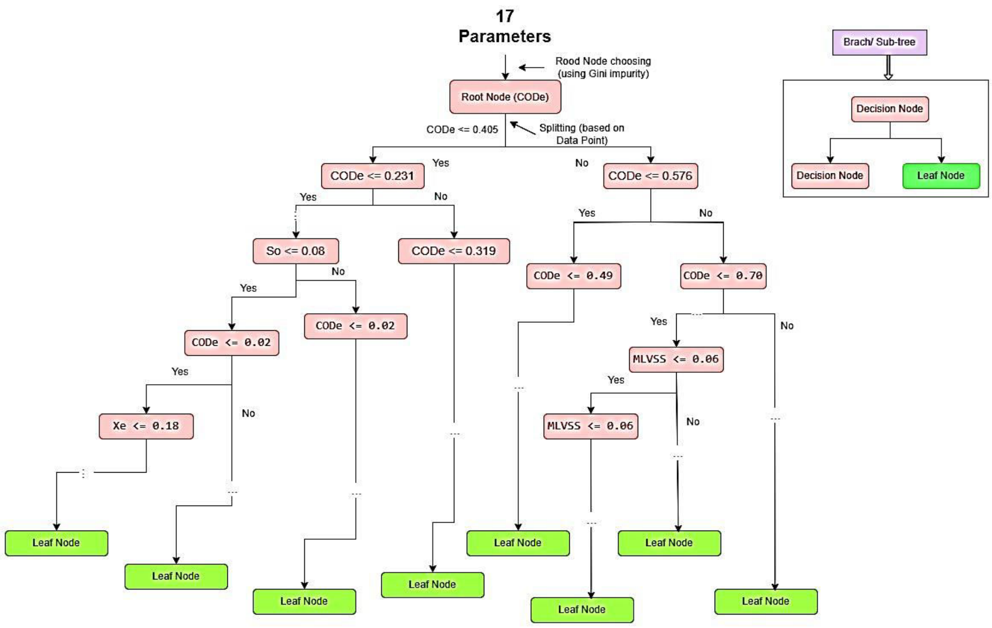

Fig. 3.1 Decision tree structure

3.5.1.2 Random Forest

Random forest, a versatile supervised machine learning algorithm, is used in classification and regression problems. Building upon decision trees on different samples, it delivers a consolidated and more accurate result, taking the average for regression and majority vote in the case of classification. With limited input parameter, this method has proven successful in wastewater discharge modeling (Cheng et al., 2023). In contrast to earlier research, this study assesses its applicability using a larger number of input variables.

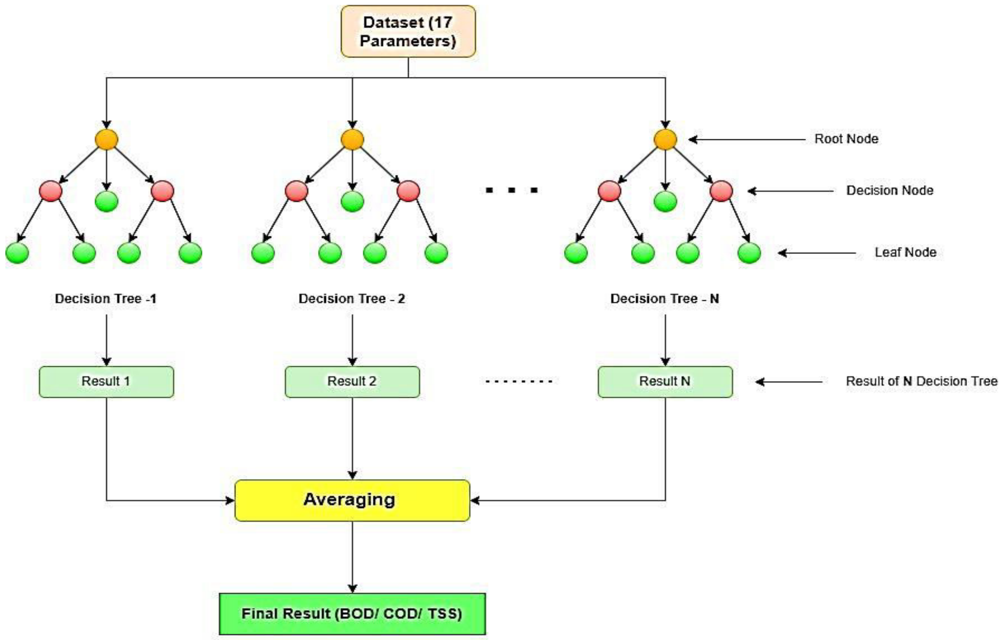

Fig. 3.2 Random forest structure

In this study, random forest creates hundreds or even thousands of decision trees, each trained on a different random sample of the training data from 94,892 data points on 18 wastewater parameters. This process is known as bootstrap aggregating or bagging. Each tree in the forest learns a slightly different collection of patterns and characteristics from the data by training on various subsets, as shown in Fig. 3.2. To further enhance the diversity of the trees, random forest introduces randomness when selecting features at each node of the tree.

Instead of considering all features for each split, it randomly selects a subset of features to evaluate. This randomness reduces the correlation between trees and improves the overall model's generalization ability. Once all the individual decision trees are trained, random forest combines their predictions through a process called ensemble averaging for target values of BOD_5, COD, and TSS.

3.5.1.3 Extra Trees

Extremely Randomized Trees, often called Extra Trees, is a well-known ensemble machine learning technique. This approach has been effective in determining discharge coefficients related to the discharge characteristics of rectangular, sharp-crested side weirs (Hameed et al., 2021). However, a few studies related to wastewater effluent modeling are available but insignificant. This study evaluates its suitability for wastewater effluent quality modeling with an increased number of input variables.

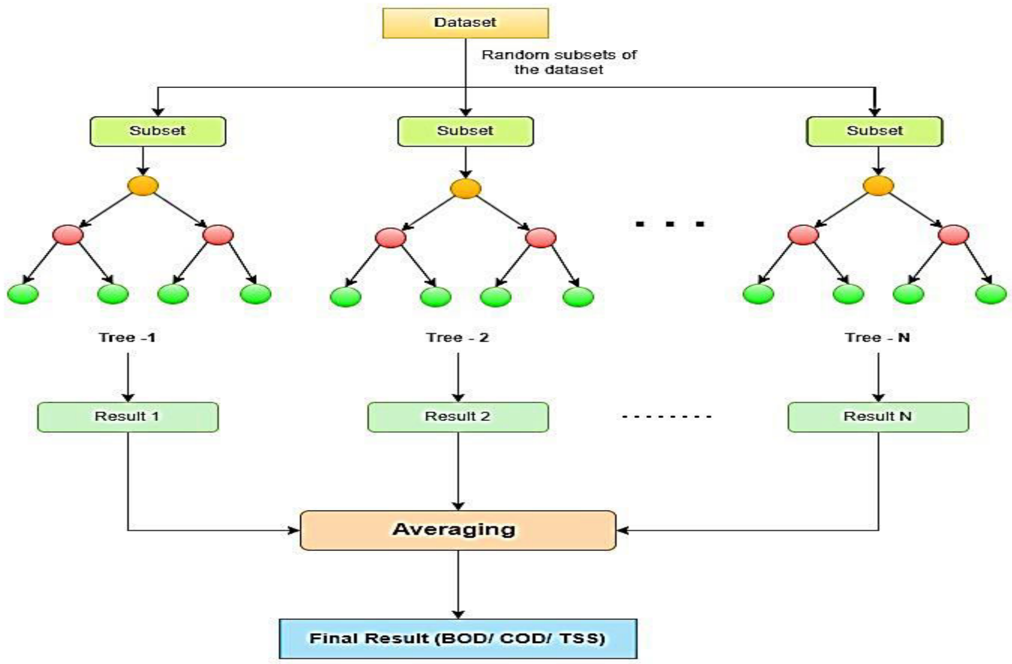

Fig. 3.3 Extra trees structure

Although it has many similarities with other tree-based algorithms, like random forest, it adds a unique twist to how individual decision trees are built inside the ensemble. First, for each decision tree in the ensemble, extra trees randomly choose portions of the training data and the features that are accessible from the 94,892 data points on 18 wastewater parameters considered in this investigation. As a result, the technique is less prone to overfitting because it does not take into account the complete dataset or all of the features when building a single tree. As seen in Fig. 3.3, Extra Trees tries to build diverse and uncorrelated decision trees by employing these random subsets. Second, Extra Trees introduces additional randomness by choosing the split thresholds randomly. This increases the noise in each tree's decision, improving the ensemble's overall performance. Like decision tree and random forest, the final prediction for a particular output, like BOD_5, COD, and TSS, is often chosen by a majority vote and averaging in classification and regression tasks, respectively.

3.5.1.4 Multivariate Linear Regression

The link between several independent variables and a single dependent variable can be modeled and analyzed via the statistical technique known as multivariate linear regression. It is an improvement of straightforward linear regression to ascertain how changes in the independent variables affect the dependent variable. With limited input parameters, this method has proven successful in wastewater effluent quality modeling (Cheng et al., 2023). In contrast to earlier research, this study assesses its applicability using a larger number of input variables.

The fundamental idea behind multivariate linear regression is to express the dependent variable as a weighted sum of the independent variables, with each independent variable assigned a coefficient that represents its impact on the dependent variable based on a dataset of 94,892 data points on 18 wastewater parameters. The equation for multivariate linear regression can be expressed in Eq. (3.11) as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad \text{Eq. (3.11)}$$

In this study, Y is the dependent variable like BOD, COD, and TSS, and the independent variables SRT, HRT, F/M, MLSS, MLVSS, etc. are indicated by X_1, X_2,..., X_n; the intercept is β_0; the regression coefficients are β_1, β_2, ..., β_n; and the error term is ε, which takes into account the dependent variable's unexplained variance.

3.5.1.5 K Nearest Neighbors Regressor

One of the most basic machine learning algorithms, K-nearest neighbor (KNN), is based on the supervised learning approach. This approach has shown efficacy in wastewater effluent modeling with few input parameters (Kim et al., 2015). This study evaluates its applicability using a larger number of input wastewater parameters.

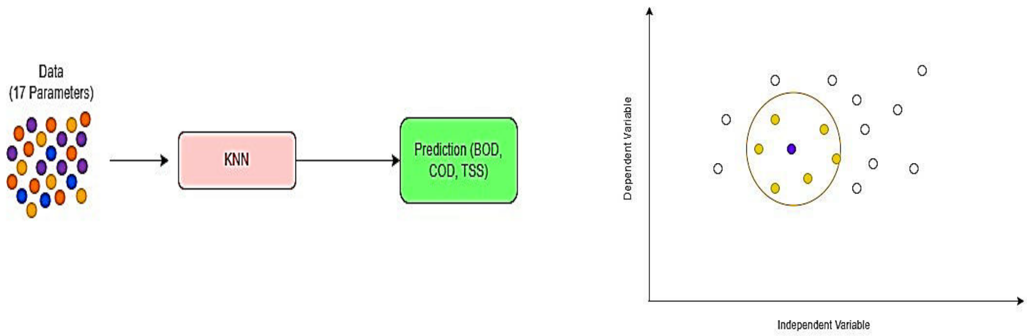

Fig. 3.4 KNN structure

The KNN algorithm stores a certain portion of 94,892 data points on 18 wastewater parameters and classifies a new data point for specific effluent parameters like BOD_5, COD, and TSS based on their similarity. It is a hyperparameter that can be tuned based on a specific problem. A smaller K value, such as 1, means that the algorithm will consider only the single nearest neighbor. In comparison, a larger K value considers more neighbors of a new input data point in the training dataset to calculate its target value when it is given for prediction. Once the K-nearest neighbors are identified, the KNN regressor predicts the target value for the new data point by averaging the target values of its K neighbors (for regression problems). As illustrated in Fig. 3.4, new data is assigned to a particular category based on the K value. The averaging mechanism makes the KNN regressor sensitive to outliers, as a single neighbor with an extreme target value can heavily influence the prediction. It is versatile and easy to implement but requires careful tuning of the K parameter and may not be suitable for high-dimensional data or datasets with outliers.

3.5.1.6 Gradient Boosting Regressor

The basic principle of gradient boosting regressor is to progressively train several weak learners, often decision trees, and then combine their predictions to minimize the prediction errors. This method has demonstrated effectiveness in modeling wastewater effluent with limited input parameters (Zhang et al., 2023). This research assesses its suitability by incorporating more input parameters related to wastewater treatment.

The initial weak learner, frequently a shallow decision tree, is where the gradient boosting regressor's mechanism starts, as seen in Fig. 3.5. This first model can approximate the target variables such as BOD_5, COD, and TSS. The system then finds gaps or residuals between the actual target values and the first learner's predictions. These residuals represent the areas where the model's predictions are most inaccurate.

In the subsequent steps, new weak learners are created, focusing on learning from these residuals. Each new learner is trained to minimize the errors that the previous learners could not capture effectively. This process continues iteratively. The prediction from each new learner is added to the ensemble, and the residuals are updated. This step-by-step learning process boosts the model's performance by correcting its errors. The final prediction in a gradient boosting regressor is obtained by summing up the weak learners' predictions.

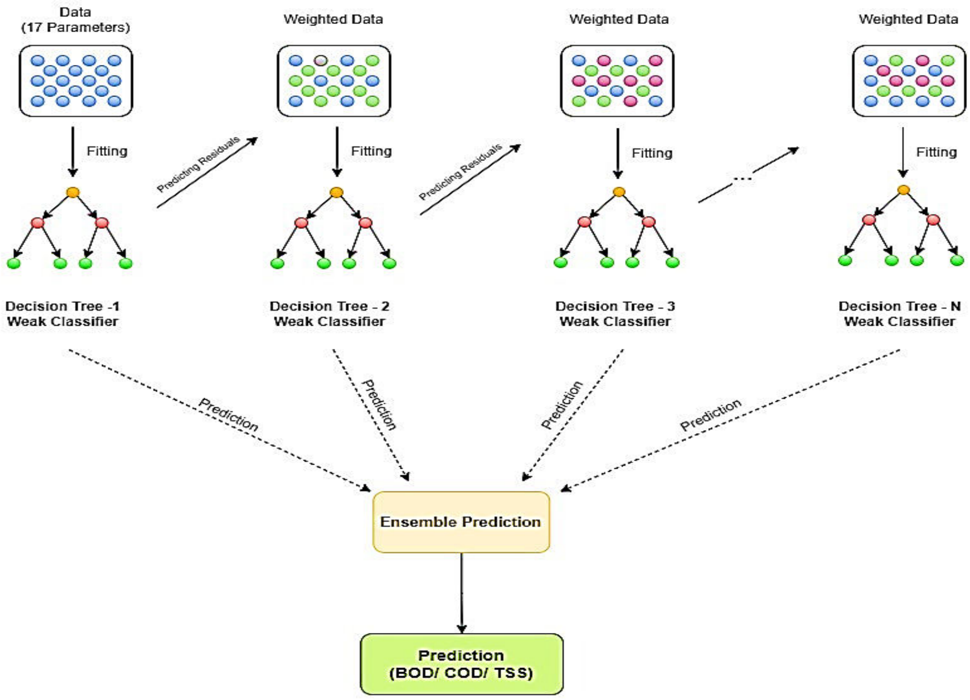

Fig. 3.5 Gradient Boosting structure

3.5.1.7 Adaboost Regressor

The adaboost regressor is a boosting algorithm that follows the stagewise addition method. It utilizes multiple weak learners to form strong learners collectively. It emphasizes weak learners' past errors and gives more weight to training samples that make inaccurate predictions. This periodic process continues until a set number of weak learners have been trained, or a predetermined accuracy level is reached. This method has demonstrated effectiveness in modeling wastewater effluent with limited input parameters (Bilali et al., 2021). This research assesses its suitability by incorporating more input parameters related to wastewater treatment.

Each training data point, a certain portion of 94,892 data points on 18 wastewater parameters, is assigned an equal weight, and a weak learner is trained on this weighted data, as illustrated in Fig. 3.6. The weak learner's goal is to minimize the error in predicting the target variable. After training the first weak learner, its performance is evaluated, and the algorithm identifies which data points were mispredicted. These misclassified data points are then given higher weights, effectively making them more important for the next weak learner.

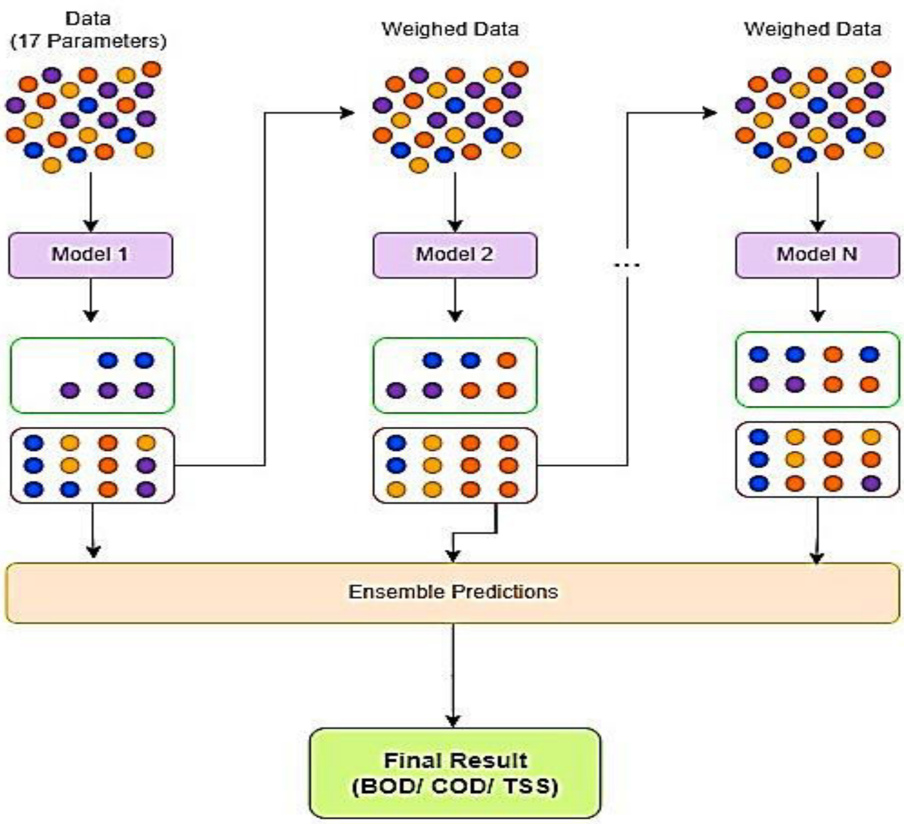

Fig. 3.6 Adaboost regressor structure

In subsequent iterations, additional weak learners are trained, and they focus more on the previously misclassified data points. The algorithm keeps track of each weak learner's performance and assigns a weight to it based on its accuracy. Weak learners that perform well are given higher weights in the final ensemble, while those with lower accuracy receive lower weights. All weak learners' predictions are merged and weighted according to performance for making predictions using the adaboost regressor. This weighted combination produces the final prediction values of the target variable like BOD_5, COD, TSS.

3.5.2 Deep Learning Models

It is possible to predict how well WWTPs will perform using Artificial Neural Networks (ANNs). Due to ANNs' exceptional prediction accuracy, the water treatment process was replicated using them (Alver & Kazan, 2020; Manu & Thalla, 2017). These studies used various ANNs as the model structure's framework for data analysis and feature extraction.

There may be obstacles to overcome during the model development such as the task of selecting the proper model parameters, i.e., the number of inputs and outputs the model must consider or the number of neurons in the hidden layer. Unfortunately, as this is an empirical operation, it must be carried out until acceptable results are produced using a trial-and-error approach. This is because there is little information to help the user choose a specific model. Therefore, it is necessary to analyze those models and assess their ability for prediction.

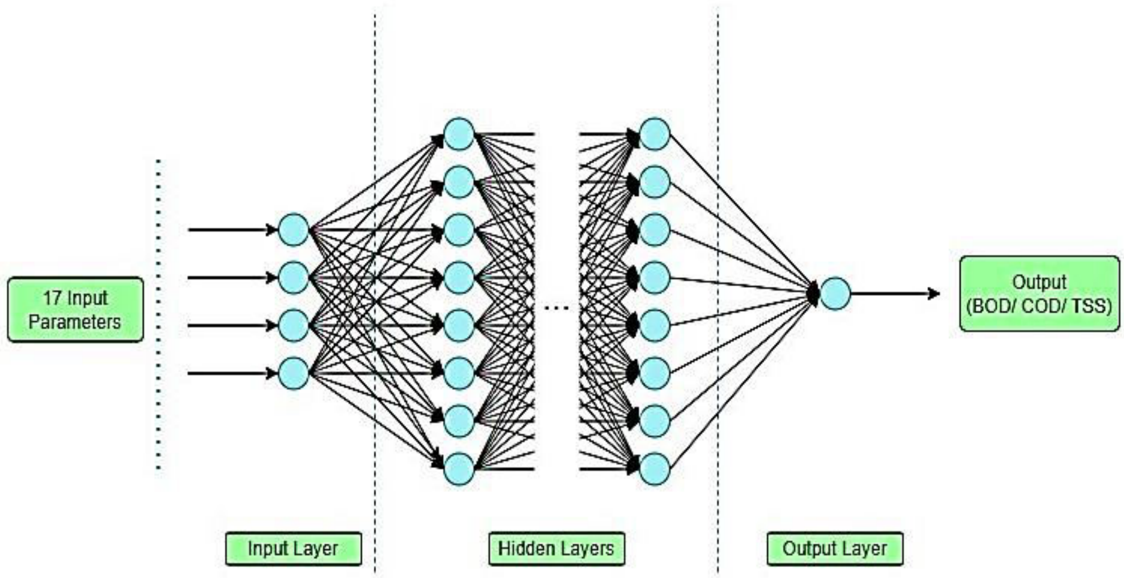

Fig. 3.7 Artificial neural network structure

The structure and operations of the human brain are the basis for ANNs, a fundamental concept in the disciplines of machine learning and artificial intelligence. ANNs are a type of computational model that aims to replicate how neurons in our brains process information. They consist of interconnected nodes called neurons, organized into layers as seen in Fig. 3.7. This investigation employed three ANN structures: ANN-I, ANN-II, and ANN-III. The optimizer and activation functions in Table 3.3 served as the basis for these variations.

The core element of an ANN is the artificial neuron, which takes in multiple inputs, assigns weights to them, processes the weighted sum, and then applies an activation function to produce an output. The weighted sum considers the relative value of each input, and the activation function provides non-linearity into the model, enabling ANNs to learn complex patterns and relationships in data.

Three different layer types are commonly present in ANNs: the input layer, one or more hidden layers, and the output layer. The input layer receives the initial data, the hidden layers perform a series of calculations to process and transform it, and the output layer provides the final result or prediction. Weights attached to the connections between neurons are modified during training, enabling the network to learn and adjust to data patterns. This study used a trial-and-error method to identify the number of hidden layers and nodes based on prediction performance. Detailed information regarding the analysis conducted with varying numbers of hidden layers and neurons can be found in Appendix A.

An optimization approach, such as gradient descent, is used to update the weights of an ANN to minimize the gap between the network's predictions and the actual labels after feeding it with labeled data (input-output pairs). A loss function, which measures the gap between forecasts and actual values, directs this process. ANNs can learn complex tasks by iteratively modifying the weights via gradient descent and backpropagation.

In the mechanism of ANN, the code initiates a sequential model, a linear stack of layers where layers are added one by one. Different ANN architecture details used in this study based on the number of hidden layers as well as dense variations in input, hidden, and output layers are illustrated in Table A1 of Appendix A. Taking ANN-I as an example from Table 3.3, the first layer comprises a dense layer with 18 neurons for 17 input features. This model includes three hidden layers with progressively increasing neuron counts (32, 64, 32).

Table 3.3. ANN variation with optimized network architecture

Model Name	Network Architecture	Optimizer	Activation Function
ANN-I	18-32-64-32-1	Adam	ReLU
ANN-II	18-64-124-64-1	SGD	Sigmoid
ANN-III	18-32-64-32-1	Adam	SELU

Each hidden layer is followed by a leaky ReLU activation function, incorporating specific alpha parameters (0.1 for the first two layers and 0.2 for the third layer) to mitigate the vanishing gradient problem. The final dense layer, containing 1 neuron, employs the ReLU activation function, indicating the model's design for regression tasks, intending to predict a single continuous numerical value. The model is compiled using the Adam optimizer, a widely used gradient descent optimization algorithm. Two callbacks are implemented: early stopping, which monitors validation loss and halts training if the loss stagnates for 2 consecutive epochs to prevent overfitting, and a model checkpoint to save the best model based on validation loss. Similar mechanisms are applied to ANN-II and ANN-III.

One of the key benefits of ANNs is their capability to learn independently and extract characteristics from raw data, making them highly versatile for various tasks. In order to perform at their best, ANNs may need to be carefully designed and tuned to address issues, including requirements for enormous quantities of data, overfitting, and the complexity of model structures. Nevertheless, ANNs continue to be an effective tool in AI and machine learning, enabling the development of intelligent systems capable of understanding and interpreting complex data.

3.6 Field Sample Collection, Preservation and Testing

In Bangladesh, there is no data regarding the Wastewater Treatment Plant (WWTP) running with the activated sludge process. From the literature review, it is clear that future wastewater treatment will be based on ASS; hence, data and information regarding essential parameters for performance evaluation are key. These data and parameters encompass various aspects of the treatment process and wastewater characteristics. These parameters collectively represent the condition of wastewater before treatment, the state of the WWTP during its operation, and the quality of wastewater after undergoing treatment. Access to this comprehensive dataset would undoubtedly enhance the present understanding of wastewater treatment processes and contribute significantly to improving environmental management in an integrated manner.

3.7 Study Area

A significant daily generation of wastewater is witnessed in Bangladesh's urban areas, particularly in densely populated cities where wastewater disposal occurs. In these areas, the opportunity exists to collect wastewater samples and measure the values of various parameters that represent the condition of the wastewater before undergoing any treatment process.

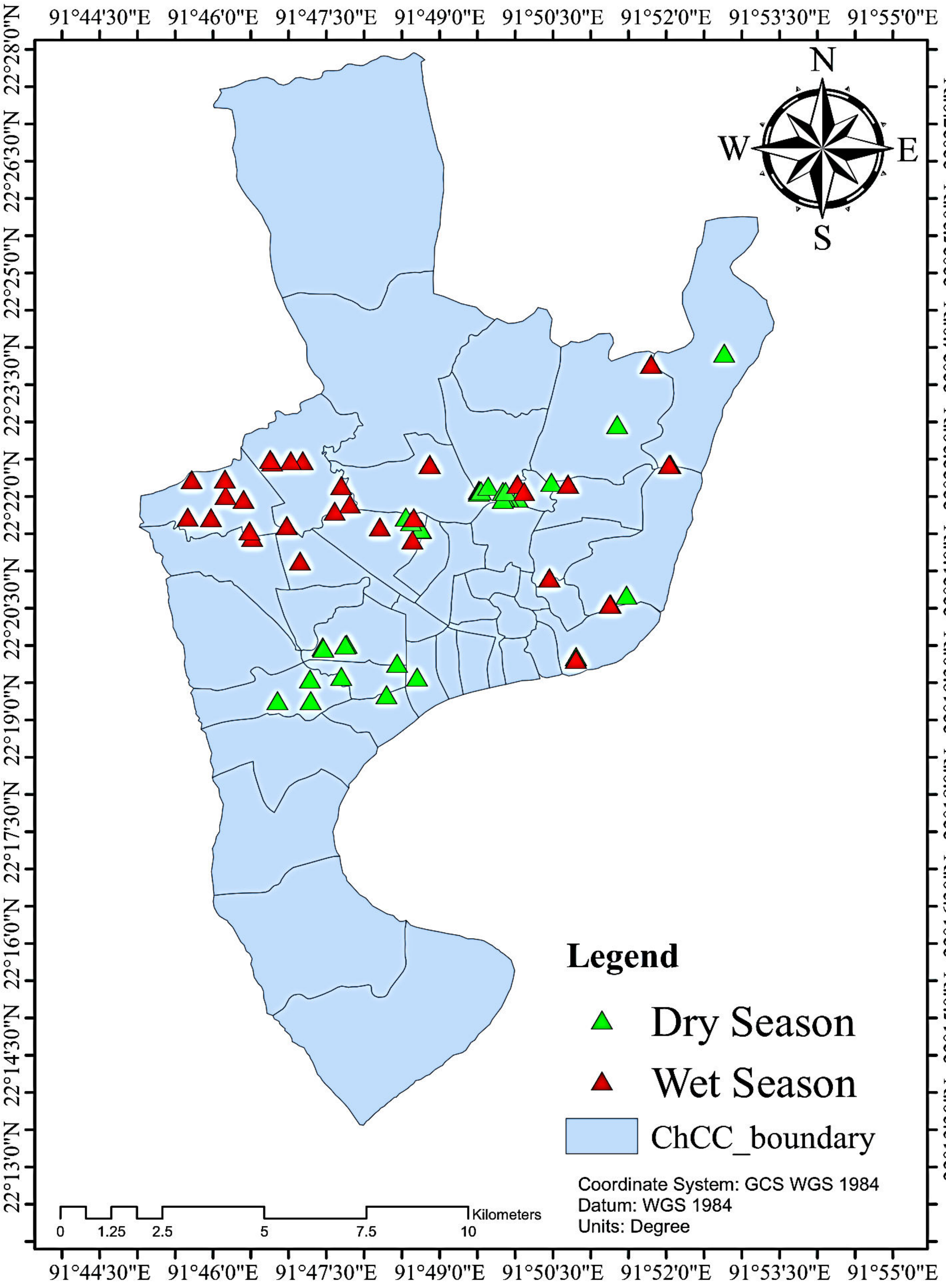

Fig. 3.8 Geographic locations for sampling in Chattogram

Gaining knowledge of the features and quality of the wastewater through the collection of real samples is essential for understanding the environmental challenges urban wastewater disposal poses. Such information can serve as a crucial foundation for developing effective strategies to address wastewater management and improve the overall environmental conditions in our city and country. Agrabad, Halishahar, Chandgaon, Suganda, and Khulshi have been chosen as study areas (as seen in Fig. 3.8) in Chittagong, where grab samples, including drains from residential areas, septic tanks, and pits, were collected.

3.7.1 Sample Collection

Understanding the dry and wet seasons is vital for wastewater treatment in Bangladesh as they significantly impact flow rates and pollutant concentrations throughout the year. Dry periods, typically occurring from November to March, reduce wastewater flow, potentially increasing pollutant levels and affecting treatment plant efficiency. Conversely, the wet season, spanning from April to September, brings heavy rainfall and flooding, causing a surge in wastewater generation. This inundation mixes rainwater with sewage, posing challenges for treatment plants in managing increased volumes. While the wet season dilutes pollutants, easing treatment in some ways, the higher flow rates stress treatment infrastructure.

Fig. 3.9 Sample collection and on-spot parameter determination

Addressing this issue, three distinct time frames cover the collection of wastewater samples: two dry periods spanning December 2021 to February 2022 and December 2022 to February 2023, along with a wet period extending from May to August 2022. A total of 89 random samples were collected during the dry season (57 samples) and the wet (monsoon) season (32 samples) to undertake a thorough assessment of the condition. Health and safety protocols are ensured before, during, and after sampling. As illustrated in Fig. 3.9, important on-the-spot information was gathered throughout the sampling procedure, including the exact location, the time of sample collection, the current ambient temperature, and the pH of the sample.

In areas where temperatures change seasonally, causing colder winters and warmer summers, biological treatment facilities may encounter fluctuations in their efficiency owing to changes in microbial activity and reaction rates. Understanding how temperature impacts biological processes is crucial for effectively designing, operating, and optimizing wastewater treatment plants. Engineers and operators often consider temperature variations and implement strategies to optimize treatment performance across different temperature ranges. That's why actual sample temperatures are included to validate and assess the resilience of various models.

3.7.2 Sample Preservation and Storage

Following the collection of samples and recording on-site parameters, the samples were placed in an icebox and transported to the laboratory on the sampling day. Subsequently, they were stored in a refrigerator until additional parameter measurements were conducted. Table 3.4 summarizes the various preservation techniques followed in this study, and the maximum suggested storage duration (UPM, 2021).

Table 3.4. Overview on sample preservation methods

Parameter	Preservation Method	Maximum Storage Duration	Source
BOD, COD	1-5 °C, dark and airtight	48 hours	Methods for the Chemical Analysis of Water and Wastes, EPA-600/4-79-020, USEPA, EMSL, 1979 ISO 5667-3:2012
TSS, VSS	1-5 °C, dark and airtight	7 days	Methods for the Chemical Analysis of Water and Wastes, EPA-600/4-79-020, USEPA, EMSL, 1979 ISO 5667-15:2009

3.7.3 Laboratory Test

In this investigation, the concentrations of BOD_5, COD, TDS, TSS, and VSS in the raw sample were assessed through three trials employing established standard methods as shown in Fig. 3.10. Table 3.5 outlines the chosen methodologies and specifies the apparatus utilized for parameter concentration determination. A rigorous Quality Assurance/Quality Control (QA/QC) program was implemented to ensure the accuracy and reliability of the laboratory analysis of water samples.

Table 3.5. List of analysed parameters and analytical procedures

Parameter	Unit	Method	Apparatus
BOD_5	mg/L	SM 5210B	BOD Bottle
COD	mg/L	SM 5220D	HACH DR 6000 UV Spectrophotometer
TSS	mg/L	SM 2540 D	Oven
TDS	mg/L	SM 2540 B-D	Oven
VSS	mg/L	SM 2540 D-E	Muffle Furnace

Fig. 3.10 Wastewater sample arrangement and testing in laboratory

- **BOD_5**

Two BOD bottles containing a sample (or a diluted sample) were filled, and the initial measurement of dissolved oxygen (DO) was promptly conducted in one bottle. The other bottle was stored in darkness at 20°C, and after 5 days, the DO (DO_f) in the sample (or diluted sample) was assessed. To determine DO, 1 mL of manganous sulfate solution was introduced into the BOD bottle using a pipette immersed just beneath the water's surface. Subsequently, 1 mL of alkaline potassium iodide solution was added to the BOD bottle. After inserting the stopper, the solution was mixed by repeatedly inverting the bottle. The precipitates were allowed to settle halfway, followed by another round of mixing and settling. Next, 1 mL of concentrated H_2SO_4 was added, and the stopper was immediately inserted, followed by mixing as before. The solution was left undisturbed for at least 5 minutes. 100 mL of solution is transferred into an Erlenmeyer flask, and immediately adding 0.025N sodium thiosulfate drop by drop from a burette is done until the yellow color almost disappears. Subsequently, approximately 1 mL of starch solution is introduced, followed by adding thiosulfate solution until the blue color disappears. The quantity of thiosulfate solution utilized (excluding any reoccurrence of the blue color) is then recorded.

- **COD**

A 2 ml sample was taken into the COD vial, which could be either high or low range depending on the wastewater quality. Then, the sample was heated for 2 hours. Following this, the COD vial was allowed to cool for 30 minutes. Finally, the determination of COD was conducted using a spectrophotometer as seen in Fig. 3.11.

Fig. 3.11 Digestion of wastewater sample in COD reactor

In biological wastewater treatment, the BOD_5/COD ratio is an essential indicator of how easily organic matter degrades. Effective treatment system design and operation depend heavily on this information. A higher BOD_5/COD ratio indicates more readily degradable organic pollutants, which increases their susceptibility to removal by microbial action in biological treatment techniques. On the other hand, a lower BOD_5/COD ratio indicates a significant concentration of materials that are either non-biodegradable or decompose more slowly. This scenario may require alternative treatment approaches beyond conventional biological methods.

Operators of treatment plants can optimize processes with the help of this ratio. It aids in selecting appropriate biological treatment systems and determining essential treatment parameters, including microbial activity, retention period, and aeration. Continuous and efficient removal of organic contaminants from wastewater is ensured by monitoring variations in the BOD_5/COD ratio over time, which enables adjustments in the treatment process to account for changes in wastewater composition.

Untreated domestic wastewater usually has a BOD_5/COD ratio of 0.3 to 0.8. Ratios between 0.92 and 1 show complete biodegradability, and ratios greater than 0.6 indicate biological treatment acceptability. Ratios less than 0.3 suggest difficulties with biological treatment, whereas ratios between 0.3 and 0.6 may necessitate adjustments. Samples collected during dry and rainy periods have an average BOD_5/COD ratio of 0.48. It suggests that biological treatment is appropriate in the given local context.

- **TSS and TDS**

A fixed sample volume was taken in a beaker and dried in the beaker containing the sample at 103–105 °C for 24 hours, and the TS was measured. A Buchner funnel was set with a filter for TSS determination, and the vacuum pump was initiated. The sample was poured into the filter. After filtration, the filter paper kept in a crucible was dried at 103–105 °C for a minimum of 2 hours and weighed to obtain the TSS. The TDS was determined by subtracting TSS from TS.

- **VSS**

For VSS determination, the remaining material from the TSS was ignited in a muffle furnace at 550°C for a duration of 20 minutes. Then the crucibles were moved to a stainless steel tray, and after allowing them to cool, the crucible weights were recorded.

3.7.4 Quality Assurance and Quality Control

The systematic management and application of procedures to guarantee that the outputs constantly satisfy predetermined quality standards is known as quality assurance. Preventing errors, shortcomings, or deviations in processes is the primary objective of quality assurance (QA), which also aims to increase the system's overall dependability and efficiency. Standard operating procedures (SOPs), routine audits, and documentation to guarantee standards and regulations are followed are a few ways to illustrate QA operations.

To make sure that processes, procedures, and outcomes adhere to predefined standards and specifications, quality control entails monitoring and assessing these elements. To guarantee the accuracy and quality of the final outcome, quality control (QC) is concerned with finding and fixing faults or deviations in the processes. In wastewater management, routine sampling and analysis, equipment calibration, and the use of control charts to track trends and deviations are scenarios involving QC operations.

Whereas quality control is outcome-oriented and concentrates on fault diagnosis, quality assurance is process-oriented and concentrates on defect prevention. Producing comprehensive analytical data that faithfully represents the waste stream from which samples are gathered is the major goal of maintaining QA/QC. In addition to providing confidence about the correctness of the data generated and ensuring compliance with environmental regulations, QA/QC methods aid in the generation and maintenance of high-quality results.

Table 3.6. Approaches and technical specifications regarding the tools employed in laboratory test

Parameters	Name of the Instrument/ Methods	Range	Accuracy	Wastewater Quality Standards (BDS)
pH	HI9814 (GroLine)	-2 to 16	± 2.0%	6.5-8.5
Temp.	HI9814 (GroLine)	-5.0 to 105.0 °C	± 0.5%	30 ^{0}C
TDS	HI9814 (GroLine)	0 to 3000 mg/L	± 2%	---
TSS	Standard Methods (APHA, 2005)			100 mg/L
DO	HI98198 (Hanna: opdo)	0 to 50 mg/L	± 1.5%	---
BOD_5	Standard Methods (APHA, 2005)			30 mg/L
COD	Standard Methods (APHA, 2005)			125 mg/L
Note:	Sewage discharge into surface and inland water bodies BDS: Bangladesh Standards (BECR, 2023)			

The entire process, from sample collection to result finalization, adhered to standard protocols to ensure the accuracy and reliability of the results. Wastewater samples were collected in non-reflectable black bottles to minimize any potential interference from external light sources (APHA, 2005; Clesceri, 2012; Karia & Christian, 2013; Tchobanoglous et al., 2003). On-site testing of crucial parameters, such as pH, temperature, and TDS, was conducted immediately, recognizing the time-sensitive nature of these properties, which can deteriorate over time. The collected samples were placed in freezing buckets, maintaining a consistent temperature and preventing any substantial degradation in quality to preserve sample integrity during transport to the laboratory (APHA, 2005; Tchobanoglous et al., 2003). The water quality samples underwent comprehensive testing upon arrival at the laboratory, using state-of-the-art equipment detailed in Table 3.6. These instruments exhibited a high level of accuracy, ranging from ±0.5% to ±2%. Before conducting the tests, it is noteworthy that all equipment underwent thorough calibration processes, ensuring their precision and reliability in measuring water quality parameters. The emphasis on immediate on-site testing, coupled with meticulous calibration practices, contributes to the overall integrity of the data generated through this wastewater quality assessment.

In addition to maintaining standard testing procedures, utmost attention was given to laboratory safety throughout the entire testing process. Laboratory personnel adhered to established safety protocols to mitigate potential risks associated with handling wastewater samples and operating testing equipment (APHA, 2005; Clesceri, 2012; Spellman & Drinan, 2003). Aprons and other appropriate personal protective equipment were consistently worn to minimize direct contact with the samples and potential exposure to hazardous substances. Furthermore, proper ventilation and fume hood usage were observed to ensure a safe working environment. Emergency response protocols and safety equipment, such as eyewash stations and fire extinguishers, were readily available and regularly inspected.

To guarantee the repeatability of analytical procedures, duplicate analyses are performed on every sample for every parameter. Every batch contains blank samples, which are used to detect contamination during analysis. The findings of the blank samples are monitored to make sure they remain within acceptable limits. In order to identify the source of the error in the analytical method, troubleshooting and appropriate corrective steps are conducted if the difference between duplicate analyses carried out on the same sample exceeds 10 percent.

3.8 System Architecture

The work in this research has been separated into four categoriess, and the choice to do so was made to comprehensively evaluate machine learning and deep learning models using various methods. Table 3.6 provides a summary of each category's details. It should be emphasized, though, that each category still has the same basic idea at its core. Therefore, a typical process for each category is shown in Fig. 3.12, along with an overview of how the machine learning and deep learning models were used.

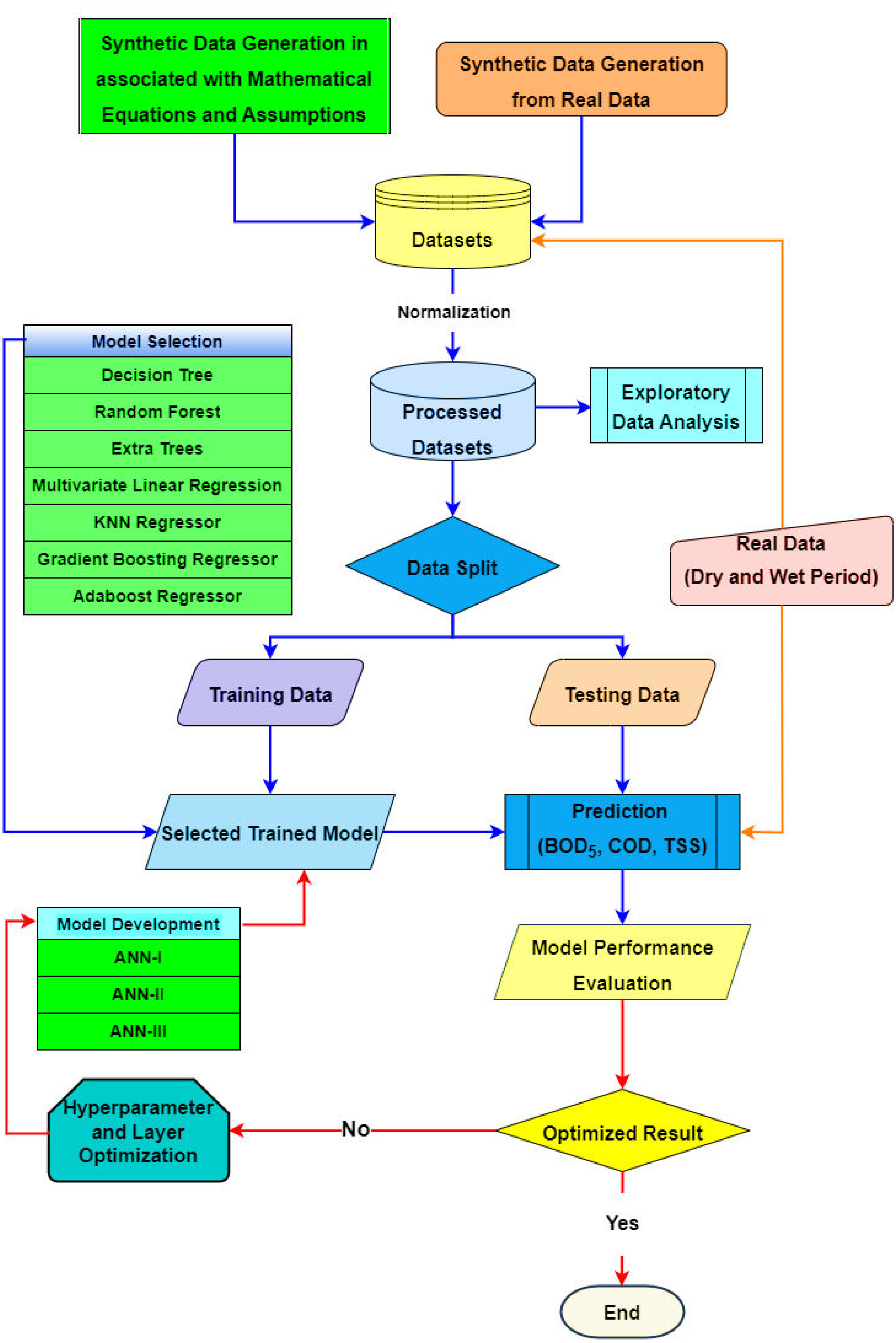

Fig. 3.12 Flow chart of the study

3.9 Model Calibration and Validation

To thoroughly assess machine learning and deep learning models using a variety of approaches, the analysis has been divided into four categories, and the details of the input, output, and variations in the training and testing data sets are highlighted in Table 3.6. The training set comprised 80% of the total data for the category-1 synthetic dataset associated with mathematical equations and assumptions, while the testing set included the remaining 20%. Data from 18 different parameters (VL, V, F/M, HRT, Q_o, MLSS, MLVSS, Q_r, X_w, SRT, Q_{ratio}, Q_w, X_o, X_e, S_o, S_e, COD and COD_e) are used to assess the prediction ability of various models for BOD_5, COD, and TSS values.

Again, to check the capability of various models, actual samples of 89, including drains from residential areas, septic tanks, and pits from various sites in Chittagong, were collected during the dry season (57 samples) and the wet (monsoon) season (32 samples).

In Category 2, aiming to evaluate how well various models, based on 18 operational and qualitative parameters of the ASS system, could accurately represent any conditions, the models were initially trained using synthetic data generated in Category 1 for BOD_5, COD, and TSS values and subsequently assessed using actual data for the same parameters.

As seen in Table 3.6, for Category 3, the dataset collected during the dry season was divided into two portions: training and testing, with 80% and 20% of the data, respectively. The dataset collected during the wet period followed the same data-splitting process. The models were then tested by predicting the wet season data using the dry season data and vice versa. The data from the dry and wet seasons were integrated into a single dataset as the final phase of this category. This combined dataset was divided into two portions, with 80% allocated for training the models and the remaining 20% reserved for testing purposes. This comprehensive approach allowed for a thorough evaluation of the models' ability to represent the observed data, both within and between the dry and wet seasons. Importantly, this analysis was carried out without utilizing any mathematical equations associated with the operation of ASS.

Table 3.7. Selection of categories to address variability in ASS modeling

<table>
<tr><th>Category</th><th>Training</th><th>Testing</th><th>Input</th><th>Output</th><th>Models</th></tr>
<tr><td>Category-1 Synthetic Data (in association with mathematical equation and assumptions)</td><td>Synthetic Data (80%)</td><td>Synthetic Data (20%)</td><td>18 parameters (VL,V, F/M, HRT, Q_o, X, MLVSS, Q_r, X_w, SRT, Q_w, Q_{ratio}, X_o, X_e, S_o, S_e, COD, and COD_e)</td><td rowspan="11">BOD_5
COD
TSS</td><td rowspan="11"><u>ML</u>
• Random Forest
• Decision Tree
• Extra Trees
• Multivariate Linear Regression
• K-Neighbors Regression
• Gradient Boosting Regressor
• Adaboost Regressor

<u>DL</u>
• ANN-I
• ANN-II
• ANN-III</td></tr>
<tr><td>Category-2</td><td>Synthetic Data (Category-1)</td><td>Real Data</td><td>BOD_5
COD
TSS</td></tr>
<tr><td rowspan="5">Category-3 (Real Data)</td><td>Dry (80%)</td><td>Dry (20%)</td><td rowspan="9">HOUR, MINUTE, Temp., pH, BOD_5, COD, TDS, TSS, VSS</td></tr>
<tr><td>Wet (80%)</td><td>Wet (20%)</td></tr>
<tr><td>Dry (100%)</td><td>Wet (100%)</td></tr>
<tr><td>Wet (100%)</td><td>Dry (100%)</td></tr>
<tr><td>Dry + Wet (80%)</td><td>Dry + Wet (20%)</td></tr>
<tr><td rowspan="4">Category-4 (Synthetic Data Generated with Real Data)</td><td>Dry_Synthetic</td><td>Dry_Real</td></tr>
<tr><td>Wet_Synthetic</td><td>Wet_Real</td></tr>
<tr><td>(Dry +Wet) Synthetic (80%)</td><td>(Dry +Wet) Synthetic (20%)</td></tr>
<tr><td>(Dry +Wet) Synthetic</td><td>(Dry+ Wet) Real</td></tr>
</table>

The importance of time series data for enhancing the performance of AI-based machine learning models in WWTP modeling studies cannot be overstated, as these data are integral for addressing significant disruptions within a WWTP. In line with this rationale, synthetic datasets consisting of 1000 samples were generated using 89 real data points for additional validation, of which 500 samples originated from 57 samples of dry period data and the remaining 500 samples developed from 32 of the wet period datasets in Category 4. To determine the extra robustness of the proposed models, these artificial datasets were subjected to several assessments. The study compared actual data from dry periods with synthetic data generated using that same dry period data. Similarly, synthetic data generated from actual wet period data was compared to real wet period data. Afterward, the combined synthetic data was split into two parts, with 80% allocated for training and 20% for testing. Finally, the blended synthetic data was assessed in relation to a real sample of 89 datasets.

3.10 Model Performance Visualization

The model can be evaluated using straightforward charts, even though the evaluation criteria are quite helpful. The model's predictions and the actual data will be compared in these charts. It indicates whether data are under- or over-fitted and show how well the model performed on both training and test data sets. This study uses a few standard visualization approaches, including violin plots of synthetic data, scatter plots of predicted and observed values, etc.

3.11 Model Performance Evaluation Criteria

After a model structure has been picked and the network has been trained, the chosen model needs to be examined. The strength of the fit between a model's outputs and the system that provided the same input actually determines how accurate a model is. For this, various validation tests must be taken into account. In most cases, a model's accuracy needs to be assessed for two sets of data samples. These data sets are the testing data set, which evaluates the network's capacity for generalization, and the training data set, which expresses learning effectiveness. The validation data set can occasionally be utilized to prevent overfitting issues in the model. It is important to note that the testing data set should ideally not have been presented before.

In this study, consideration has been given to the specified evaluation criteria. This allowed to check every model to be used practically. R^2, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) calculations were made to assess each model's BOD_5, COD, and TSS prediction performance.

- R^2, often referred to as the coefficient of determination in statistics, measures how well a regression model's independent variables consider the variance in the dependent variable. A higher R^2 indicates that a more significant proportion of the variance is explained, while a lower R^2 suggests that the model does not explain much of the variance, and there may be unaccounted factors influencing the dependent variable. R^2 typically ranges from 0 to 1, with 1 being the model's perfect fit, which completely explains all variance. However, negative values are possible if the model poorly fits the data. So, higher R^2 value means the model may effectively make predictions based on the given independent variables. The R^2 value of each model is calculated using Eq. (3.12).

$$R^2 = 1 - \frac{SSR}{SST} \qquad \text{Eq. (3.12)}$$

Where, SSR (Sum of Squared Residuals) = the sum of the squared differences between the predicted values of the regression model and the actual values of the dependent variable, and SST (Total Sum of Squares) = the sum of the squared differences between the dependent variable's actual values and its mean.

- RMSE is a measure of the spread or dispersion of the errors between the predicted and actual values. RMSE clearly indicates how well a predictive model's forecasts align with the actual data. Smaller RMSE values indicate better model performance, while larger RMSE values signify less accurate predictions. The RMSE value of each model is calculated using Eq. (3.13).

$$RMSE = \sqrt{\overline{(f - o)^2}} \qquad \text{Eq. (3.13)}$$

Where, f = forecast value, and o = actual value.

- In statistics and machine learning, MAE quantifies how far off predictions are from the true values on average, without considering the direction of the errors (i.e., whether they are overestimations or underestimations). Similar to RMSE, the lower the MAE, the better the model predicts. The MAE value of each model is calculated using Eq. (3.14).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|o_i - f_i| \qquad \text{Eq. (3.14)}$$

Where, n is the total number of data points, o_i represents the actual (observed) value for the i^{th} data point, f_i represents the predicted value for the i^{th} data point and | | represents the absolute value.

Chapter 4: RESULTS AND DISCUSSIONS

4.1 General

The detailed findings have been organized for discussion with previously published literature in accordance with the study's aims and the approaches taken to accomplish those objectives. Variations and comparisons are provided systematically. It has been described how to model an activated sludge system utilizing various machine learning and deep learning tools in order to predict BOD_5, COD, and TSS, assuming that biological treatments' would take care of the reduction of BOD_5, COD, and TSS values in effluent, addressing the influent's BOD_5, COD, and TSS variability, by seasonal variations in wastewater quality.

4.2 Capability of AI Based Model to Generate Wastewater Synthetic Data

The formulae and assumptions mentioned in Section 3.4.1 were used to generate almost 1,50,000 datasets of different variables of wastewater related to ASS. The Python module CTGANSynthesizer, which uses a conditional generative adversarial network (GAN)-based deep learning technique, was used to generate this synthetic tabular data. The generated dataset was visualized using Violin Plot (Fig 4.1). The violin plot is an excellent data visualization tool that assists in understanding and developing new insights by plotting a large amount of data of various variables on a single figure representing the distribution. Similar to a box plot, but with a rotated plot on each side that provides more details about the estimated density on the y-axis. A violin-like image is produced by mirroring and flipping the density, then filling in the resulting shape. The benefit of a violin plot is that it can reveal characteristics in the distribution that a boxplot cannot. This plots display trends and patterens of synthetic data of different parameters.

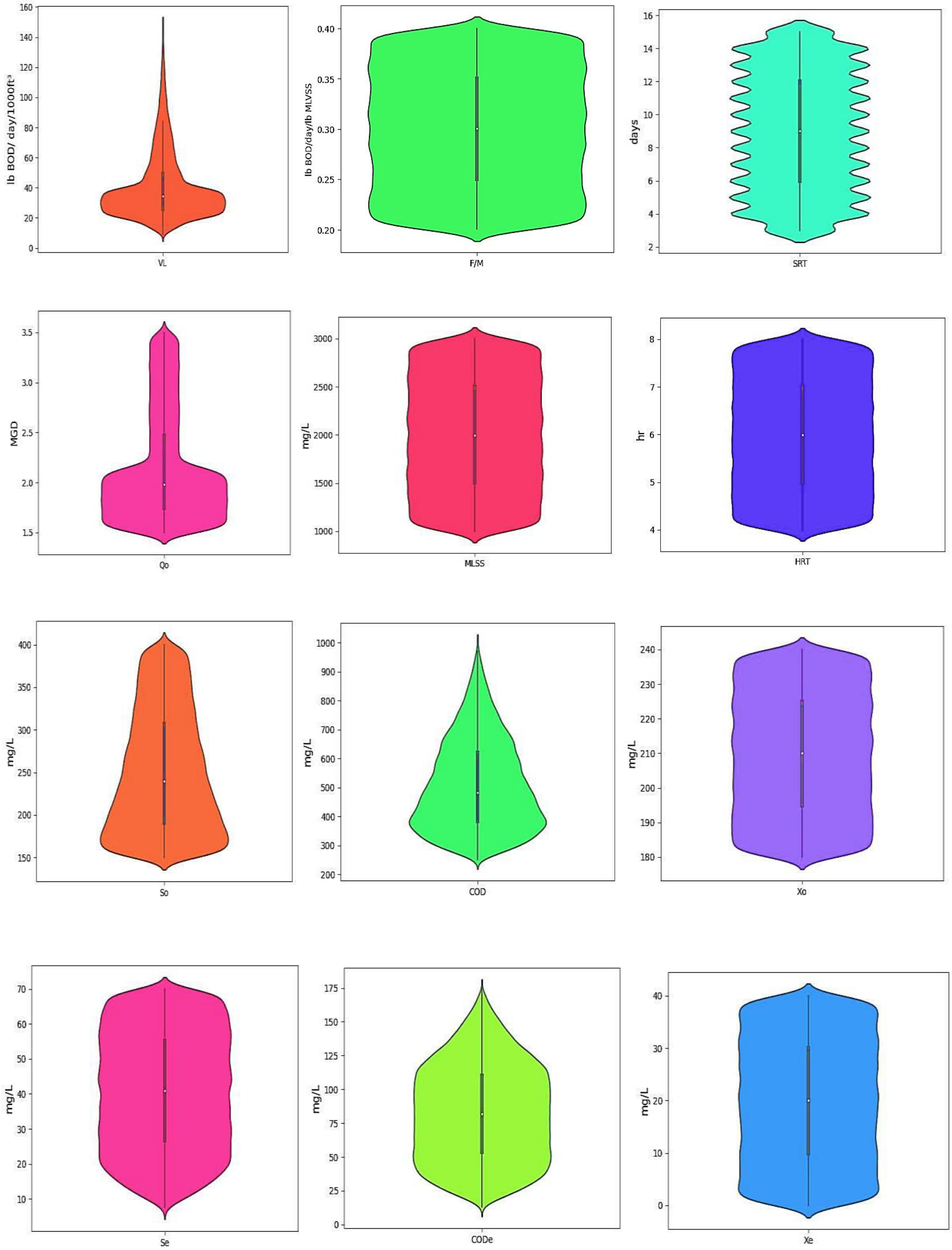

Fig. 4.1 Violin plot of the generated dataset

Four elements make up a violin plot. In Fig. 4.1, the distribution's median is indicated by a white dot with a white center in the plot's center. The distribution's quartile range is shown inside the plot as a narrow bar. The remaining part of the distribution is represented by a long, thin line extending from the bar and is determined by the formulas Q_1-1.5 IQR for the lower range and Q_3+1.5 IQR for the upper range. The points that stay outside of this line are regarded as outliers. The boundary of the violin plot reflecting the distribution of data points is defined by a line dividing the plot portraying the Kernel Density Estimation (KDE) plot. Here, using Univariate Analysis, violin plots are utilized to show the distribution of each variable. These plots, which often incorporate a kernel density plot and a mirrored histogram, show the density estimation of the variable's values. The density of data points at various values is represented by the width of the violin, with broader sections denoting more density and skinnier sections representing lower concentrations of data points.

The synthetic data patterns exhibit irregularity, as seen in Fig. 4.1, which is expected due to the complex nature of wastewater composition. Moreover, certain variables are influenced by operational factors like Q_w, S_e, VL, and Q_r. These values change based on influent and plant operation, so their variability is not surprising. Upon examining Fig. 4.1, it is apparent that the AI-generated SRT values, for instance, have a range of 3 to 15 and exhibit a symmetric distribution with a uniform density but fluctuation in the data represented in the KDE plot, indicating a strong resemblance to the real field. In contrast, variables like FM and HRT have symmetric distributions; however, the density of VL and Q_o appears elongated and non-uniform in the KDE plot, suggesting a skewed distribution that aligns well with real field variations. This provides guidance for selecting suitable ranges of different parameters for practical considerations, such as 0.2 to 0.4 for FM, 4 to 8 for HRT, and so on.

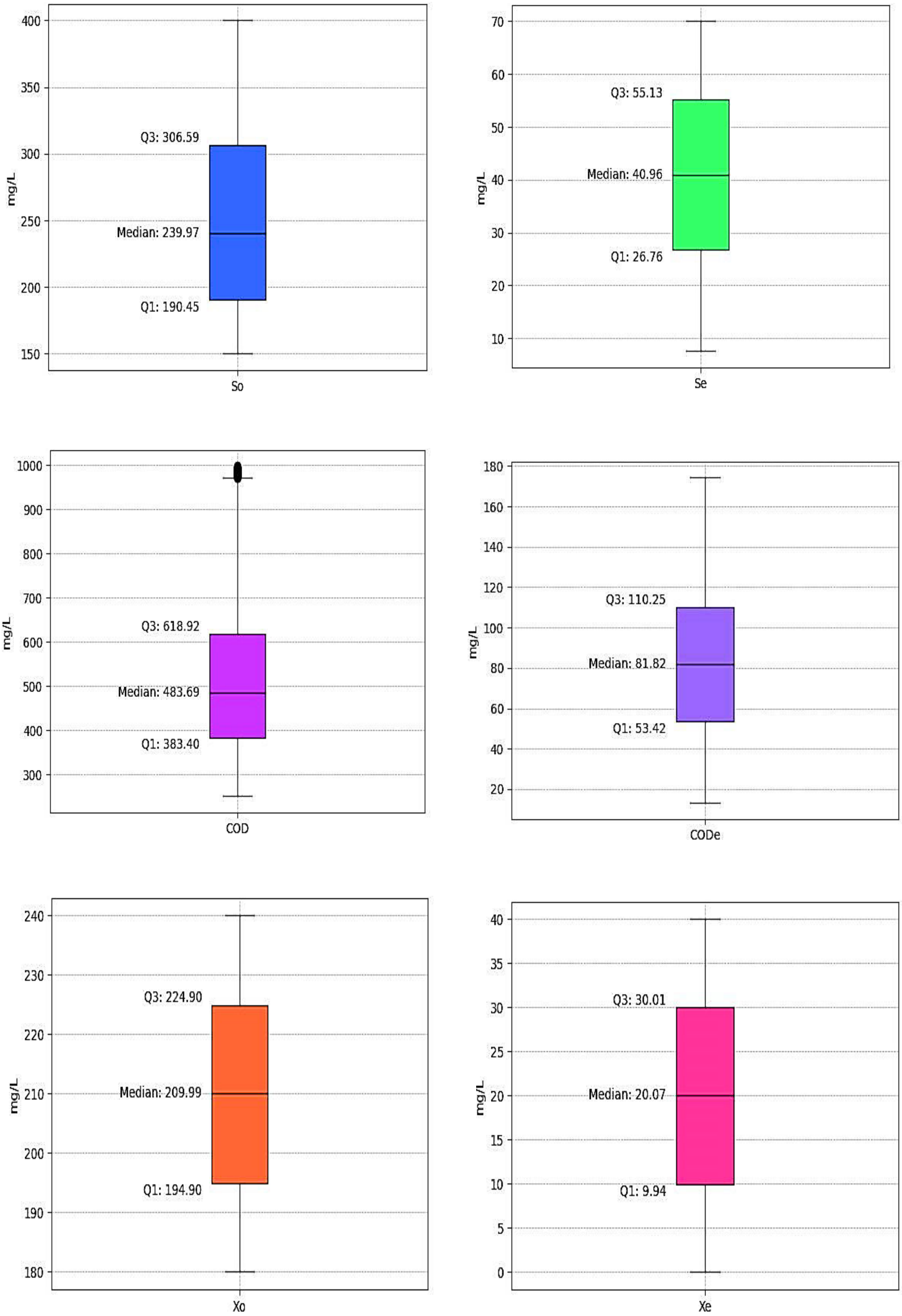

Fig. 4.2 Box plot of the generated dataset

Again, box plots are an essential tool for exploratory data analysis because they facilitate rapid comprehension of a dataset's key properties and the ability of analysts, researchers, and decision-makers to draw well-informed conclusions. They offer an in-depth understanding of the data distribution and serve as a helpful supplement to other statistical metrics and graphical representations. Box plots provide a straightforward and easily understood way of visualizing the distribution of the data set illuminating central tendency, spread, and skewness of the data. The symmetric nature of a distribution is evident when the median aligns with the middle of the box and the whiskers are symmetrical. Positive skewness (skewed right) is observed when the median is closer to the bottom of the box, with shorter whiskers on the lower end. Conversely, negative skewness (skewed left) is indicated by a median closer to the top of the box, accompanied by shorter whiskers on the upper end.

Illustrating this with a specific example, in Fig. 4.2, the boxplot for variable S_o reveals a median of 239.97 mg/L. This signifies that half of the data values are below 239.97, while the other half are above this value. The first quartile is 190.45 mg/L, and the third quartile is 306.59 mg/L, resulting in an interquartile range (IQR) of 116.14 mg/L, covering the middle 50% of the data. The whiskers extend from the box to the minimum and maximum values, which are 150 mg/L and 400 mg/L, respectively. This dataset is notable for having no outliers. Overall, the distribution of this dataset is right-skewed, indicating a predominant concentration toward higher values. This observation aligns with the patterns depicted in the Violin plot presented in Fig. 4.1. Furthermore, synthetic data and the capabilities of AI-based modeling have been assessed using various statistical parameters, as discussed in section 4.2.1.

4.2.1 Basic Statistics of Synthetic Data

Furthermore, Table 4.1 listed the basic statistics of the variables for the estimation of the minimum, maximum, mean, standard deviation, coefficient of variation, skewness and kurtosis values for the generated dataset for insight pattern and association.

A higher coefficient of variation (CV) implies a more diverse distribution, while a lower CV suggests a more consistent distribution. Table 4.1 reveals a range of CV values from 8.2 to 57.8, highlighting substantial variability in the dataset concerning various wastewater parameters. The highest CV for X_e is 57.8%, showing a significant level of variability, with the standard deviation being 57.8% of the mean. This suggests that the variability in this parameter is largely influenced by both the quantity of recycling and the quality of the effluent produced. On the other hand, the lower CV value of 8.2% for X_o (primary effluent TSS), with the standard deviation being only 8.2% of the mean, reflects a consistency in effluent production driven by effective plant operations to meet specific targets. Similar relative patterns in relation to CV are observed for other wastewater parameters.

A statistical metric known as skewness evaluates the asymmetry of a probability distribution. It measures how much the data is moved to one side or skewed. Skewness aids in understanding a dataset's form and outliers. Depending on the model, skewness may violate model assumptions or make it more difficult to understand the significance of a particular independent variable (feature) when its values are skewed.

Positive skewness indicates that the distribution's tail is longer on the right side. Extremely positive skewness is undesirable in a distribution because it might lead to inaccurate results. Mean > Median > Mode is the basis for positive skewness. The skewness is negative when the left side of the distribution's tail is longer than the right side's tail. Mode > Median > Mean is a necessary condition for negative skewness. Another name for a "symmetric distribution" is zero skewness. It denotes that the data distribution is symmetrical around the mean and devoid of large tails at either end. Mean = Mode = Median is the condition for zero skewness.

Table 4.1. Statistical description of the generated synthetic dataset

Parameter	Unit	Min	Max	Mean	SD	CV (%)	Skewness	Kurtosis
VL	lb BOD/ day/1000ft^3	9.1	148.8	41.1	22.5	54.7	1.54	2.3
F/M	lb BOD/day/lb MLVSS	0.2	0.4	0.3	0.06	19.3	0.006	-1.21
HRT	hr	4	8	6	1.16	19.3	0.006	-1.2
Q_o	MGD	1.5	3.5	2.15	0.54	25.2	0.96	-0.21
X (MLSS)	mg/L	1000	3000	2000	575	28.8	-0.003	-1.2
Q_r	MGD	0.4	2.5	1.45	0.61	41.8	-0.004	-1.2
X_w	mg/L	5000	10000	7498	1443	19.2	-0.001	-1.12
SRT	days	3	15	9	3.5	38.8	0.006	-1.17
Q_{ratio}	---	0.25	0.75	0.5	0.14	28.9	-0.005	-1.2
MLVSS	mg/L	700	2400	1500	436	29	0.03	-1.14
S_o	mg/L	150	400	252	70	27.9	0.41	-0.99
S_e	mg/L	7.6	70	41	16.6	40.6	-0.023	-1.15
COD	mg/L	250	997	510	155	30.4	0.58	-0.42
COD_e	mg/L	13	174	83	35.3	42.6	0.17	-0.87
X_o	mg/L	180	240	210	17.3	8.2	0.004	-1.2
X_e	mg/L	0.0005	40	20	11.6	57.8	-0.005	-1.2

A good rule of thumb is that the data are almost symmetrical if the skewness is between -0.5 and 0.5. The data are slightly skewed if the skewness is between -1 and -0.5 (negative skewed) or between 0.5 and 1 (positive skewed). The data are considered to be highly skewed if the skewness is less than -1 (negative skewed) or higher than 1 (positive skewed). Greater skewness in the data reflects a wider disparity, while lower skewness indicates a more consistent distribution.

Table 4.1 reveals that certain output parameters, such as COD, COD_e, and X_o, exhibit positive skewness. In contrast, the output parameters S_e (effluent BOD) and X_e demonstrate negative skewness, with values of -0.023 and -0.005, respectively. The remaining wastewater parameters show minimal skewness, except for Q_o, which has a skewness value of 0.96. VL stands out with a highly positive skewness of 1.54, indicating a long tail on the right side of the distribution, signifying the presence of outliers or extreme values that pull the mean to the right. This is also evident in the graphical distribution with the violin plot in Fig. 4.1, which appears stretched or skewed towards the right, with the majority of data points clustered on the left. This observation also aligns with the coefficient of variation (CV), which is greatly affected by the amount of recycling as well as the effluent's quality.

Another statistical term known as kurtosis characterizes a probability distribution's form. Compared to a normal distribution, it reveals information about its tails and peaks. While negative kurtosis suggests lighter tails and a flatter distribution, positive kurtosis implies heavier tails and a more peaked distribution. Kurtosis aids in the analysis of a dataset's properties and outliers. Since the kurtosis of normal distributions is 3, the excess kurtosis is computed by deducting the kurtosis by 3.

Long and thick tails are characteristic of leptokurtic (Kurtosis > 3), which increases the possibility of outliers. Positive values of kurtosis suggest a peaked distribution with thick tails. Extremely positive kurtosis denotes a distribution in which more data points are distributed away from the mean and toward the tails. Since the platykurtic distribution (Kurtosis < 3) has a thin tail and is spread outward from the center, the majority of the data points are concentrated close to the mean. The platykurtic distribution is flatter (less peaked) than the normal distribution. Kurtosis is close to zero when a distribution is mesokurtic (Kurtosis = 3), which is the same as the normal distribution. Mesokurtic has curves with medium-sized peaks and moderate breadth distribution.

Kurtosis is used to assess the shape of a distribution, specifically whether it is more peaked or flatter than a normal distribution and if it contains more or fewer extreme values (outliers). As indicated in Table 4.1, VL displays a markedly elevated positive kurtosis (platykurtic) at 2.3. This indicates that the datasets associated with VL are more widely dispersed and possess flatter distributions, resulting in a diminished likelihood of extreme values or outliers. This suggests that VL occasionally features values that are exceptionally high or low, influenced by both the amount of recycling and the quality of the effluent produced. This characteristic is also apparent in the violin plot presented in Fig. 4.1, illustrating a peaked distribution with heavy tails, thereby confirming the coefficient of variation (CV) and skewness values. Conversely, the distributions of other wastewater parameters show negative (platykurtic) kurtosis values.

In summary, the skewness and kurtosis of data have implications for the planning of wastewater treatment facilities and the strategies employed to control processes. When dealing with data that is either positively or negatively skewed, operators must ensure that the plant or system can effectively handle occasional spikes or drops in wastewater parameter values. Additionally, when kurtosis is significantly positive, it serves as an indicator of outliers or extreme data points in the dataset, highlighting unusual events or problems within the wastewater system. This information can help in identifying potential issues with wastewater treatment facilities and allow for necessary corrective measures to be taken.

4.2.2 Correlation of Synthetic Data

A correlation heatmap is a visual tool that illustrates the relationships between various parameters in wastewater samples by displaying them as a color-coded matrix based on their correlations. In this heatmap, each parameter is represented both as a row and a column, and the cells in the matrix reveal the strength and direction of the correlations between them. The color of each cell shows the degree of correlation; darker colors indicate greater correlations.

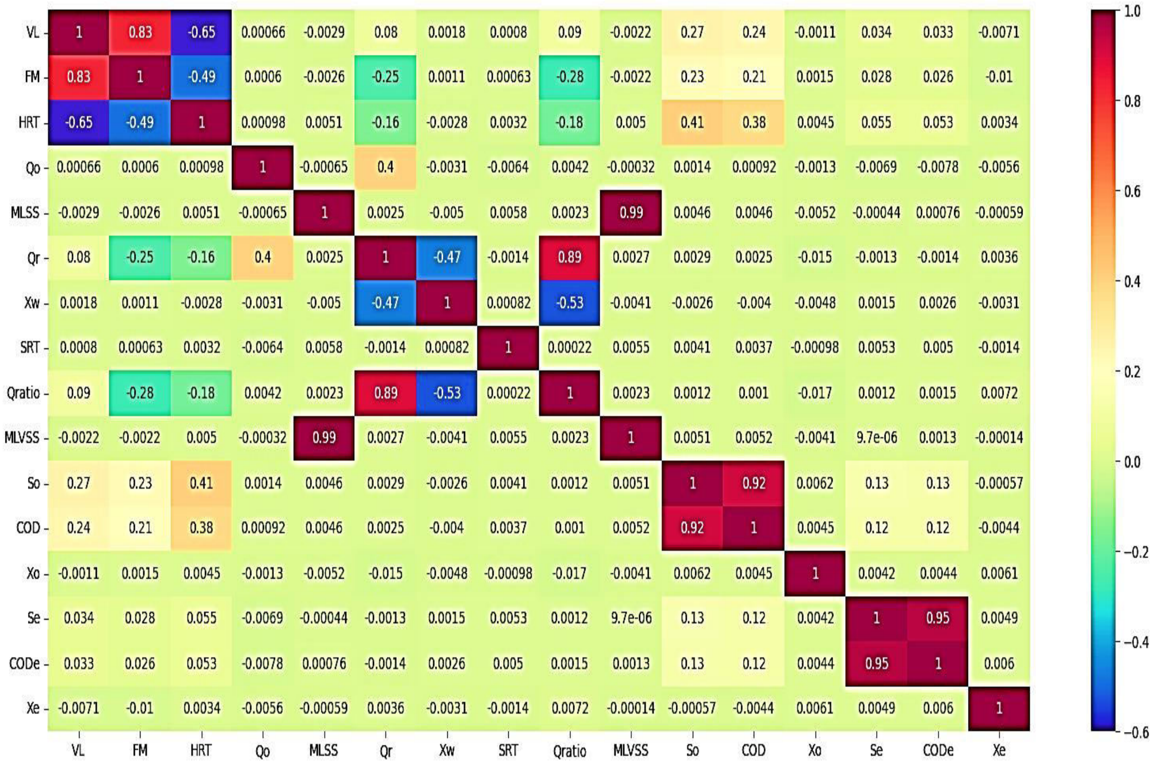

Fig. 4.3 Heatmap showing correlation among wastewater parameters

In Fig. 4.3, a correlation heatmap generated using Pearson's correlation coefficient was created from a dataset comprising 94,892 instances, each described by 18 parameters. A linear relationship between two variables is quantified by Pearson's correlation coefficient, which scales from -1 to 1, indicating perfect negative correlation and perfect positive correlation, respectively, and 0, denoting no correlation. To create this heatmap, the Python data visualization library Seaborn, which is built on Matplotlib, was employed.

The correlation coefficients were converted into a matrix, comparing each parameter to every other parameter. This matrix is symmetric, with diagonal values set to 1 (as each parameter correlates perfectly with itself). Off-diagonal values represent the correlations between parameters. The color of each cell signifies the strength and direction of the correlation.

Typically, color scales transition from one color to another, such as a gradient from blue (indicating negative correlation) to red (showing positive correlation), with white representing no correlation. Strong positive correlations (closer to 1) are depicted with warm colors like dark red or orange, suggesting that as one parameter increases, the other tends to increase as well. Strong negative correlations (closer to -1) are shown with dark blue or green colors, suggesting that as one parameter increases, the other tends to decrease. Weak or no correlations (closer to 0) are represented by white or pale cells, indicating a lack of a meaningful relationship between the parameters.

In this study, as observed in Fig. 4.3, the correlation coefficient values were found to be significantly higher than those in other studies conducted elsewhere. This was expected due to the variations in wastewater composition. For instance, a strong positive correlation ($r = 0.92$) was observed between COD and BOD_5 in this synthetic data, depicted as a dark red cell in the heatmap. Similarly, a positive correlation ($r = 0.83$) was found between F/M and VL, represented by an orange cell. Additionally, there were medium negative correlations between Q_w and SRT ($r = -0.52$) and F/M ($r = -0.39$), as indicated by blue cells, and a positive correlation between Q_w and HRT ($r = 0.48$), represented by a light orange cell.

Notably, the values reported here were notably higher, even though 18 parameters were considered, in contrast to the findings from daily wastewater data of 312 records encompassing eight parameters (pH, BOD_5, COD, TSS, TN, TP, Temperature, and Conductivity) at the Nicosia wastewater treatment plant in North Cyprus, Turkey, for the years 2014 to 2016, where the reported correlation value between BOD_5 and COD was only 0.1793 (Elkiran & Abba, 2017).

The heatmap helps identify which wastewater parameters tend to move in the same or opposite directions, providing insights into the complex interactions within the wastewater system. This information is valuable for making decisions related to wastewater treatment, quality control, optimizing processes, controlling pollution, and monitoring environmental conditions.

Based on a thorough examination of the generated synthetic wastewater data in relation to the ASS system in terms of coefficient of variation, skewness, kurtosis, and heatmap, it can be concluded that, in the absence of real data, it is possible to create synthetic data that closely resembles the operations and attributes of the Wastewater Treatment Plant (WWTP) by considering the specific local conditions.

4.3 Category 1: Synthetic Data Associated with Mathematical Equation and Assumptions

The availability of time series data is essential for the performance of WWTP modeling studies in the AI-based machine learning process because they are critical disruptions in a WWTP. Such real time series data are not accessible in Bangladesh. After a rigorous search and analysis of the literature, it is found that there is no reliable online dataset containing extra instances with different parameter values collected from WWTP. Even though specific datasets are accessible online, they have many restrictions, such as, for example, fewer wastewater parameters and fewer parameter relationships. Again, acquiring sufficiently lengthy and high-quality time series data has grown more complicated and challenging in terms of security. Artificial intelligence (AI) based tools have been used to create synthetic data in this line. The category-1 synthetic dataset made up of mathematical equations and assumptions was split into two sets, the training set making up 80% and the testing set 20% of the total data, respectively. To create output values reflecting BOD_5, COD, & TSS values, the model required input from 18 parameters (VL, V, F/M, HRT, Q_o, X, MLVSS, Q_r, X_w, SRT, Q_w, Q_{ratio}, X_o, X_e, S_o, S_e, COD and COD_e).

Table 4.2. Category 1 (Synthetic data associated with mathematical equation and assumptions): Performance evaluation of ML and DL for wastewater effluent (BOD_5, COD and TSS)

Model	R^2			RMSE			MAE		
	BOD_5	**COD**	**TSS**	**BOD_5**	**COD**	**TSS**	**BOD_5**	**COD**	**TSS**
Random Forest	0.93	0.91	0.60	0.072	0.066	0.182	0.058	0.053	0.14
Decision Tree	0.85	0.82	0.12	0.101	0.092	0.272	0.077	0.07	0.202
Extra Trees	0.93	0.91	0.57	0.072	0.067	0.189	0.058	0.053	0.147
Multivariate Linear Regression	0.91	0.91	0.05	0.079	0.065	0.282	0.064	0.052	0.243
K-Neighbors Regression	0.85	0.85	-0.19	0.102	0.083	0.315	0.082	0.066	0.267
Gradient Boosting Regressor	0.93	0.91	0.33	0.071	0.065	0.236	0.058	0.052	0.196
Adaboost Regressor	0.92	0.90	0.11	0.074	0.067	0.272	0.061	0.055	0.232
ANN-I	0.92	0.91	0.93	0.077	0.066	0.077	0.062	0.053	0.051
ANN-II	0.92	0.90	0.90	0.073	0.067	0.089	0.06	0.054	0.025
ANN-III	0.92	0.91	0.94	0.073	0.066	0.068	0.059	0.052	0.051

Optimized performance metrics for different AI-based models that predict BOD_5, COD, and TSS in an activated sludge system are provided in Table 4.2. The outcomes of various trials can be found in Tables B1–B3 of Appendix B. R^2, RMSE, and MAE are some of the measures. The models were assessed, which included three different Artificial Neural Networks (ANNs) and seven machine learning models (Gradient Boosting Regressor, Adaboost Regressor, Multivariate Linear Regression, Decision Tree, Extra Trees, and K-Neighbors Regression).

The measurements reveal that multiple models accurately predict BOD_5, COD, and TSS, as seen by strong R^2 scores (close to 1), low RMSE, and low MAE values. Notably, the three ANNs (ANN-I, ANN-II, and ANN-III) appeared exceptionally good. The variations of these three ANN structures were based on the optimizer and activation functions, as in Table 3.4. Three different layer types are commonly present in ANNs: the input layer, one or more hidden layers, and the output layer. During training, the weights on the connections between neurons are changed, allowing the network to recognize and adapt to data patterns. The weights of ANNs were updated using optimization techniques like gradient descent to minimize the difference between the predictions and the actual labels after feeding it with labeled data (input-output pairs) and changing the number of both hidden layers and their nodes to perform at their best. This process is controlled by a loss function, which calculates the gap between predictions and actual results. Based on the accuracy of the predictions, this study employed a trial-and-error method to determine the optimum number of hidden layers and nodes. The attached appendices show how altering the hidden layers and the neurons within each layer affect the performance of the models. Eighteen neurons in the input layer for all ANNs were found to generate the best results after numerous tries. Again, optimum ANN architectures had three hidden layers with different neurons for each layer, and it emerged that the 18-32-64-32-1, 18-64-124-64-1, and 18-32-64-32-1 were the most effective for ANN-I, ANN-II, and ANN-III respectively.

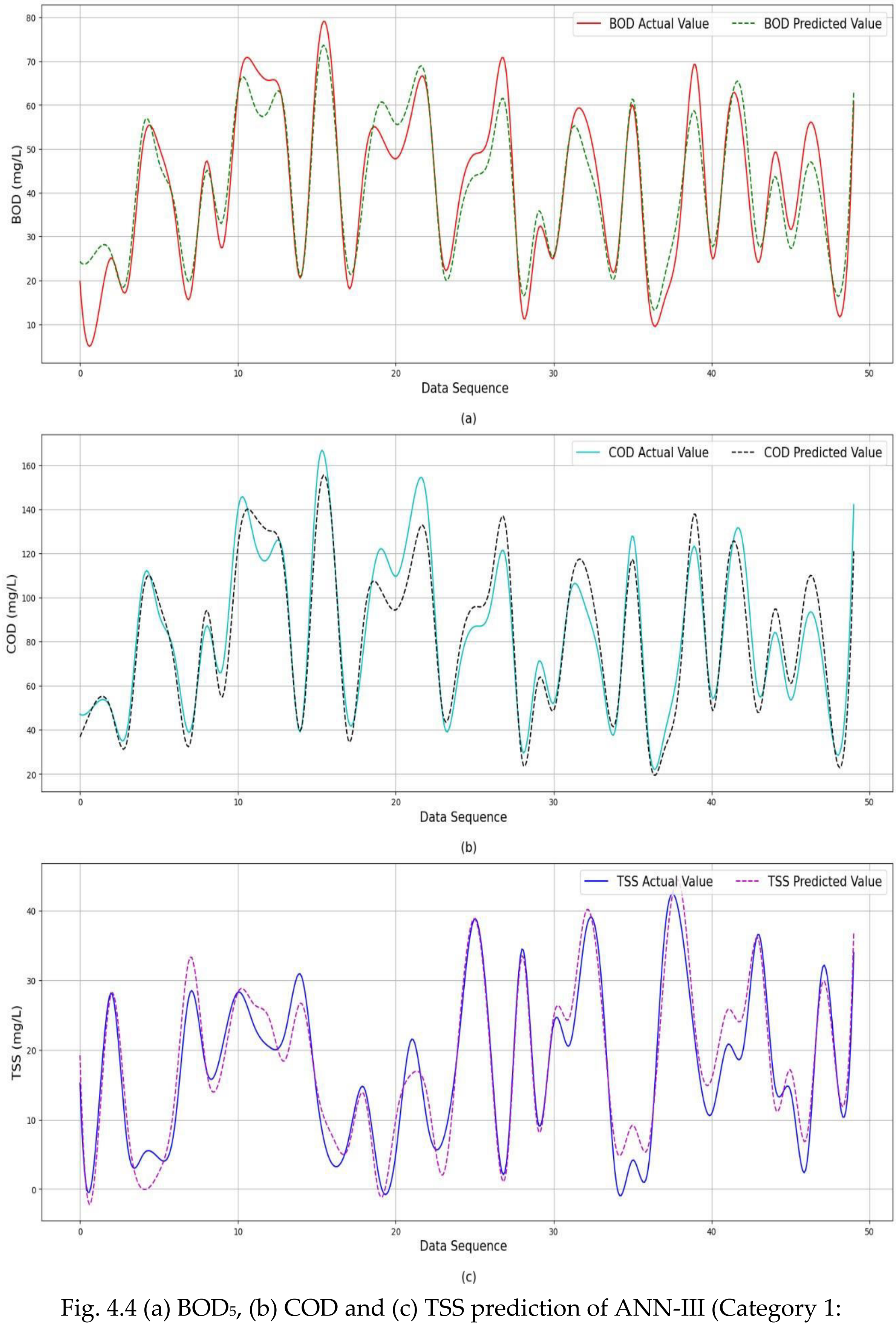

Fig. 4.4 (a) BOD_5, (b) COD and (c) TSS prediction of ANN-III (Category 1: Synthetic data associated with mathematical equation and assumptions)

The three ANNs deliver commendable results, with high R^2 values, with particular prominence in three outputs (BOD_5, COD & TSS). Their relatively low RMSE and MAE values suggest precise predictive accuracy. ANNs are renowned for their capability to capture intricate non-linear data relationships, which is particularly advantageous in this context. The closest agreement between predicted and actual values for BOD_5, COD, and TSS is also shown in Fig. 4.4. It is to be noted that for better representation, the first few data points were used for graphical representation. The R^2, RMSE, and MAE values reported here are consistent with the research conducted by (Zhao et al., 2016), demonstrating their applicability in effluent prediction capabilities even in the context of Bangladesh, where long-term time-series data are not yet available. Additionally, due to the variations in input parameters and the quantity of data, the agreement of predicted outcomes is shown to be slightly higher and, in some cases, somewhat lower than a few studies reported elsewhere.

However, this variation cannot be overlooked but rather discussed, considering that the wastewater quality is found to vary widely on different scales and geographic locations. Moreover, this study contributes to the field by incorporating a more extensive set of input variables, surpassing the scope of previous research. This expanded input dataset enhances the comprehensiveness of these models and strengthens their predictive capabilities.

Ensemble methods such as Random Forest and Extra Trees often yield strong performance across various tasks. They exhibit high R^2 values, attaining an R^2 of greater than 0.9 in the case of BOD_5 and COD, signifying robust correlations between the predicted and actual values, and they display low RMSE and MAE values. But in the case of TSS prediction, the performance of ML models is far behind that of ANN models, as clearly observed in Fig. 4.5. To produce a more accurate illustration, only a portion of the data points were included. These outcomes suggest their effective capture of underlying data patterns and relationships. However, their performance in predicting TSS lags behind even negative in the case of K-Neighbors Regression that of ANNs which is clearly depicted in Table 4.2, showing less neighborhood between predicted and actual values. The values, however, are in line with other studies elsewhere (Wang et al., 2021).

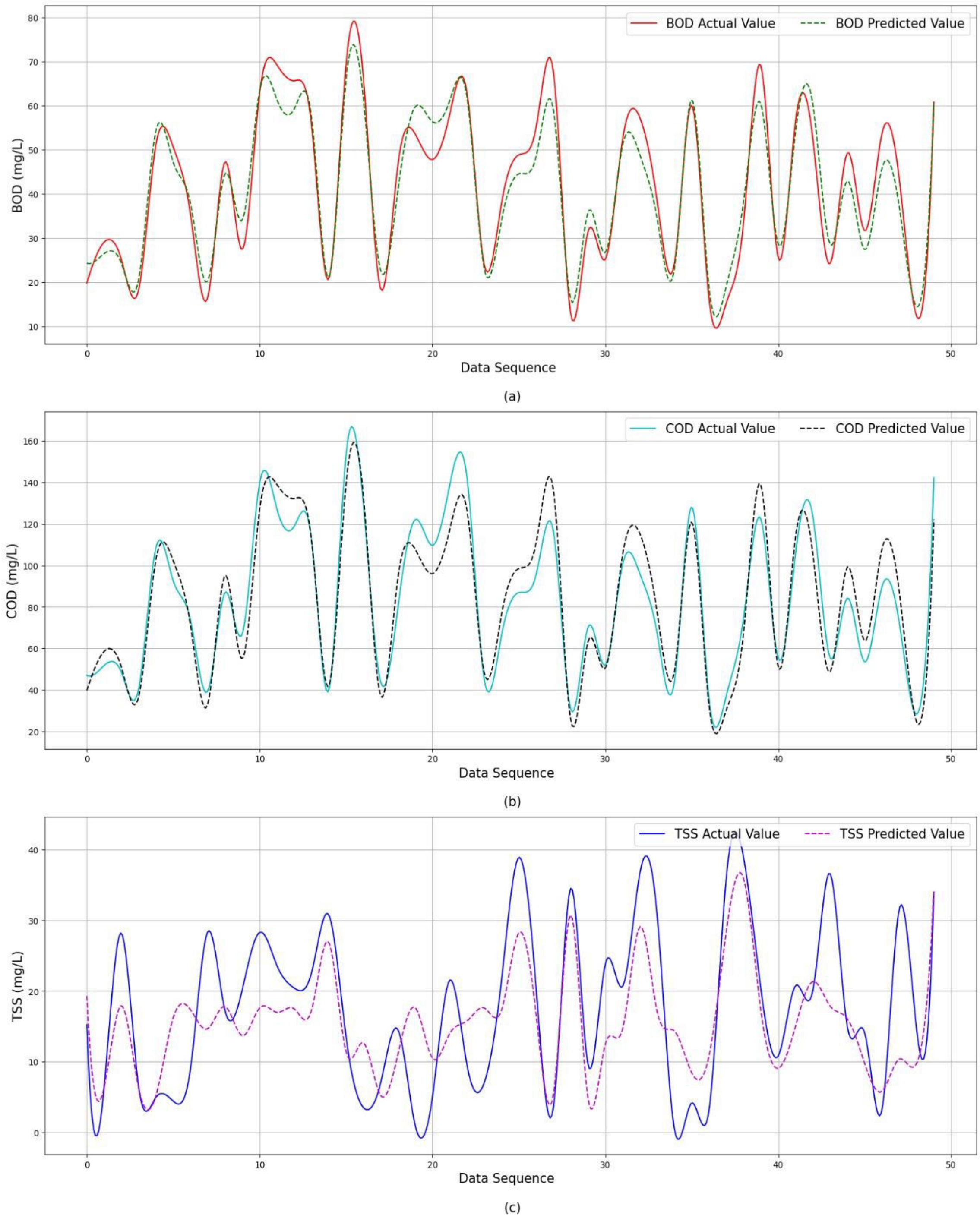

Fig. 4.5 (a) BOD_5, (b) COD and (c) TSS prediction of Random Forest (Category 1: Synthetic data associated with mathematical equation and assumptions)

An excellent general model for projecting BOD_5, COD, and TSS in the activated sludge system can be determined depending on the project's needs. To make a final choice, issues including model training and inference times, resource requirements, and the accessibility of supplemental data should be taken into account. Which models offer the most insightful understanding of the system can be determined by additional analysis, such as ratings of feature relevance and partial dependence plots. Additionally, if available, the validation of the models using extra test data might improve the overall model selection process. However, based on basic statistics (R^2, RMSE and MAE), the ANNs emerge as strong contenders if capturing complicated non-linear data correlations is crucial.

4.4 Category 2: Real Wastewater Data against Generated Synthetic Data

In Category 2, samples of 89 were taken from septic tanks, pits, and residential area drains throughout Chittagong to determine whether various models built using the 18 operational and qualitative parameters of the ASS system could accurately represent the situation. For this, only the BOD_5, COD, and TSS values from synthetic and real data are employed. The outcomes demonstrated in Table 4.3 show how well different models predict BOD_5, COD, and TSS levels in an activated sludge system. The results from different trials are available in Tables C1–C3 of Appendix C. R^2, RMSE, and MAE are three evaluation criteria used to assess these models.

Several machine learning models display significantly high values for R^2, according to the data. The R^2 values of the following models: Random Forest, Extra Trees, Gradient Boosting Regressor, and Adaboost Regressor are all greater than 0.92, demonstrating an excellent capacity to capture the variance in the target variables. Random Forest, Extra Trees, and Gradient Boosting Regressor all exhibit a comparable pattern in terms of RMSE and MAE, with consistently low RMSE and MAE values for the three target variables (BOD_5, COD, and TSS). This shows that the predictions made by these models are pretty accurate. As shown in Fig. 4.6, BOD_5, COD, and TSS had the best agreement between predicted and actual values. The R^2, RMSE, and MAE values presented here agree with the findings of other research, proving their relevance in effluent prediction abilities even in the geographical area of Bangladesh, where long-term time-series data are not yet available (Baki et al., 2019b).

Table 4.3. Category 2 (Real wastewater data against synthetic data): Performance evaluation of ML and DL for wastewater effluent (BOD_5, COD and TSS)

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
Random Forest	0.93	0.91	0.98	0.072	0.066	0.087	0.058	0.053	0.048
Decision Tree	0.85	0.82	0.74	0.101	0.092	0.098	0.077	0.07	0.033
Extra Trees	0.93	0.91	0.74	0.072	0.067	0.083	0.058	0.053	0.056
Multivariate Linear Regression	0.91	0.91	0.82	0.079	0.065	0.085	0.064	0.052	0.085
K-Neighbors Regression	0.85	0.85	0.81	0.102	0.083	0.092	0.082	0.066	0.059
Gradient Boosting Regressor	0.93	0.91	0.90	0.071	0.065	0.091	0.058	0.052	0.091
Adaboost Regressor	0.92	0.90	0.83	0.074	0.067	0.085	0.061	0.055	0.086
ANN-I	0.92	0.91	0.92	0.077	0.066	0.092	0.062	0.053	0.047
ANN-II	0.92	0.90	0.86	0.073	0.067	0.088	0.06	0.054	0.043
ANN-III	0.92	0.91	0.82	0.073	0.066	0.084	0.059	0.052	0.067

Additionally, each model performs better than other studies reported elsewhere because of the differences in input parameters and the volume of data. However, the abundance of data could be responsible for this variation. Furthermore, this study advances the area by using a limited number of factors from a collection of input variables, which have rarely been employed in earlier studies.

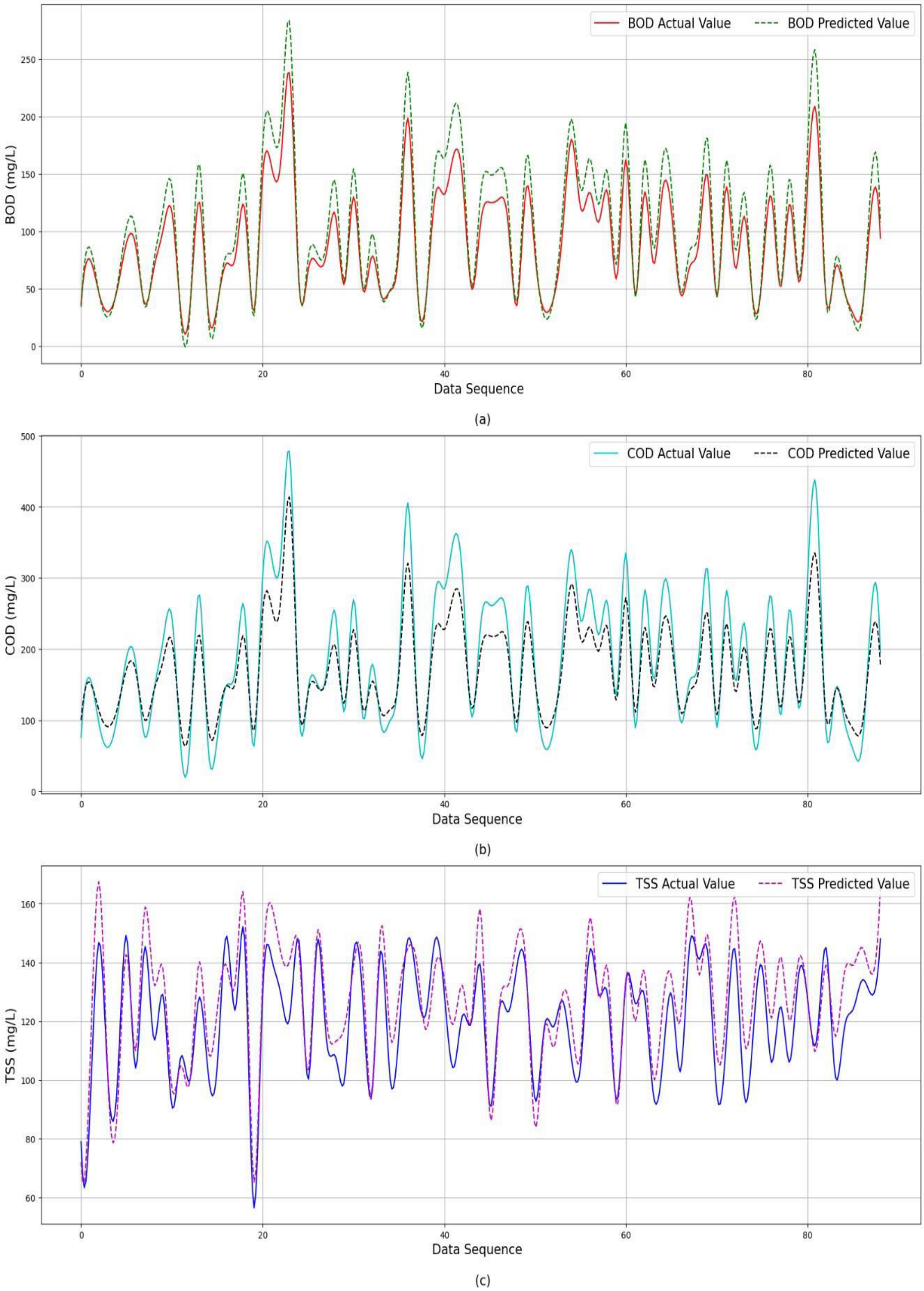

Fig. 4.6 (a) BOD_5, (b) COD and (c) TSS prediction of Random Forest (Category 2: Real wastewater data against synthetic data)

The three ANNs, ANN-I, ANN-II, and ANN-III, on the other hand, looked to be equally effective and produced praiseworthy results, with high R^2 values over 0.9 and special prominence in three outputs (BOD_5, COD, and TSS), with the exception of ANN-II and ANN-III in the case of TSS displaying below 0.9.

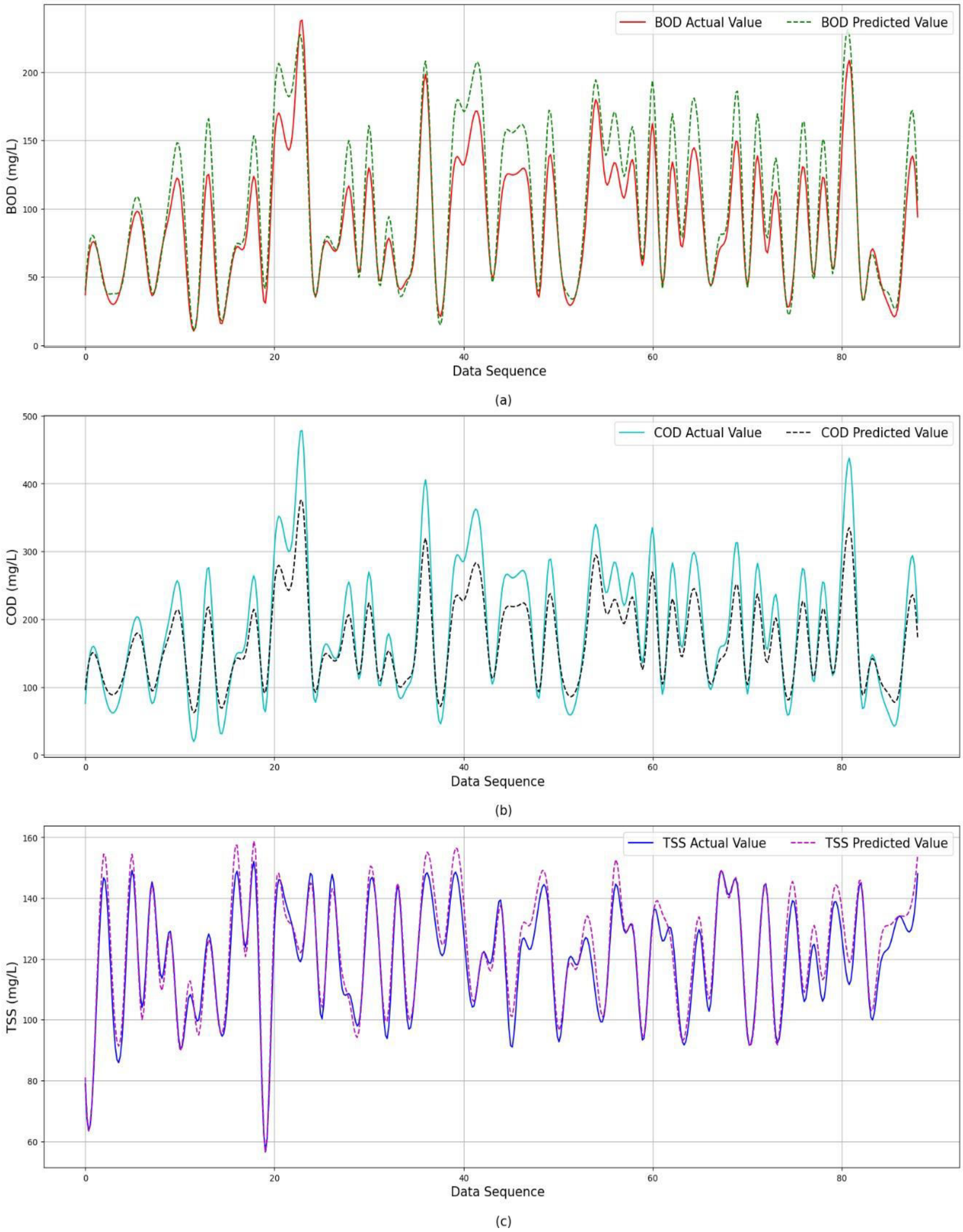

Fig. 4.7 (a) BOD_5, (b) COD and (c) TSS prediction of ANN-I (Category 2: Real wastewater data against synthetic data)

They have consistently low RMSE and MAE values for the three target variables (BOD_5, COD, and TSS). It is especially useful in this situation as ANNs are recognized for capturing complex non-linear data correlations. Fig. 4.7 also displays the BOD_5, COD, and TSS values where predicted and actual values were most closely aligned. The R^2, RMSE, and MAE values of this work are consistent with earlier studies, demonstrating their applicability to effluent prediction in Bangladesh (Alsulaili & Refaie, 2020). Due to input parameters, data volume, and the use of a small number of input variables, the model performs well.

Overall, combining these metrics reveals that while Random Forest, Extra Trees, and Gradient Boosting Regressor models are performing remarkably well in predicting BOD_5, COD, and TSS levels within the activated sludge system, exhibiting strong explanatory power (high R^2) and low prediction errors (low RMSE and MAE), the ANNs are also capable of capturing these variations and even outperform them if they were fed with modifications in their structural elements like epoch number and neurons variation. However, it is crucial to keep in mind that other practical factors, such as processing resources, model interpretability, and ease of implementation, may also influence the decision between these models.

4.5 Category 3: Seasonal Variation Based Performance

In Category 3, it was checked whether different models created using the 18 operational and qualitative ASS parameters could accurately represent the limited real data that had been collected from the various sites (grab samples) during both dry and wet seasons. This was done without the use of any mathematical equations that were connected to the operation of ASS. The improved results for the R^2, RMSE, and MAE evaluation criteria are shown in Table 4.4, Table 4.5, and Table 4.6, which illustrate how well various models predict BOD_5, COD, and TSS respectively, using the proposed machine learning and deep learning models under different dataset settings. The findings of BOD_5, COD, and TSS performance derived from diverse trials are presented in Tables D1–D15 of Appendix D. For training and testing purposes, the dataset of the dry period was split into 80% and 20%, respectively. The dataset for the wet period was likewise handled in a similar way. After that, dry period data was tested against wet period data, and vice versa. In the end, all data from the dry and wet periods were combined and divided into two parts, 80% of which were used for training and 20% for testing.

Random Forest, Decision Tree, Extra Trees, Multivariate Linear Regression, Gradient Boosting Regressor and Adaboost Regressor models consistently achieve excellent results, demonstrating robustness across all scenarios of BOD_5 (Table 4.4), COD (Table 4.5) and TSS (Table 4.6) showcasing high R^2 values exceeding 0.9, low RMSE, and low MAE, making them top-performing models overall. However, Decision Tree struggles in some scenarios like only dry condition (R^2 around 0.88) in the case of BOD_5, wet to dry condition (R^2 around 0.88) in the case of COD, particularly in the case of TSS (R^2 less than 0.8). K-Neighbors Regression performs relatively poorly compared to other models, especially in scenarios with Wet conditions in all pollutant cases (BOD_5, COD, TSS). It has lower R^2 values (sometimes less than 0.5) and higher RMSE and MAE values, indicating weaker predictive power. The predicted values for BOD_5, COD, and TSS showed the best correlation with the actual values, as shown in Fig. 4.8. Here, the predicted values are seen to be very close to the actual value, indicating that the BOD_5 value is being predicted very effectively by the Multivariate Linear Regression model. The plots in the COD and TSS cases are pretty similar. This study's results are consistent with those of earlier research, proving its adaptability to regional circumstances. Despite the little amount of data, each model performs almost identically to other studies published elsewhere (Zhao et al., 2016).

Table 4.4. Category 3 (Seasonal variation): Performance evaluation of ML and DL for wastewater effluent (BOD_5)

Model	Dry (80%) --- Dry (20%)			Wet (80%) --- Wet (80%)			Dry (100%) --- Wet (100%)			Wet (100%) ---- Dry (100%)			Dry+Wet (80%) --- Dry+Wet (20%)		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Random Forest	0.97	0.063	0.049	0.99	0.029	0.024	0.99	0.032	0.026	0.88	0.079	0.063	0.99	0.016	0.012
Decision Tree	0.89	0.116	0.078	0.95	0.068	0.049	0.97	0.045	0.031	0.93	0.063	0.053	0.97	0.03	0.019
Extra Trees	0.98	0.043	0.034	1	0.02	0.015	0.99	0.027	0.023	0.93	0.061	0.052	1	0.011	0.007
Multivariate Linear Regression	0.99	0.032	0.028	1	0.007	0.005	0.99	0.027	0.024	0.96	0.049	0.041	1	0.011	0.009
K-Neighbors Regression	0.72	0.184	0.146	0.46	0.218	0.2	0.68	0.158	0.131	0.51	0.162	0.113	0.67	0.104	0.087
Gradient Boosting Regressor	0.98	0.052	0.044	0.98	0.044	0.03	0.98	0.036	0.026	0.94	0.057	0.05	0.99	0.013	0.01
Adaboost Regressor	0.99	0.042	0.032	0.99	0.025	0.018	0.99	0.029	0.025	0.92	0.065	0.054	0.99	0.013	0.009
ANN-I	0.98	0.054	0.040	0.99	0.028	0.023	0.72	0.148	0.105	0.76	0.116	0.099	0.96	0.036	0.03
ANN-II	0.84	0.138	0.113	0.80	0.134	0.11	0.97	0.044	0.038	0.83	0.099	0.079	0.79	0.083	0.066
ANN-III	0.54	0.234	0.171	0.75	0.15	0.115	0.85	0.108	0.087	0.79	0.111	0.096	0.86	0.067	0.052

Table 4.5. Category 3 (Seasonal variation): Performance evaluation of ML and DL for wastewater effluent (COD)

Model	Dry (80%) --- Dry (20%)			Wet (80%) --- Wet (80%)			Dry (100%) --- Wet (100%)			Wet (100%) ---- Dry (100%)			Dry+Wet (80%) --- Dry+Wet (20%)		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Random Forest	0.98	0.024	0.019	0.99	0.031	0.027	0.98	0.035	0.029	0.91	0.068	0.052	0.99	0.022	0.017
Decision Tree	0.97	0.033	0.025	0.95	0.067	0.042	0.98	0.036	0.03	0.89	0.076	0.057	0.99	0.022	0.016
Extra Trees	0.99	0.019	0.017	0.99	0.022	0.018	0.99	0.028	0.023	0.96	0.045	0.039	0.99	0.015	0.011
Multivariate Linear Regression	0.99	0.016	0.014	1	0.007	0.005	0.99	0.026	0.024	0.95	0.048	0.041	1	0.012	0.01
K-Neighbors Regression	0.55	0.122	0.104	0.46	0.218	0.196	0.68	0.156	0.128	0.56	0.149	0.094	0.68	0.106	0.082
Gradient Boosting Regressor	0.99	0.02	0.016	0.98	0.041	0.029	0.98	0.035	0.028	0.94	0.057	0.048	0.99	0.016	0.013
Adaboost Regressor	0.98	0.025	0.02	0.97	0.048	0.043	0.98	0.039	0.03	0.92	0.064	0.05	0.99	0.021	0.016
ANN-I	0.96	0.036	0.03	0.96	0.06	0.041	0.96	0.055	0.044	0.71	0.121	0.099	0.94	0.044	0.035
ANN-II	0.87	0.065	0.054	0.91	0.086	0.061	0.98	0.041	0.036	0.93	0.059	0.047	0.98	0.024	0.022
ANN-III	0.85	0.07	0.044	0.61	0.186	0.144	0.94	0.065	0.051	0.74	0.116	0.094	0.74	0.096	0.077

Table 4.6. Category 3 (Seasonal variation): Performance evaluation of ML and DL for wastewater effluent (TSS)

Model	Dry (80%) --- Dry (20%)			Wet (80%) ·--- Wet (80%)			Dry (100%) --- Wet (100%)			Wet (100%) ---- Dry (100%)			Dry+Wet (80%) --- Dry+Wet (20%)		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Random Forest	0.98	0.039	0.033	0.98	0.041	0.035	0.97	0.051	0.038	0.96	0.061	0.045	0.95	0.049	0.025
Decision Tree	0.79	0.125	0.091	0.98	0.034	0.028	0.86	0.102	0.077	0.89	0.098	0.063	0.95	0.049	0.033
Extra Trees	0.99	0.029	0.023	0.96	0.054	0.04	0.99	0.032	0.026	0.97	0.047	0.036	0.97	0.039	0.016
Multivariate Linear Regression	1	0.013	0.01	1	0.012	0.01	1	0.011	0.009	0.98	0.042	0.034	1	0.009	0.007
K-Neighbors Regression	0.77	0.131	0.107	0.61	0.169	0.137	0.48	0.199	0.172	0.70	0.161	0.111	0.84	0.085	0.067
Gradient Boosting Regressor	0.99	0.028	0.019	0.97	0.049	0.04	0.97	0.05	0.038	0.98	0.041	0.03	0.96	0.04	0.019
Adaboost Regressor	0.96	0.053	0.034	0.97	0.046	0.037	0.90	0.087	0.062	0.95	0.065	0.047	0.95	0.048	0.028
ANN-I	0.74	0.14	0.103	0.93	0.069	0.055	0.68	0.156	0.133	0.85	0.114	0.101	0.93	0.054	0.036
ANN-II	0.81	0.057	0.021	0.95	0.061	0.053	0.84	0.111	0.094	0.92	0.081	0.06	0.92	0.059	0.04
ANN-III	0.82	0.114	0.095	0.88	0.094	0.089	0.64	0.166	0.144	0.70	0.158	0.127	0.95	0.046	0.039

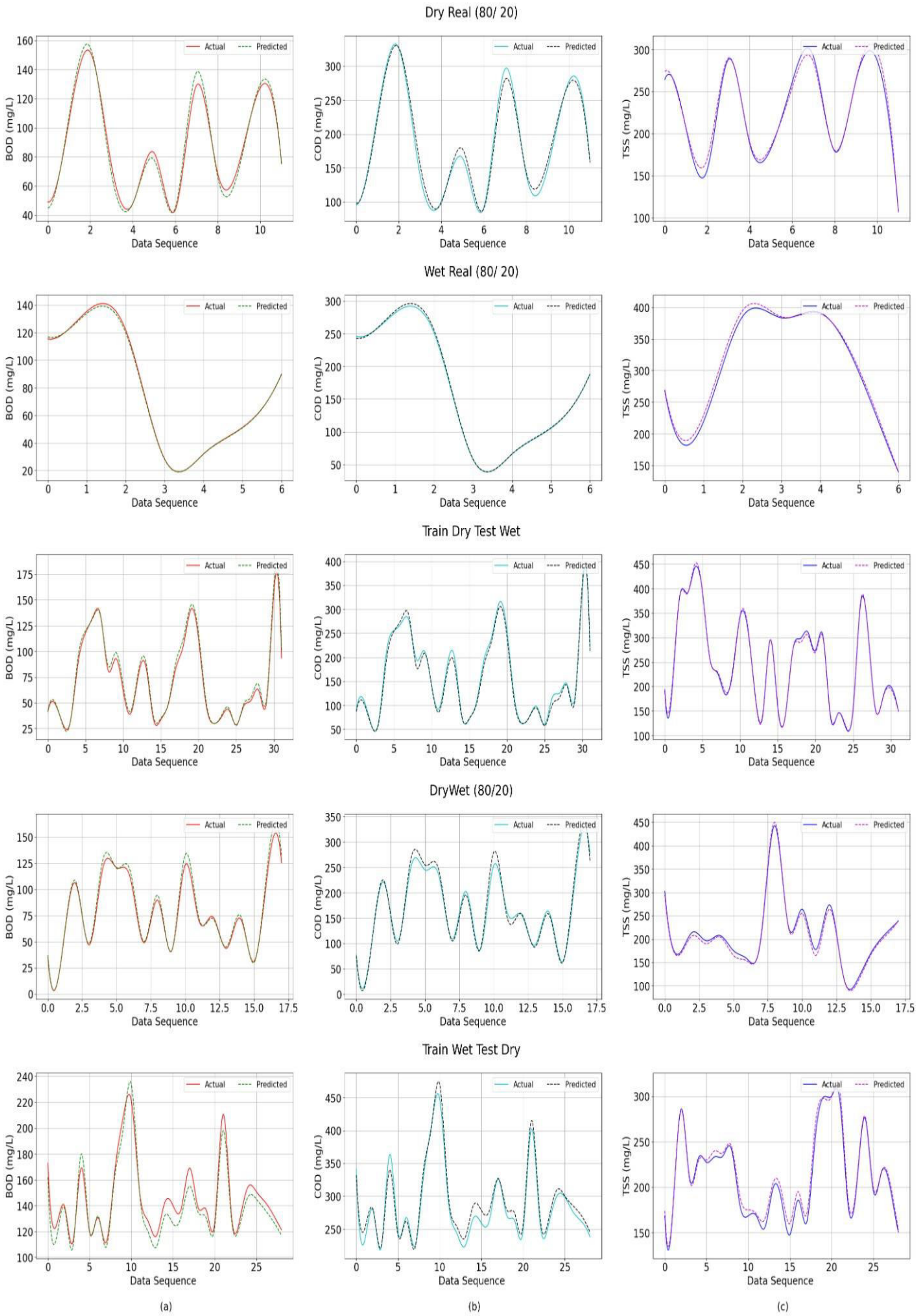

Fig. 4.8 (a) BOD_5, (b) COD and (c) TSS prediction of Multivariate Linear Regression for different scenarios (Category 3: Seasonal variation)

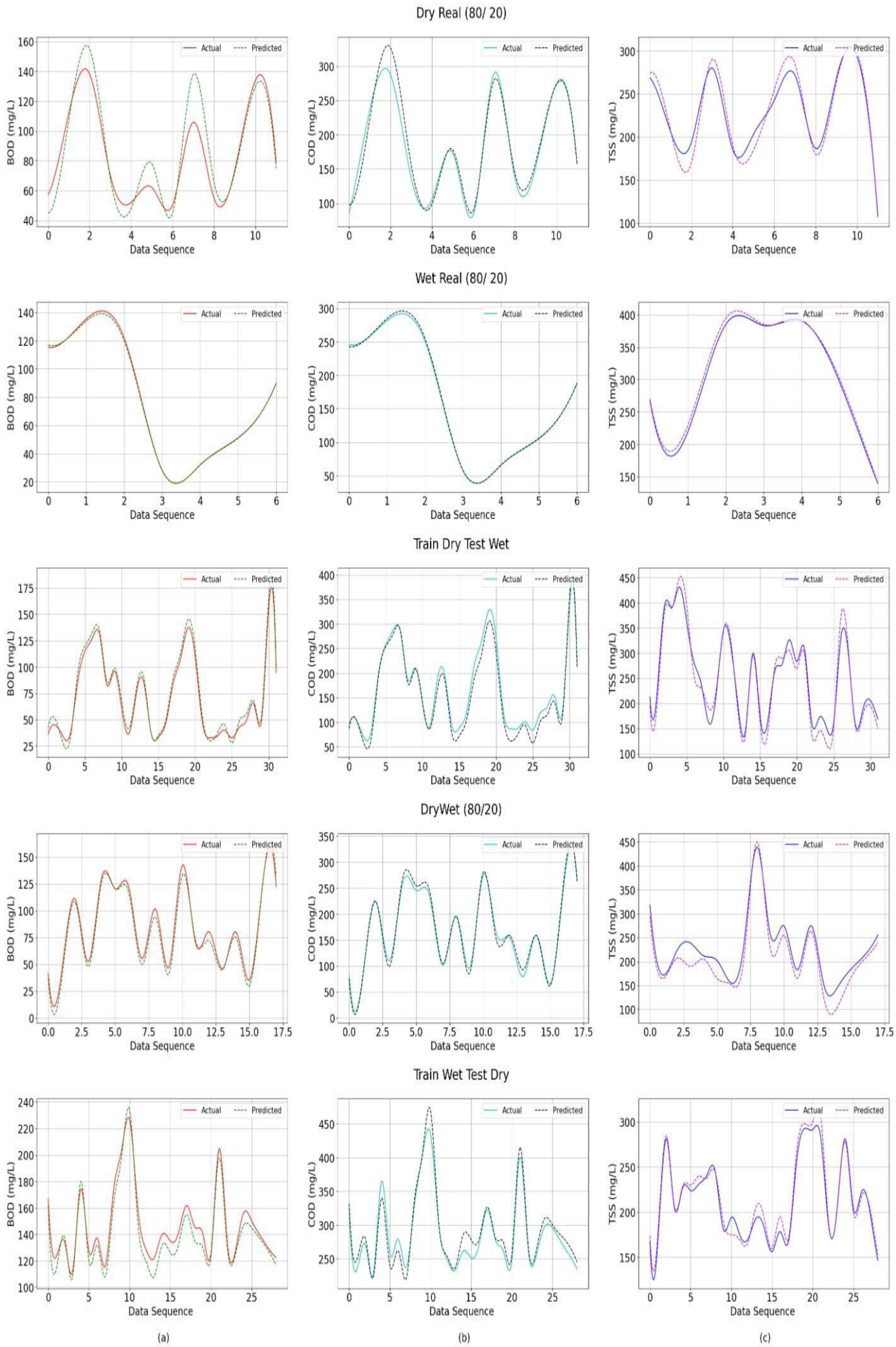

Fig. 4.9 (a) BOD_5, (b) COD and (c) TSS prediction of ANN-II for different scenarios (Category 3: Seasonal variation)

Depending on the context, the performance of ANN models varies greatly but satisfactory. It is seen that although RMSE and MAE values are pretty similar to the values that are found in machine learning models, relatively lower values of R^2 are found in comparison to machine learning models. In all pollutant cases (BOD_5, COD, and TSS), ANN-II, ANN-I and ANN-III perform well under certain conditions (R^2 between 0.8 and 0.95) but moderately under other situations (R^2 between 0.7 and 0.8) and sometimes falling below 0.6. This difference is due to the limited number of input data sets, which may hinder the model's ability to understand the intrinsic heterogeneous behavior of wastewater quality. Fig. 4.9 shows that the predicted values for BOD_5, COD, and TSS best agreed with the actual values. It indicates that the ANN-II model predicts the BOD_5 levels accurately but with some deviations, which reflects the comparatively lower performance of deep learning models. An analogous plot is also shown for COD and TSS in Fig. 4.9. The R^2, RMSE, and MAE values of this study agree with other studies, demonstrating their applicability to local contexts. Additionally, each model performs almost similarly to other studies reported elsewhere, despite the small volume of data (Qiu et al., 2016).

In summary, the overall excellent models based on limited real data for predicting BOD_5, COD and TSS are Extra Trees, Multivariate Linear Regression, Random Forest, and Adaboost Regressor. These models consistently demonstrate high R^2 values, low RMSE, and low MAE across various scenarios, indicating strong predictive power and robustness. Multivariate Linear Regression, in particular, stands out as the top-performing model with consistently outstanding results in all scenarios. Deep learning model performance, however, also meets acceptance criteria.

4.6 Category 4: Synthetic Data Generation from Real Wastewater Data

The availability of time series data is crucial for the performance of WWTP modeling studies in the AI-based machine learning process because they are critical disruptions in a WWTP. Following this line of reasoning, synthetic data of 1000 samples was created using real data of 89, collected from septic tanks, pits, and residential area drains throughout various locations in Chittagong, of which 500 samples originated from 57 samples of dry period data and the remaining 500 samples developed from 32 of the wet period dataset in Category 4.

The created synthetic data sets were evaluated in various ways to test the additional robustness of the proposed models. Real data from the dry period was judged against synthetic data that was based on that data. Again, synthetic data based on wet period real data was assessed against real data of the wet period. Then, after combining and dividing it into two portions, 80% of the artificial data from the dry and wet periods was used for training and 20% for testing. In conclusion, merged artificial data was evaluated in relation to a real sample of 89 datasets. Tables 4.7, 4.8, and 4.9 demonstrate how effectively suggested machine learning and deep learning models predict BOD_5, COD, and TSS, respectively, concerning performance metrics (R^2, RMSE, and MAE) under various dataset conditions. The BOD_5, COD, and TSS performance results obtained from a variety of trials are presented in Tables E1–E12 within Appendix E.

Extra Trees, one of the machine learning models, produced low RMSE and MAE values in the context of BOD_5 and high R^2 values (e.g., 0.99 for Synthetic to Real case of dry period data and applicable for other scenarios). This demonstrates that Extra Trees offered an excellent data fit and gave precise BOD_5 level forecasts. High R^2 values and low error metrics were also displayed by Decision Trees, Random Forest, Multivariate Linear Regression, Gradient Boosting Regressor, and Adaboost Regressor, all of which performed well. On the other hand, K-Neighbors Regression exhibited lower R^2 values and greater error metrics, indicating less accurate predictions in the BOD_5 scenario.

Table 4.7. Category 4 (Synthetic data from seasonal real wastewater data):

Performance evaluation of ML and DL for wastewater effluent (BOD_5)

Model	Dry_Synthetic --- Dry_Real			Wet_Synthetic --- Wet_Real			(Dry +Wet) Synthetic (80%) --- (Dry +Wet) Synthetic (20%)			(Dry +Wet) Synthetic ---- (Dry+ Wet) Real		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Random Forest	0.96	0.047	0.035	1	0.015	0.01	0.95	0.024	0.007	1	0.013	0.006
Decision Tree	0.96	0.045	0.032	1	0.004	0.001	0.97	0.018	0.006	1	0.001	0
Extra Trees	0.99	0.029	0.022	1	0.017	0.007	0.97	0.018	0.005	1	0.005	0.002
Multivariate Linear Regression	0.98	0.033	0.028	1	0.005	0.004	1	0.007	0.004	1	0.015	0.01
K-Neighbors Regression	0.24	0.208	0.176	0.67	0.159	0.14	0.73	0.056	0.019	0.87	0.082	0.06
Gradient Boosting Regressor	0.98	0.037	0.027	1	0.017	0.006	0.97	0.017	0.005	1	0.006	0.002
Adaboost Regressor	0.97	0.044	0.032	0.99	0.023	0.019	0.97	0.019	0.007	1	0.014	0.011
ANN-I	0.89	0.079	0.068	0.99	0.03	0.024	0.97	0.02	0.011	0.99	0.02	0.016
ANN-II	0.99	0.026	0.022	1	0.01	0.008	0.94	0.105	0.04	1	0.014	0.011
ANN-III	0.94	0.059	0.049	1	0.007	0.006	0.98	0.015	0.009	1	0.014	0.011

Table 4.8. Category 4 (Synthetic data from seasonal real wastewater data):

Performance evaluation of ML and DL for wastewater effluent (COD)

Model	Dry_Synthetic --- Dry_Real			Wet_Synthetic --- Wet_Real			(Dry +Wet) Synthetic (80%) --- (Dry +Wet) Synthetic (20%)			(Dry +Wet) Synthetic ---- (Dry+ Wet) Real		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Random Forest	0.95	0.057	0.043	0.96	0.056	0.046	0.95	0.025	0.006	1	0.012	0.007
Decision Tree	0.95	0.051	0.041	0.95	0.059	0.044	0.97	0.02	0.007	1	0.001	0
Extra Trees	0.98	0.033	0.026	0.99	0.028	0.016	0.97	0.019	0.005	1	0.004	0.001
Multivariate Linear Regression	0.98	0.032	0.028	1	0.01	0.007	1	0.007	0.004	1	0.016	0.011
K-Neighbors Regression	0.27	0.207	0.179	0.68	0.156	0.134	0.74	0.056	0.018	0.86	0.087	0.065
Gradient Boosting Regressor	0.97	0.044	0.037	0.96	0.054	0.045	0.97	0.019	0.005	1	0.006	0.003
Adaboost Regressor	0.96	0.048	0.038	0.96	0.052	0.04	0.97	0.018	0.007	1	0.016	0.012
ANN-I	0.93	0.064	0.053	0.97	0.045	0.036	0.99	0.013	0.01	1	0.021	0.016
ANN-II	0.98	0.034	0.03	0.99	0.029	0.025	0.95	0.026	0.01	1	0.02	0.016
ANN-III	0.95	0.057	0.049	0.98	0.034	0.027	0.99	0.012	0.006	1	0.019	0.014

Table 4.9. Category 4 (Synthetic data from seasonal real wastewater data):

Performance evaluation of ML and DL for wastewater effluent (TSS)

Model	Dry_Synthetic --- Dry_Real			Wet_Synthetic --- Wet_Real			(Dry +Wet) Synthetic (80%) --- (Dry +Wet) Synthetic (20%)			(Dry +Wet) Synthetic ---- (Dry+ Wet) Real		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Random Forest	0.88	0.08	0.06	0.95	0.06	0.047	0.99	0.008	0.003	1	0.01	0.005
Decision Tree	0.85	0.09	0.063	0.93	0.071	0.058	0.98	0.009	0.003	1	0.004	0.001
Extra Trees	0.96	0.043	0.033	0.99	0.029	0.024	0.99	0.006	0.003	1	0.002	0.001
Multivariate Linear Regression	0.99	0.022	0.017	0.99	0.031	0.026	1	0.005	0.003	1	0.008	0.006
K-Neighbors Regression	0.16	0.21	0.167	0.68	0.157	0.137	0.78	0.031	0.011	0.89	0.064	0.05
Gradient Boosting Regressor	0.89	0.075	0.055	0.94	0.069	0.052	0.99	0.008	0.003	1	0.003	0.002
Adaboost Regressor	0.84	0.092	0.061	0.94	0.066	0.055	0.96	0.014	0.008	0.99	0.016	0.012
ANN-I	0.94	0.057	0.045	0.93	0.071	0.06	0.94	0.016	0.01	1	0.018	0.012
ANN-II	0.96	0.048	0.038	0.95	0.062	0.056	0.96	0.014	0.009	0.97	0.033	0.028
ANN-III	0.97	0.041	0.031	0.94	0.07	0.061	0.95	0.015	0.008	1	0.009	0.007

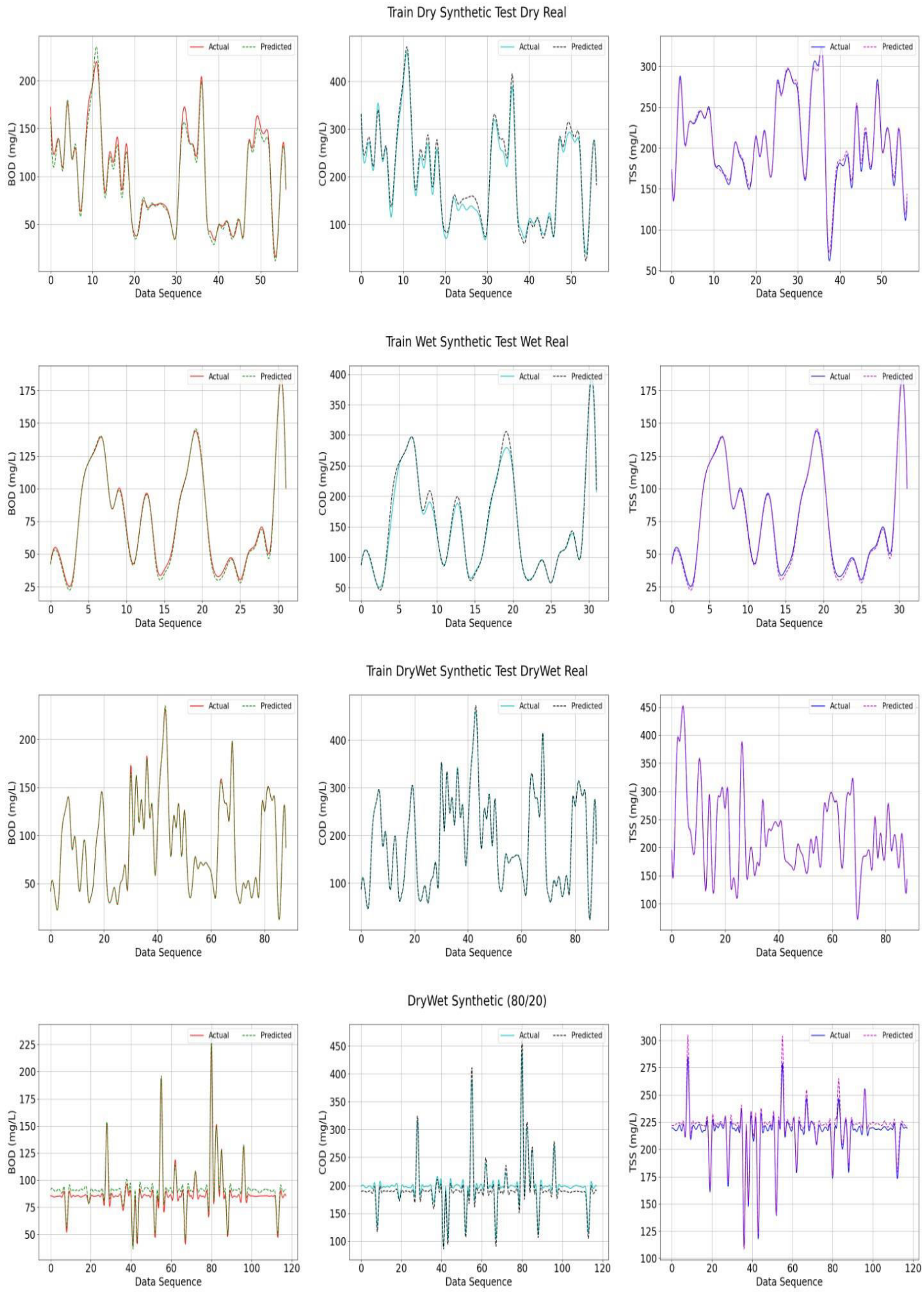

Fig. 4.10 (a) BOD_5, (b) COD and (c) TSS prediction of Multivariate Linear Regression for different scenarios (Category 4: Synthetic data from seasonal real wastewater data)

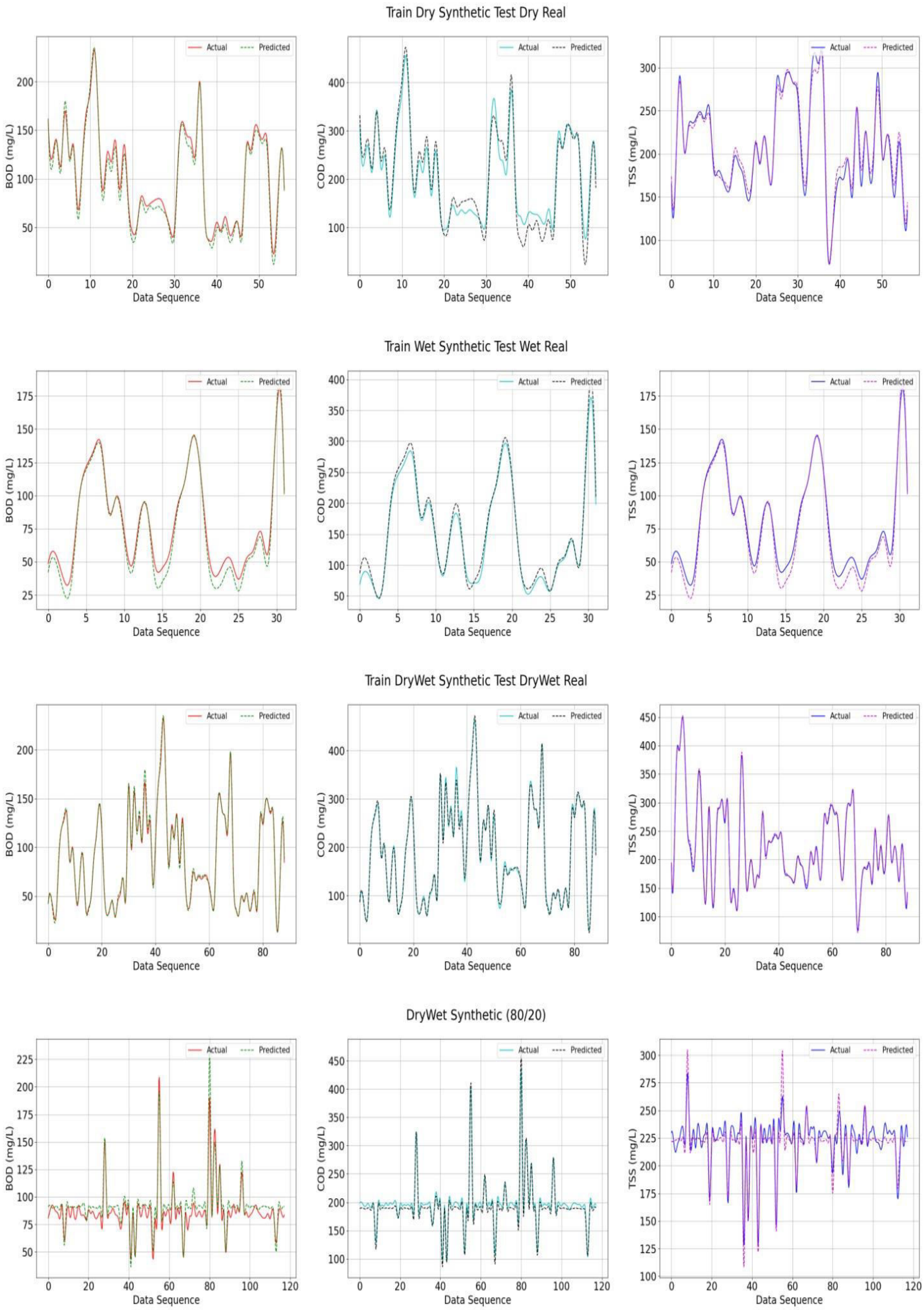

Fig. 4.11 (a) BOD_5, (b) COD and (c) TSS prediction of ANN-III for different scenarios (Category 4: Synthetic data from seasonal real wastewater data)

Multivariate Linear Regression replaces Extra Trees as the top-performing machine learning model for COD (Table 4.8) and TSS (Table 4.9) prediction, obtaining remarkably high R^2 values and extremely low RMSE and MAE values across all scenarios. However, with high R^2 values and low error metrics, Decision Trees, Random Forest, Extra Trees, Gradient Boosting Regressor, and Adaboost Regressor all show excellent performance. Lower R^2 values and larger error metrics, which indicate less accurate predictions in the context of COD and TSS, are characteristics of K-Neighbors Regression, which exhibits a similar trend to BOD_5 prediction. In Fig. 4.10, it was observed that the predicted values are pretty similar to the actual value, showing how well the Multivariate Linear Regression model predicts the BOD_5 value. The COD and TSS illustrations show a similar plot (Fig. 4.10) for Multivariate Linear Regression. The results of this study support the usage of it in regional contexts. They are consistent with extreme quality compared to earlier research (Zhao et al., 2016).

Based on artificial data sets, all three ANN models (ANN-I, ANN-II, and ANN-III) perform exceptionally well in predicting BOD_5, COD, and TSS. In all circumstances, they consistently attain high R^2 values (over 0.9) and low RMSE and MAE values. The consistency of R^2 values across all ANNs shows that they can explain a substantial portion of the variance in the target variables. Additionally, the low RMSE and MAE values imply that the prediction errors in these models are minimal. The limitations associated with category 3's small data set were tackled in this category across all situations as the size of the data set increased, demonstrating the improved ability of ANNs to understand the inherently diversified behavior of wastewater quality fully. The values predicted by the ANN-III model and the actual values are illustrated in Fig. 4.11. The ANN-III model accurately predicts the BOD_5 levels, which reflects the superior performance of deep learning models, as shown by the close positioning of the green line to the red line. An analogous plot is also illustrated for COD and TSS in the same Fig. 4.10.

The results of this study support the suitability of its implementation in regional contexts by correlating with earlier research findings (Qiu et al., 2016). It should be noted that the limitations associated with category 3's small data set were removed in this category across all scenarios as the size of the data set increased, demonstrating the improved ability of ANNs to thoroughly comprehend the naturally diverse behavior of wastewater quality.

In summary, the accuracy of the models varies depending on the specific pollutant (BOD_5, COD, or TSS) predicted and the data circumstances. Overall, employing the three ANN models to predict BOD_5, COD, and TSS, high R^2 values (over 0.9) and low RMSE and MAE values are consistently attained under all circumstances. The limitations imposed by category 3's small data set were removed in this category across all scenarios as the size of the data set increased, demonstrating the improved capacity of ANNs to comprehend the naturally variable behavior of wastewater quality thoroughly. Similarly, machine learning models, Extra Trees and Multivariate Linear Regression consistently outperformed all situations. These models excelled in capturing and forecasting the levels of these water quality indicators, exhibiting high R^2 values, low RMSE, and low MAE.

4.7 Summary of the Findings

Following the aims and methodologies employed to achieve the objectives of this study, the results, obtained through a comprehensive examination and a detailed analysis, have been presented in Table 4.10.

- This analysis suggests that when real data is unavailable, it is feasible to create synthetic data analogous to the operations and attributes of a WWTP using an AI-based machine learning model with mathematical equations, necessary assumptions and consideration of local conditions.

Table 4.10. Performance comparison between ML and DL (all categories and scenarios) for wastewater effluent (BOD_5, COD and TSS)

Category	Training	Testing	ML	DL	Superiority
Category-1 Synthetic Data (in association with mathematical equation and assumptions)	Synthetic Data (80%)	Synthetic Data (20%)	Random Forest > Extra Trees (Not All ML Models satisfactory)	ANN-III > ANN-I > ANN-II (almost same)	DL
Category-2	Synthetic Data (Category-1)	Real Data	Random Forest > Gradient Boosting Regressor (All ML Models satisfactory)	ANN-I > ANN-II > ANN-III (almost same)	ML and DL (almost same)
Category-3 (Real Data)	Dry (80%)	Dry (20%)	Multivariate Linear Regression > Extra Trees (Not All ML Models satisfactory)	ANN-II > ANN-I > ANN-III (All DLs almost equally satisfactory)	Selected ML (due to limited data) > all DLs
	Wet (80%)	Wet (20%)			
	Dry (100%)	Wet (100%)			
	Wet (100%)	Dry (100%)			
	Dry + Wet (80%)	Dry + Wet (20%)			
Category-4 (Synthetic Data Generated with Real Data)	Dry_Synthetic	Dry_Real	Multivariate Linear Regression > Extra Trees (Not All ML Models satisfactory)	ANN-III > ANN-I > ANN-II (almost same)	Selected ML and all DLs (almost same)
	Wet_Synthetic	Wet_Real			
	(Dry +Wet) Synthetic (80%)	(Dry +Wet) Synthetic (20%)			
	(Dry +Wet) Synthetic	(Dry+ Wet) Real			

- In Category 1, dealing with AI-based synthetic wastewater data, three ANNs exhibit strong predictive performance with R^2 values exceeding 0.9

and high precision, indicated by low RMSE (0.066-0.073) and MAE (0.051-0.059) values for BOD_5, COD, and TSS in the order ANN-III > ANN-I > ANN-II. In the case of ML, Random Forest and Extra Trees models perform well only for BOD_5 and COD, except TSS, showing, in some cases, negative results.

- Similar to Category 1, the three ANNs appeared to be equally effective in Category 2 in the context of real wastewater data versus synthetic data, with noteworthy results showing high R^2 values above 0.9 and consistently low RMSE (0.066-0.084) and MAE (0.052-0.067) values for the three outputs (BOD_5, COD, and TSS), except ANN-II and ANN-III, with R^2 values for TSS falling just below the 0.9 threshold, a clear indication of avoiding overfitting the data, which is a crucial consideration when working with complex real wastewater data sets. Although Random Forest and Gradient Boosting Regressor have almost identical performance with ANN, the choice of the priority ML model differs from Category 1, which is fully satisfied in the case of ANN.

- Multivariate Linear Regression and Extra Trees are the best-performing models in Category 3 based on a limited number of seasonal real wastewater data, demonstrating high R^2 values above 0.9 and low RMSE and MAE. However, they struggle in certain situations, especially in the TSS scenario, with incredibly poor performance in wet conditions. Conversely, ANNs have slightly lower R^2 values than ML models but are acceptable, with ANN-II, ANN-I, and ANN-III performing well in some circumstances but moderately in others. This inconsistency may be due to a lack of sufficient data sets, making it difficult to fully understand the naturally varied behavior of wastewater quality.

- In Category 4, assessing the generation of synthetic seasonal data from limited real wastewater data, all three ANN models consistently achieve the best results in terms of R^2 exceeding 0.9 and low RMSE (0.009-0.019) and MAE

(0.007-0.011) when predicting BOD_5, COD, and TSS under all conditions, and the restrictions associated with the limited data set in Category 3 were eliminated in this category across all scenarios due to the increased data set, revealing the improved ability of ANNs to understand the naturally diverse behavior of wastewater quality comprehensively. In contrast, the best-performing ML models are different for BOD_5, COD, and TSS, and some ML models encounter difficulties in certain conditions, similar to Category 1 and Category 3, particularly in the TSS scenario.

Chapter 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 General

The performance of wastewater treatment facilities using data-driven and artificial intelligence-based techniques has become prominent in recent years. These approaches do not require any knowledge about the structural details or current condition of the system. However, the effectiveness of these strategies is heavily influenced by the data quality. AI models have the advantage of being able to estimate effluent concentrations without having any prior knowledge of the system. Furthermore, there is no need to make any assumptions about the mathematical relationships between inputs and outputs. These models are capable of identifying the connections between the input and output elements without requiring explicit consideration of the underlying physics of the process. This study aims to assess the effectiveness of artificial intelligence-driven machine learning tools in simulating activated sludge wastewater treatment plants. ASS modeling using various AI-based machine learning and deep learning techniques in association with addressing variability in wastewater characteristics is evaluated. The most effective AI-based ML and DL algorithms for the best ASS performance were revealed.

5.2 Conclusions of the Study

After a thorough examination and in-depth analysis involving various scenarios, including seasonal fluctuations, this study concludes the following:

- AI-based machine learning tools demonstrate significant potential for synthetic data generation, similar to WWTP operations and attributes, tested by basic statistics and the correlation of wastewater parameters when real data is unavailable, to model the biological treatment process for treating domestic wastewater.

- The variability of wastewater composition and their reduction potential performances by ASS using ANN are found to be fairly appropriate for the prediction of effluent quality parameters.

- The ANN-III model is revealed as the most effective predictive tool in this investigation, demonstrating exceptional performance across a wide range of conditions, including seasonal variations with R^2 values exceeding 0.9, as well as maintaining low values of RMSE (0.009-0.084) and MAE (0.007-0.067) in almost all variabilities addressed.

Utilizing AI-based ML and DL algorithms, in particular Artificial Neural Networks (ANN), has proven to be an efficient method for generating synthetic data when real data is unavailable because of its capacity to capture intricate non-linear correlations within the data.

5.3 Implications of this Study

The effectiveness of WWTP modeling studies greatly depends on the presence of time-series data, as this data is essential for understanding disruptions in WWTPs caused by various factors, such as changes in wastewater composition due to dietary habits, environmental conditions, climate, and other local factors. Unfortunately, in Bangladesh, real-time series data is often unavailable, making it challenging to gather sufficient information. In the absence of real data, it is possible to create synthetic wastewater data utilizing AI-based machine learning tools that closely resemble the operations and characteristics of the WWTP by considering the specific local conditions. Simultaneously, ML and DL algorithms, particularly ANN, can capture complex non-linear correlations among wastewater parameters. This surely helps wastewater professionals monitor WWTPs effectively, promptly identify issues, take the remedial action needed for existing or newly developed WWTPs, and make decisions related to wastewater treatment, quality control, process optimization, and environmental pollution control and management.

5.4 Recommendations for Future Study

While this study has achieved notable success, specific aspects have been recognized as areas for potential further investigation. In the continuation of this research, future studies can explore the following issues:

- This study primarily concentrates on tackling the multiple-input-single-output (MISO) situation; however, the use of a multiple-input-multiple-output (MIMO) technique for evaluating the resilience of artificial neural networks (ANNs) can be explored with the developed models.

- Testing the model with real-time series and seasonal operational data from wastewater treatment plants is a recommended way to assess the precision, homogeneity, and dependability of the model.

- Another suggested strategy is to use the same techniques with synthetic data, considering diverse parameters beyond BOD_5, COD, and TSS to assess the potential for 3Rs (reduce, reuse and recycle) in accordance with the guidelines outlined in Bangladesh Standards (BECR, 2023).

- One potential avenue for promising future research involves utilizing feature extraction methods like PCA on the same dataset and then conducting a comparative analysis with the models that have been developed.

It is vital to remember that the choice of a neural network structure for activated sludge wastewater treatment system depends on the specific circumstances, data availability, and modeling goals. Suitable data pre-processing and validation approaches are also essential for developing precise and trustworthy ANN models for wastewater treatment operations.

Bibliography

Abolpour, B., Najaf Tarqi, M., Askari Dolatabad, Y., & Salahi Sarbijan, F. (2021). Risk evaluation and distribution of arsenic concentration in groundwater resources of villages in Jiroft County Kerman Province Iran. *Caspian Journal of Environmental Sciences*, *19*(3). https://doi.org/10.22124/cjes.2021.4944

Alani, R., & Khudhair, B. (2019). Prediction of Sediment Accumulation Model for Trunk Sewer Using Multiple Linear Regression and Neural Network Techniques. *Civil Engineering Journal*, *5*, 82. https://doi.org/10.28991/cej-2019-03091227

Alleman, J. E., Kennedy, M. S., Mines, R. O., Weber, A. S., & Sherrard, J. H. (1984). Activated Sludge. *Journal (Water Pollution Control Federation)*, *56*(6), 568–577.

Alsulaili, A., & Refaie, A. (2020). Artificial neural network modeling approach for the prediction of five-day biological oxygen demand and wastewater treatment plant performance. *Water Supply*, *21*(5), 1861–1877. https://doi.org/10.2166/ws.2020.199

Alver, A., & Kazan, Z. (2020). Prediction of full-scale filtration plant performance using artificial neural networks based on principal component analysis. *Separation and Purification Technology*, *230*, 115868. https://doi.org/10.1016/j.seppur.2019.115868

Andrews, J. F. (2012, May 30). *Dynamic Control of Wastewater Treatment Plants* (world) [Research-article]. ACS Publications; American Chemical Society. https://doi.org/10.1021/es00058a722

APHA. (2005). *American Public Health Association (APHA), 2005. Standard Methods for the Examination of Water and Wastewater, twenty-first ed. American Public Health Association, Washington, D.C., p. 46.*

Bagheri, M., Mirbagheri, S. A., Bagheri, Z., & Kamarkhani, A. M. (2015). Modeling and optimization of activated sludge bulking for a real wastewater treatment plant using hybrid artificial neural networks-genetic algorithm approach. *Process Safety and Environmental Protection*, *95*, 12–25. https://doi.org/10.1016/j.psep.2015.02.008

Baki, O. T., Aras, E., Akdemir, U. O., & Yilmaz, B. (2019a). Biochemical oxygen demand prediction in wastewater treatment plant by using different regression analysis models. *DESALINATION AND WATER TREATMENT*, *157*, 79–89. https://doi.org/10.5004/dwt.2019.24158

Baki, O. T., Aras, E., Akdemir, U. O., & Yilmaz, B. (2019b). Biochemical oxygen demand prediction in wastewater treatment plant by using different regression analysis models. *DESALINATION AND WATER TREATMENT*, *157*, 79–89. https://doi.org/10.5004/dwt.2019.24158

Barnett, M. W., & Andrews, J. F. (1992). Expert System for Anaerobic-Digestion-Process Operation. *Journal of Environmental Engineering*, *118*(6), 949–963. https://doi.org/10.1061/(ASCE)0733-9372(1992)118:6(949)

BECR, B. (2023). *Bangladesh Environment Conservation Rules*.

Bekkari, N., & Zeddouri, A. (2018). 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), 593–608. https://doi.org/10.1108/MEQ-04-2018-0084

Borzooei, S., Campo, G., Cerutti, A., Meucci, L., Panepinto, D., Ravina, M., Riggio, V., Ruffino, B., Scibilia, G., & Zanetti, M. (2019). Optimization of the wastewater treatment plant: From energy saving to environmental impact mitigation. *Science of The Total Environment*, *691*, 1182–1189. https://doi.org/10.1016/j.scitotenv.2019.07.241

Buaisha, M., Balku, S., & Özalp-Yaman, Ş. (2020). Heavy Metal Removal Investigation in Conventional Activated Sludge Systems. *Civil Engineering Journal*, *6*, 470–477. https://doi.org/10.28991/cej-2020-03091484

Buhr, H. O., & Miller, S. B. (1983). A dynamic model of the high-rate algal-bacterial wastewater treatment pond. *Water Research*, *17*(1), 29–37. https://doi.org/10.1016/0043-1354(83)90283-X

Cao, W., & Yang, Q. (2020). Online sequential extreme learning machine based adaptive control for wastewater treatment plant. *Neurocomputing*, *408*, 169–175. https://doi.org/10.1016/j.neucom.2019.05.109

Capodaglio, A. G. (1994). *Evaluation of modelling techniques for wastewater treatment plant automation.* https://www.zotero.org/msicecuet/items/P79Q6MTZ/item-list

Celik, U., Yurtay, N., & SERTKAYA, C. (2013). Wastewater Effluent Prediction Based on Decision Tree. *Journal of Selcuk University Natural and Applied Science*, *ICOEST Conf.*, 138–148.

Chan, C. W., & Huang, G. H. (2003). Artificial intelligence for management and control of pollution minimization and mitigation processes. *Engineering Applications of Artificial Intelligence*, *16*(2), 75–90. https://doi.org/10.1016/S0952-1976(03)00062-9

Cheng, Q., Chunhong, Z., & Qianglin, L. (2023). Development and application of random forest regression soft sensor model for treating domestic wastewater in a sequencing batch reactor. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-023-36333-8

Clesceri, L. S. (2012). *Standard Methods for the Examination of Water and Wastewater, 22nd Edition* (E. W. Rice, R. B. Baird, A. D. Eaton, & L. S. Clesceri, Eds.; 22nd edition). American Water Works Association.

Courtens, E., De Clippeleir, H., Vlaeminck, S., Boon, N., & Verstraete, W. (2011). Simple strategies to control the oxygen budget of an OLAND rotating biological contactor. *Microbial Resource Management in Biotechnology, 1st International Symposium, Abstracts*. 1st International symposium on

Microbial Resource Management in Biotechnology (MRM 2011) : Concepts and applications. http://hdl.handle.net/1854/LU-2008219

Davis, M. L., & Cornwell, D. A. (2013). *Introduction to environmental engineering* (5th ed). McGraw-Hill.

Eckenfelder, W. W., & Barnhart, E. L. (1963). Performance of a High Rate Trickling Filter Using Selected Media. *Journal (Water Pollution Control Federation)*, *35*(12), 1535–1551.

EL Bilali, A., Taleb, A., Bahlaoui, M. A., & Brouziyne, Y. (2021). An integrated approach based on Gaussian noises-based data augmentation method and AdaBoost model to predict faecal coliforms in rivers with small dataset. *Journal of Hydrology*, *599*, 126510. https://doi.org/10.1016/j.jhydrol.2021.126510

Elkiran, G., & Abba, S. I. (2017, August). *6. Effluent prediction of chemical oxygen demand from the astewater treatment plant using artificial neural network application.pdf | My Library | Zotero*. https://www.zotero.org/msicecuet/items/HRR5B3ZX/library

Elmaadawy, K., Elaziz, M. A., Elsheikh, A. H., Moawad, A., Liu, B., & Lu, S. (2021). Utilization of random vector functional link integrated with manta ray foraging optimization for effluent prediction of wastewater treatment plant. *Journal of Environmental Management*, *298*, 113520. https://doi.org/10.1016/j.jenvman.2021.113520

El-Rawy, M., Abd-Ellah, M. K., Fathi, H., & Ahmed, A. K. A. (2021). Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. *Journal of Water Process Engineering*, *44*, 102380. https://doi.org/10.1016/j.jwpe.2021.102380

Erdirencelebi, D., & Yalpir, S. (2011). Adaptive network fuzzy inference system modeling for the input selection and prediction of anaerobic digestion effluent quality. *Applied Mathematical Modelling*, *35*(8), 3821–3832. https://doi.org/10.1016/j.apm.2011.02.015

Fenu, A., Guglielmi, G., Jimenez, J., Spèrandio, M., Saroj, D., Lesjean, B., Brepols, C., Thoeye, C., & Nopens, I. (2010). Activated sludge model (ASM) based modelling of membrane bioreactor (MBR) processes: A critical review with special regard to MBR specificities. *Water Research*, *44*(15), 4272–4294. https://doi.org/10.1016/j.watres.2010.06.007

Frigon, D., Biswal, B. K., Mazza, A., Masson, L., & Gehr, R. (2013). Biological and Physicochemical Wastewater Treatment Processes Reduce the Prevalence of Virulent Escherichia coli. *Applied and Environmental Microbiology*, *79*(3), 835–844. https://doi.org/10.1128/AEM.02789-12

Friha, I., Karray, F., Feki, F., Jlaiel, L., & Sayadi, S. (2014). Treatment of cosmetic industry wastewater by submerged membrane bioreactor with consideration of microbial community dynamics. *International Biodeterioration & Biodegradation*, *88*, 125–133. https://doi.org/10.1016/j.ibiod.2013.12.015

Fu, C., & Poch, M. (1995). System identification and real-time pattern recognition by neural networks for an activated sludge process. *Environment International*, *21*(1), 57–69. https://doi.org/10.1016/0160-4120(94)00024-2

Garg, S. K. (2014). *ENVIRONMENTAL ENGINEERING: SEWAGE DISPOSAL AND AIR POLLUTION ENGINEERING VOL .II,31/ED*. KHANNA PUBLISHERS.

Girgin, S., Polat, F., & Alhajj, R. (2010). Improving reinforcement learning by using sequence trees. *Machine Learning*, *81*(3), 283–331. https://doi.org/10.1007/s10994-010-5182-y

Gleick, P. H. (2000). A Look at Twenty-first Century Water Resources Development. *Water International*, *25*(1), 127–138. https://doi.org/10.1080/02508060008686804

Granata, F., Papirio, S., Esposito, G., Gargano, R., & De Marinis, G. (2017). Machine Learning Algorithms for the Forecasting of Wastewater Quality Indicators. *Water*, *9*(2), Article 2. https://doi.org/10.3390/w9020105

Gujer, W., Henze, M., Mino, T., & Loosdrecht, M. van. (1999). Activated sludge model No. 3. *Water Science and Technology*, *39*(1), 183–193. https://doi.org/10.1016/S0273-1223(98)00785-9

Gujer, W., Henze, M., Mino, T., Matsuo, T., Wentzel, M. C., & Marais, G. v. R. (1995). The Activated Sludge Model No. 2: Biological phosphorus removal. *Water Science and Technology*, *31*(2), 1–11. https://doi.org/10.1016/0273-1223(95)00175-M

Guo, C., Chen, Y., Xia, W., Qu, X., Yuan, H., Xie, S., & Lin, L.-S. (2020). Eutrophication and heavy metal pollution patterns in the water suppling lakes of China's south-to-north water diversion project. *Science of The Total Environment*, *711*, 134543. https://doi.org/10.1016/j.scitotenv.2019.134543

Guo, H., Jeong, K., Lim, J., Jo, J., Kim, Y. M., Park, J., Kim, J. H., & 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. https://doi.org/10.1016/j.jes.2015.01.007

Hameed, M. M., AlOmar, M. K., Khaleel, F., & Al-Ansari, N. (2021). An Extra Tree Regression Model for Discharge Coefficient Prediction: Novel, Practical Applications in the Hydraulic Sector and Future Research Directions. *Mathematical Problems in Engineering*, *2021*, e7001710. https://doi.org/10.1155/2021/7001710

Harremoës, P., Capodaglio, A. G., Hellström, B. G., Henze, M., Jensen, K. N., Lynggaard-Jensen, A., Otterpohl, R., & Søeberg, H. (1993). Wastewater Treatment Plants under Transient Loading – Performance, Modelling and Control. *Water Science and Technology*, *27*(12), 71–115. https://doi.org/10.2166/wst.1993.0292

Hassen, E., & Asmare, A. (2019). *Predictive performance modeling of Habesha brewery wastewater treatment plant using artificial neural network*. https://doi.org/10.31221/osf.io/k6bvj

Hauduc, H., Rieger, L., Ohtsuki, T., Shaw, A., Takács, I., Winkler, S., Héduit, A., Vanrolleghem, P., & Gillot, S. (2011). Activated Sludge Modelling: Development and Potential Use of a Practical Applications Database. *Water Science and Technology : A Journal of the International Association on Water Pollution Research*, *63*, 2164–2182. https://doi.org/10.2166/wst.2011.368

Henze, M., Grady Jr, L., Gujer, W., Marais, G., & Matsuo, T. (1987). Activated Sludge Model No 1. *Wat Sci Technol*, *29*.

Howell, J. A., & Atkinson, B. (1976). Sloughing of microbial film in trickling filters. *Water Research*, *10*(4), 307–315. https://doi.org/10.1016/0043-1354(76)90172-X

Hu, B., Xue, J., Zhou, Y., Shao, S., Fu, Z., Li, Y., Chen, S., Qi, L., & Shi, Z. (2020). Modelling bioaccumulation of heavy metals in soil-crop ecosystems and identifying its controlling factors using machine learning. *Environmental Pollution*, *262*, 114308. https://doi.org/10.1016/j.envpol.2020.114308

Iturmendi, F., Estrada, V., Paz Ochoa, M., Hoch, P. M., & Soledad Diaz, M. (2012). Biological Wastewater Treatment: Dynamic Global Sensitivity Analysis and Parameter Estimation in a System of Waste Stabilization Ponds. In I. D. L. Bogle & M. Fairweather (Eds.), *Computer Aided Chemical Engineering* (Vol. 30, pp. 212–216). Elsevier. https://doi.org/10.1016/B978-0-444-59519-5.50043-5

Ivanciu, P., Ungureanu, N., & Vladut, V. (2019, November 7). *Applications of rotating biological contactors in wastewater treatment*.

Jeppsson, U. (2012). *Modelling Aspects of Wastewater Treatment Processes*.

Kang, J.-H., Song, J., Yoo, S. S., Lee, B.-J., & Ji, H. W. (2020). Prediction of Odor Concentration Emitted from Wastewater Treatment Plant Using an Artificial Neural Network (ANN). *Atmosphere*, *11*(8), Article 8. https://doi.org/10.3390/atmos11080784

Karia, G. L., & Christian, R. A. (2013). *WASTEWATER TREATMENT: Concepts and Design Approach*. PHI Learning.

Kasem, R., ALabdeh, D., Noori, R., & Karbassi, A. (2018). A software sensor for in-situ monitoring of the 5-day biochemical oxygen demand. *Rudarsko-Geološko-Naftni Zbornik*, *33*(1), Article 1. https://doi.org/10.17794/rgn.2018.1.3

Keskitalo, J., & Leiviskä, K. (2015). Artificial Neural Network Ensembles in Hybrid Modelling of Activated Sludge Plant. In P. Angelov, K. T. Atanassov, L. Doukovska, M. Hadjiski, V. Jotsov, J. Kacprzyk, N. Kasabov, S. Sotirov, E. Szmidt, & S. Zadrożny (Eds.), *Intelligent Systems'2014* (pp. 683–694). Springer International Publishing. https://doi.org/10.1007/978-3-319-11313-5_60

Kim, M., Kim, Y., Kim, H., Piao, W., & Kim, C. (2015). Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant. *Frontiers of Environmental Science & Engineering*, *10*. https://doi.org/10.1007/s11783-015-0825-7

Koné, D., & Strauss, M. (2004). *Low-cost Options for Treating Faecal Sludges (FS) in Developing Countries – Challenges and Performance*.

Kristensen, N. R., Madsen, H., & Jørgensen, S. B. (2004). A Method for Systematic Improvement of Stochastic Grey-Box Models. *Computers & Chemical Engineering*, *28*(8), 1431–1449. https://doi.org/10.1016/j.compchemeng.2003.10.003

Laskov, P., & Lippmann, R. (2010). Machine learning in adversarial environments. *Machine Learning*, *81*(2), 115–119. https://doi.org/10.1007/s10994-010-5207-6

Lee, D. S., Jeon, C. O., Park, J. M., & Chang, K. S. (2002). Hybrid neural network modeling of a full-scale industrial wastewater treatment process. *Biotechnology and Bioengineering*, *78*(6), 670–682. https://doi.org/10.1002/bit.10247

Lessard, P., & Beck, M. B. (1988). Dynamic Modeling of Primary Sedimentation. *Journal of Environmental Engineering*, *114*(4), 753–769. https://doi.org/10.1061/(ASCE)0733-9372(1988)114:4(753)

Lessard, P., & Beck, M. B. (1993). Dynamic modelling of the activated sludge process: A case study. *Water Research*, *27*(6), 963–978. https://doi.org/10.1016/0043-1354(93)90060-U

Liu, W., Zhang, J., Liu, H., Guo, X., Zhang, X., Yao, X., Cao, Z., & Zhang, T. (2021). A review of the removal of microplastics in global wastewater treatment plants: Characteristics and mechanisms. *Environment International*, *146*, 106277. https://doi.org/10.1016/j.envint.2020.106277

Lizarralde, I., Fernández-Arévalo, T., Brouckaert, C., Vanrolleghem, P., Ikumi, D. S., Ekama, G. A., Ayesa, E., & Grau, P. (2015). A new general methodology for incorporating physico-chemical transformations into multi-phase wastewater treatment process models. *Water Research*, *74*, 239–256. https://doi.org/10.1016/j.watres.2015.01.031

Lu, H., & Ma, X. (2020). Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere*, *249*, 126169. https://doi.org/10.1016/j.chemosphere.2020.126169

Lubensky, J., Ellersdorfer, M., & Stocker, K. (2019). Ammonium recovery from model solutions and sludge liquor with a combined ion exchange and air stripping process. *Journal of Water Process Engineering*, *32*, 100909. https://doi.org/10.1016/j.jwpe.2019.100909

Mannina, G., Rebouças, T. F., Cosenza, A., Sànchez-Marrè, M., & Gibert, K. (2019). Decision support systems (DSS) for wastewater treatment plants – A review of the state of the art. *Bioresource Technology*, *290*, 121814. https://doi.org/10.1016/j.biortech.2019.121814

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. https://doi.org/10.1007/s13201-017-0526-4

Mara, D. D. (2004). *Domestic wastewater treatment in developing countries*. Earthscan Publications.

Mjalli, F. S., Al-Asheh, S., & Alfadala, H. E. (2007). Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance. *Journal of Environmental Management*, *83*(3), 329–338. https://doi.org/10.1016/j.jenvman.2006.03.004

Moral, H., Aksoy, A., & Gokcay, C. F. (2008). Modeling of the activated sludge process by using artificial neural networks with automated architecture screening. *Computers & Chemical Engineering*, *32*(10), 2471–2478. https://doi.org/10.1016/j.compchemeng.2008.01.008

Moreno, M. D. (1990). A tracer study of the hydraulics of facultative stabilization ponds. *Water Research*, *24*(8), 1025–1030. https://doi.org/10.1016/0043-1354(90)90125-P

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: A guide for data scientists* (First edition). O'Reilly Media, Inc.

Nadiri, A. A., Shokri, S., Tsai, F. T.-C., & Asghari Moghaddam, A. (2018). Prediction of effluent quality parameters of a wastewater treatment plant using a supervised committee fuzzy logic model. *Journal of Cleaner Production*, *180*, 539–549. https://doi.org/10.1016/j.jclepro.2018.01.139

Najafzadeh, M., & Ghaemi, A. (2019). Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. *Environmental Monitoring and Assessment*, *191*. https://doi.org/10.1007/s10661-019-7446-8

Najar, Z. A., & Engin, S. N. (2019). Development of Model for Activated Sludge Process. *International Journal of Scientific Research*.

Nasr, M., Moustafa, M., Seif, H., & Elkobrosy, G. (2012). Application of Artificial Neural Network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-EGYPT. *AEJ - Alexandria Engineering Journal*, *51*, 37–43. https://doi.org/10.1016/j.aej.2012.07.005

Newhart, K. B., Holloway, R. W., Hering, A. S., & Cath, T. Y. (2019). *Data-driven performance analyses of wastewater treatment plants: A review—ScienceDirect*. https://www.sciencedirect.com/science/article/abs/pii/S0043135419302490

Nezhad, M., Mehrdadi, N., Torabian, A., & Behboudian, S. (2015). Artificial neural network modeling of the effluent quality index for municipal wastewater treatment plants using quality variables: South of Tehran wastewater treatment plant. *Journal of Water Supply: Research and Technology - Aqua*, *65*, jws2015030. https://doi.org/10.2166/aqua.2015.030

Nopens, I., Batstone, D. J., Copp, J. B., Jeppsson, U., Volcke, E., Alex, J., & Vanrolleghem, P. A. (2009). An ASM/ADM model interface for dynamic plant-wide simulation. *Water Research*, *43*(7), 1913–1923. https://doi.org/10.1016/j.watres.2009.01.012

Olsson, G., Nielsen, M., Yuan, Z., Lynggaard-Jensen, A., & Steyer, J.-P. (2005). *Instrumentation, Control and Automation in Wastewater Systems*. IWA Publishing.

Orhon, D., Karahan, O., Zengin, G. E., Olsson, O., Bauer, M., & Development (OECD), O. for E. C.-O. and. (2005). *Mechanism and Design of Sequencing Batch Reactors for Nutrient Removal*. IWA Publishing.

Otterpohl, R., & Freund, M. (1992). *Dynamic Models for Clarifiers of Activated Sludge Plants with Dry and Wet Weather Flows—ProQuest*. https://www.proquest.com/openview/5b12fc710bbfa0618e162b20c1de05be/1?pq-origsite=gscholar&cbl=2044520

Paraskevas, P. A., Pantelakis, I. S., & Lekkas, T. D. (1999). An advanced integrated expert system for wastewater treatment plants control. *Knowledge-Based Systems*, *12*, 355–361. https://doi.org/10.1016/s0950-7051(99)00040-4

Qiu, Y., Liu, Y., & Huang, D. (2016). Date-Driven Soft-Sensor Design for Biological Wastewater Treatment Using Deep Neural Networks and Genetic Algorithms. *Journal of Chemical Engineering of Japan*, *49*(10), 925–936. https://doi.org/10.1252/jcej.16we016

Rivas, A., Irizar, I., & Ayesa, E. (2008). Model-based optimisation of Wastewater Treatment Plants design. *Environmental Modelling & Software*, *23*(4), 435–450. https://doi.org/10.1016/j.envsoft.2007.06.009

Rose C., Parker A., Jefferson B., & Cartmell E. (2015). *The Characterization of Feces and Urine: A Review of the Literature to Inform Advanced Treatment Technology—PMC*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4500995/

Rout, P. R., Zhang, T. C., Bhunia, P., & Surampalli, R. Y. (2021). Treatment technologies for emerging contaminants in wastewater treatment plants: A review. *Science of The Total Environment*, *753*, 141990. https://doi.org/10.1016/j.scitotenv.2020.141990

Rustum, R. (2009). *Modelling activated sludge wastewater treatment plants using artificial intelligence techniques (fuzzy logic and neural networks)* [Thesis, Heriot-Watt University]. https://www.ros.hw.ac.uk/handle/10399/2207

Rustum, R., & Adeloye, A. (2007a). Replacing Outliers and Missing Values from Activated Sludge Data Using Kohonen Self-Organizing Map. *Journal of Environmental Engineering-Asce - J ENVIRON ENG-ASCE, 133*. https://doi.org/10.1061/(ASCE)0733-9372(2007)133:9(909)

Rustum, R., & Adeloye, A. (2007b, September). *Modelling activated sludge process using Adaptive Network-Based Fuzzy Inference System (ANFIS): 10th IWA Specialised Conference on Design, Operation and Economics of Large Wastewater Treatment Plants*.

Safder, U., Loy-Benitez, J., Nguyen, H.-T., & Yoo, C. (2022). A hybrid extreme learning machine and deep belief network framework for sludge bulking monitoring in a dynamic wastewater treatment process. *Journal of Water Process Engineering*, *46*, 102580. https://doi.org/10.1016/j.jwpe.2022.102580

Safeer, S., Pandey, R. P., Rehman, B., Safdar, T., Ahmad, I., Hasan, S. W., & Ullah, A. (2022). A review of artificial intelligence in water purification and wastewater treatment: Recent advancements. *Journal of Water Process Engineering*, *49*, 102974. https://doi.org/10.1016/j.jwpe.2022.102974

Sah, L., Rousseau, D. P. L., & Hooijmans, C. M. (2012). Numerical Modelling of Waste Stabilization Ponds: Where Do We Stand? *Water, Air, & Soil Pollution*, *223*(6), 3155–3171. https://doi.org/10.1007/s11270-012-1098-4

Schuetze, M., Butler, D., & Beck, M. (2002). *Modelling, Simulation and Control of Urban Wastewater Systems*. https://doi.org/10.1007/978-1-4471-0157-4

Séguret, F., Racault, Y., & Sardin, M. (2000). Hydrodynamic behavior of full scale trickling filters. *Water Research - WATER RES*, *34*, 1551–1558. https://doi.org/10.1016/S0043-1354(99)00317-6

Sharafati, A., Asadollah, S. B. H. S., & Hosseinzadeh, M. (2020). The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. *Process Safety and Environmental Protection*, *140*, 68–78. https://doi.org/10.1016/j.psep.2020.04.045

Sharghi, E., Nourani, V., Ashrafi, A. A., & Gökçekuş, H. (2019). Monitoring effluent quality of wastewater treatment plant by clustering based artificial neural network method. *DESALINATION AND WATER TREATMENT*, *164*, 86–97. https://doi.org/10.5004/dwt.2019.24385

Shi, J., Xu, C., Han, Y., & Han, H. (2021). Case study on wastewater treatment technology of coal chemical industry in China. *Critical Reviews in Environmental Science and Technology*, *51*(10), 1003–1044. https://doi.org/10.1080/10643389.2020.1742530

Shi, S., Xu, G., Yu, H., & Zhang, Z. (2018). Strategies of valorization of sludge from wastewater treatment. *Journal of Chemical Technology & Biotechnology*, *93*(4), 936–944. https://doi.org/10.1002/jctb.5548

Siegrist, H., & Gujer, W. (1987). Demonstration of mass transfer and pH effects in a nitrifying biofilm. *Water Research*, *21*(12), 1481–1487. https://doi.org/10.1016/0043-1354(87)90131-X

Spellman, F. R., & Drinan, J. (2003). *Wastewater Treatment Plant Operations Made Easy: A Practical Guide for Licensure*. DEStech Publications, Inc.

Stropky, D., Pougatch, K., Nowak, P., Salcudean, M., Pagoria, P., Gartshore, I., & Yuan, J. (2007). RTD (residence time distribution) predictions in large mechanically aerated lagoons. *Water Science and Technology*, *55*(11), 29–36. https://doi.org/10.2166/wst.2007.346

Takeguchi, M. M., Simpson, L. B., Smith, D. K., Morse, L. R., Ramsey, W. S., Messing, R. A., Martin, P. J., & Sharma, B. P. (1981). Rapid estimation of biochemical oxygen demand. *Applied Biochemistry and Biotechnology*, *6*(4), 273–282. https://doi.org/10.1007/BF02798278

Tchobanoglous, G., Burton, F. L., Stensel, H. D., & Inc, M. & E. (2003). *Wastewater Engineering: Treatment and Reuse*. McGraw-Hill Education.

The Business Standard. (2023, August 28). *New sewerage project proposed in Chattogram*. The Business Standard. https://www.tbsnews.net/bangladesh/new-sewerage-project-proposed-chattogram-690490

UPM. (2021). *Sampling for faecal sludge and other liquid... | Documents | Gates Open Research*. https://gatesopenresearch.org/documents/5-27

Vaughan, G. M., & Holder, G. A. (1984). Substrate Removal in the Trickling Filter Process. *Journal (Water Pollution Control Federation)*, *56*(5), 417–423.

Vayenas, D. V., & Lyberatos, G. (1994). A novel model for nitrifying trickling filters. *Water Research*, *28*(6), 1275–1284. https://doi.org/10.1016/0043-1354(94)90292-5

Verma, A., Wei, X., & Kusiak, A. (2013). Predicting the total suspended solids in wastewater: A data-mining approach. *Engineering Applications of Artificial Intelligence*, *26*(4), 1366–1372. https://doi.org/10.1016/j.engappai.2012.08.015

Verma, S., Daverey, A., & Sharma, A. (2017). Slow sand filtration for water and wastewater treatment – a review. *Environmental Technology Reviews*, *6*(1), 47–58. https://doi.org/10.1080/21622515.2016.1278278

Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., Tysklind, M., & Souihi, N. (2021). A machine learning framework to improve effluent quality control in wastewater treatment plants. *Science of The Total Environment*, *784*, 147138. https://doi.org/10.1016/j.scitotenv.2021.147138

Wang, L., Long, F., Liao, W., & Liu, H. (2020). Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning algorithms. *Bioresource Technology*, *298*, 122495. https://doi.org/10.1016/j.biortech.2019.122495

Weijers, S. R., & Vanrolleghem, P. A. (1997). *A procedure for selecting best identifiable parameters in calibrating activated sludge model no. 1 to full-scale plant data | Water Science & Technology | IWA Publishing*. https://iwaponline.com/wst/article/36/5/69/5080/A-procedure-for-selecting-best-identifiable

Wik, T. (2003). Trickling Filters and Biofilm Reactor Modelling. *Reviews in Environmental Science and Bio/Technology*, *2*, 193–212. https://doi.org/10.1023/B:RESB.0000040470.48460.bb

Wood, M. G., Howes, T., Keller, J., & Johns, M. R. (1998). Two dimensional computational fluid dynamic models for waste stabilisation ponds. *Water Research*, *32*(3), 958–963. https://doi.org/10.1016/S0043-1354(97)00316-3

Zhang, W., Li, R., Zhao, J., Wang, J., Meng, X., & Li, Q. (2023). Miss-gradient boosting regression tree: A novel approach to imputing water treatment data. *Applied Intelligence*, *53*, 1–21. https://doi.org/10.1007/s10489-023-04828-6

Zhao, L., Dai, T., Qiao, Z., Sun, P., Hao, J., & Yang, Y. (2020). Application of artificial intelligence to wastewater treatment: A bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse. *Process Safety and Environmental Protection*, *133*, 169–182. https://doi.org/10.1016/j.psep.2019.11.014

Zhao, Y., Guo, L., Liang, J., & Zhang, M. (2016). Seasonal artificial neural network model for water quality prediction via a clustering analysis method in a wastewater treatment plant of China. *Desalination and Water Treatment*, *57*(8), 3452–3465. https://doi.org/10.1080/19443994.2014.986202

Appendices

Appendix A: Different types of ANN structures

Table A1: ANN variation with different network architecture

Model Name	**Optimizer**	**Activation Function**	**Network Architecture**		
			Trial-1	**Trial-2**	**Trial-3**
ANN-I	Adam	ReLU	8-16-8-1	8-64-8-1	32-64-32-1
ANN-II	SGD	Sigmoid	8-16-32-16-8-1	8-16-64-16-8-1	128-64-32-16-8-1
ANN-III	Adam	SELU	8-16-32-64-32-16-8-1	8-16-32-128-32-16-8-1	128-16-32-64-32-16-8-1

Appendix B: Category-1: Synthetic data in association with mathematical equations and assumptions [Training (80%), Testing (20%)]

Table B1. Trial–1: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS)

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.85	0.90	0.62	0.101	0.066	0.186	0.077	0.053	0.142
ANN-II	0.80	0.87	0.63	0.072	0.078	0.185	0.098	0.055	0.146
ANN-III	0.88	0.90	0.70	0.079	0.068	0.174	0.082	0.053	0.132

Table B2. Trial–2: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS)

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.82	0.80	0.67	0.173	0.108	0.178	0.16	0.153	0.135
ANN-II	0.82	0.80	0.60	0.174	0.109	0.189	0.162	0.155	0.15
ANN-III	0.81	0.80	0.70	0.181	0.109	0.174	0.164	0.154	0.131

Table B3. Trial–3: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS)

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.88	0.86	0.83	0.106	0.097	0.099	0.106	0.084	0.086
ANN-II	0.88	0.85	0.62	0.103	0.099	0.186	0.105	0.085	0.147
ANN-III	0.88	0.85	0.69	0.108	0.099	0.175	0.102	0.083	0.132

Appendix C: Category-2 (Training: Synthetic data of Category-1 and Testing: Real wastewater data)

Table C1. Trial–1: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS)

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.75	0.86	0.86	0.115	0.088	0.087	0.089	0.071	0.038
ANN-II	0.74	0.83	0.83	0.115	0.096	0.085	0.09	0.079	0.036
ANN-III	0.73	0.82	0.86	0.118	0.1	0.088	0.087	0.081	0.038

Table C2. Trial–2: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS)

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.75	0.84	0.86	0.118	0.094	0.087	0.082	0.075	0.038
ANN-II	0.75	0.81	0.83	0.114	0.102	0.085	0.09	0.088	0.036
ANN-III	0.75	0.83	0.86	0.114	0.098	0.088	0.086	0.079	0.038

Table C3. Trial–3: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS)

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.76	0.82	0.60	0.111	0.1	0.198	0.081	0.08	0.147
ANN-II	0.76	0.84	0.63	0.112	0.094	0.186	0.086	0.079	0.136
ANN-III	0.77	0.84	0.64	0.111	0.094	0.186	0.085	0.076	0.136

Appendix D: Category-3: Seasonal variation based performance

Table D1. Trial–1: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Dry Real (80/20)]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.82	0.84	0.86	0.104	0.119	0.103	0.084	0.096	0.083
ANN-II	0.67	0.81	0.43	0.241	0.123	0.202	0.213	0.087	0.184
ANN-III	0.83	0.86	0.46	0.088	0.109	0.201	0.089	0.083	0.156

Table D2. Trial–2: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Dry Real (80/20)]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.43	0.67	0.88	0.25	0.187	0.094	0.212	0.109	0.069
ANN-II	0.66	0.64	0.58	0.209	0.192	0.296	0.182	0.159	0.251
ANN-III	0.70	0.68	0.43	0.185	0.124	0.206	0.176	0.181	0.181

Table D3. Trial–3: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Dry Real (80/20)]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.87	0.82	0.70	0.107	0.084	0.15	0.036	0.069	0.124
ANN-II	0.85	0.54	0.67	0.137	0.292	0.295	0.306	0.259	0.249
ANN-III	0.85	0.51	0.43	0.134	0.301	0.205	0.06	0.246	0.17

Table D4. Trial–1: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Wet Real (80/20)]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.89	0.84	0.86	0.129	0.119	0.103	0.117	0.096	0.083
ANN-II	0.81	0.81	0.83	0.145	0.123	0.92	0.357	0.107	0.74
ANN-III	0.80	0.86	0.46	0.1217	0.109	0.201	0.109	0.083	0.156

Table D5. Trial–2: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Wet Real (80/20)]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.86	0.83	0.89	0.128	0.096	0.122	0.107	0.054	0.093
ANN-II	0.80	0.63	0.75	0.387	0.231	0.261	0.064	0.189	0.325
ANN-III	0.81	0.47	0.63	0.344	0.243	0.226	0.269	0.209	0.203

Table D6. Trial–3: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Wet Real (80/20)]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.87	0.86	0.86	0.057	0.174	0.14	0.036	0.111	0.104
ANN-II	0.85	0.87	0.72	0.107	0.172	0.204	0.106	0.109	0.188
ANN-III	0.85	0.81	0.86	0.094	0.136	0.148	0.86	0.167	0.161

Table D7. Trial–1: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Dry Wet (80/20)]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.79	0.66	0.85	0.224	0.264	0.098	0.184	0.225	0.076
ANN-II	0.78	0.81	0.80	0.275	0.153	0.175	0.134	0.117	0.137
ANN-III	0.79	0.62	0.86	0.244	0.258	0.095	0.197	0.203	0.085

Table D8. Trial–2: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Dry Wet (80/20)]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.74	0.36	0.60	0.185	0.284	0.16	0.148	0.225	0.132
ANN-II	0.80	0.51	0.57	0.136	0.253	0.172	0.133	0.217	0.136
ANN-III	0.90	0.52	0.61	0.113	0.258	0.167	0.093	0.213	0.154

Table D9. Trial–3: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Dry Wet (80/20)]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.79	0.84	0.95	0.205	0.087	0.058	0.153	0.67	0.049
ANN-II	0.81	0.81	0.57	0.189	0.156	0.272	0.129	0.125	0.234
ANN-III	0.85	0.87	0.60	0.187	0.084	0.239	0.121	0.063	0.211

Table D10. Trial–1: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: Dry, Testing: Wet]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.86	0.80	0.45	0.138	0.124	0.283	0.109	0.093	0.243
ANN-II	0.46	0.74	0.45	0.298	0.125	0.296	0.261	0.107	0.257
ANN-III	0.62	0.58	0.47	0.171	0.297	0.268	0.138	0.252	0.212

Table D11. Trial–2: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: Dry, Testing: Wet]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.55	0.80	0.76	0.33	0.124	0.166	0.3	0.093	0.119
ANN-II	0.63	0.74	0.68	0.28	0.105	0.287	0.26	0.21	0.262
ANN-III	0.67	0.68	0.46	0.22	0.197	0.334	0.186	0.152	0.271

Table D12. Trial–3: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: Dry, Testing: Wet]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.42	0.61	0.65	0.29	0.276	0.283	0.261	0.223	0.259
ANN-II	0.66	0.71	0.58	0.268	0.182	0.312	0.221	0.264	0.266
ANN-III	0.62	0.79	0.45	0.28	0.212	0.309	0.238	0.26	0.256

Table D13. Trial–1: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: Wet, Testing: Dry]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.89	0.63	0.76	0.081	0.247	0.142	0.064	0.208	0.109
ANN-II	0.90	0.47	0.72	0.078	0.274	0.295	0.063	0.218	0.252
ANN-III	0.70	0.49	0.86	0.134	0.257	0.081	0.095	0.2	0.078

Table D14. Trial–2: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: Wet, Testing: Dry]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.69	0.65	0.84	0.13	0.169	0.149	0.102	0.291	0.102
ANN-II	0.88	0.72	0.75	0.119	0.127	0.169	0.092	0.267	0.16
ANN-III	0.80	0.86	0.82	0.144	0.094	0.153	0.095	0.065	0.108

Table D15. Trial–3: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: Wet, Testing: Dry]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.79	0.65	0.80	0.151	0.189	0.13	0.164	0.131	0.102
ANN-II	0.82	0.68	0.72	0.113	0.184	0.149	0.187	0.125	0.139
ANN-III	0.73	0.69	0.53	0.135	0.178	0.199	0.197	0.123	0.155

Appendix E: Category-4: Synthetic data from real wastewater data

Table E1. Trial–1: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Dry Wet Synthetic (80/20)]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.65	0.87	0.41	0.175	0.035	0.283	0.16	0.027	0.273
ANN-II	0.51	0.83	0.50	0.137	0.169	0.157	0.112	0.156	0.144
ANN-III	0.70	0.79	0.49	0.132	0.132	0.212	0.113	0.121	0.203

Table E2. Trial–2: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Dry Wet Synthetic (80/20)]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.84	0.72	0.61	0.027	0.144	0.183	0.022	0.128	0.173
ANN-II	0.54	0.83	0.62	0.139	0.137	0.132	0.114	0.119	0.113
ANN-III	0.77	0.78	0.77	0.149	0.046	0.179	0.123	0.034	0.171

Table E3. Trial–3: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Dry Wet Synthetic (80/20)]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.85	0.66	0.55	0.066	0.236	0.12	0.061	0.219	0.11
ANN-II	0.66	0.38	0.68	0.084	0.153	0.134	0.081	0.136	0.117
ANN-III	0.68	0.12	0.51	0.083	0.144	0.073	0.08	0.117	0.06

Table E4. Trial–1: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: Dry Synthetic, Testing: Dry Real]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.53	0.54	0.53	0.165	0.269	0.319	0.137	0.223	0.253
ANN-II	0.48	0.54	0.79	0.249	0.247	0.261	0.209	0.212	0.196
ANN-III	0.73	0.93	0.64	0.125	0.064	0.293	0.099	0.049	0.231

Table E5. Trial–2: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: Dry Synthetic, Testing: Dry Real]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.53	0.57	0.92	0.236	0.203	0.063	0.194	0.251	0.043
ANN-II	0.55	0.56	0.74	0.245	0.249	0.244	0.208	0.213	0.185
ANN-III	0.52	0.53	0.85	0.224	0.246	0.088	0.187	0.203	0.074

Table E6. Trial–3: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: Dry Synthetic, Testing: Dry Real]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.75	0.71	0.80	0.234	0.134	0.172	0.187	0.16	0.156
ANN-II	0.78	0.77	0.87	0.248	0.151	0.137	0.211	0.116	0.184
ANN-III	0.80	0.86	0.84	0.227	0.025	0.175	0.184	0.039	0.117

Table E7. Trial–1: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: Wet Synthetic, Testing: Wet Real]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.83	0.42	0.70	0.076	0.427	0.247	0.063	0.327	0.181
ANN-II	0.69	0.70	0.72	0.249	0.274	0.206	0.22	0.238	0.164
ANN-III	0.87	0.87	0.82	0.05	0.051	0.078	0.037	0.042	0.066

Table E8. Trial–2: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: Wet Synthetic, Testing: Wet Real]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.45	0.86	0.86	0.333	0.054	0.103	0.233	0.045	0.092
ANN-II	0.68	0.77	0.67	0.288	0.284	0.298	0.253	0.247	0.257
ANN-III	0.84	0.87	0.89	0.07	0.051	0.092	0.058	0.035	0.084

Table E9. Trial–3: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: Wet Synthetic, Testing: Wet Real]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.85	0.92	0.87	0.061	0.043	0.049	0.045	0.035	0.041
ANN-II	0.66	0.64	0.68	0.286	0.279	0.301	0.251	0.244	0.259
ANN-III	0.86	0.91	0.72	0.054	0.038	0.079	0.048	0.032	0.07

Table E10. Trial–1: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: DryWet Synthetic, Testing: DryWet Real]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.87	0.83	0.88	0.024	0.094	0.03	0.031	0.017	0.023
ANN-II	0.74	0.90	0.83	0.233	0.022	0.096	0.2	0.013	0.057
ANN-III	0.89	0.89	0.86	0.022	0.019	0.036	0.016	0.015	0.025

Table E11. Trial–2: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: DryWet Synthetic, Testing: DryWet Real]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.89	0.89	0.89	0.017	0.023	0.023	0.014	0.017	0.016
ANN-II	0.62	0.81	0.86	0.23	0.133	0.187	0.201	0.212	0.153
ANN-III	0.89	0.89	0.87	0.019	0.017	0.032	0.015	0.013	0.021

Table E12. Trial–3: Performance evaluation of DL for wastewater effluent (BOD_5, COD, TSS) [Training: DryWet Synthetic, Testing: DryWet Real]

Model	R^2			RMSE			MAE		
	BOD_5	COD	TSS	BOD_5	COD	TSS	BOD_5	COD	TSS
ANN-I	0.55	0.89	0.57	0.378	0.017	0.214	0.302	0.013	0.166
ANN-II	0.64	0.51	0.72	0.224	0.233	0.192	0.195	0.202	0.156
ANN-III	0.89	0.90	0.89	0.017	0.016	0.012	0.012	0.012	0.009