Development of Al-Based Diagnostic Tool for Enhanced Operational Control of Energy Efficient Wastewater Treatment Systems

By

Sidra Yasin

(Registration No: 00000399708)

Department of Energy Systems Engineering

US-Pakistan Center for Advanced Studies in Energy

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

(2024)

Development of AI-Based Diagnostic Tool for Enhanced Operational Control of Energy Efficient Wastewater Treatment System

By

Sidra Yasin

(Registration No: 00000399708)

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in
Energy Systems Engineering

Supervisor: Dr. Abeera Ayaz Ansari

US-Pakistan Center for Advanced Studies in Energy

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

(2024)

THESIS ACCEPTANCE CERTIFICATE

Certified that the final copy of MS/MPhil thesis written by Ms. Sidra Yasin (Registration No. 00000399708), of ESE, USPCASE has been vetted by undersigned, found complete in all respects as per NUST Statues/Regulations, is within the similarity indices limit and is accepted as partial fulfillment for the award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: ______________________

Name of Supervisor: Dr. Abeera Ayaz Ansari

Date: 26/09/2024

Signature (HoD): ______________________

Date: 27/09/2024

Signature (Dean/Principal): ______________________

Date: 27/09/2024

TH-4 form

FORM TH-4

National University of Sciences & Technology

MASTER'S THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by (Student Name & Regn No.) **Sidra Yasin ; 00000399708**

Titled: **Development of AI-Based Diagnostic Tool for Enhanced Operational Control of Energy Efficient Waste Water Treatment System** be accepted in partial fulfillment of the requirements for the award of **MS Energy Systems Engineering** degree with (A) grade.

Examination Committee Members

1. Name: Engr. Abdul Kashif Janjua — Signature:
2. Name: Dr. Mustafa Anwar — Signature:
3. Name: Dr. Muhammad Hassan — Signature:

Supervisor's name: Dr. Abeera Ayaz Ansari — Signature:

Date: 06/09/2024

Prof. Dr. Naseem Iqbal
Head of Department

Signature

27/09/2024
Date

COUNTERSIGNED

Date: 27/09/2024

Dean/Principal

US-Pakistan Center for Advanced Studies in Energy (USPCASE)

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in this thesis, entitled "Development of Al-Based Diagnostic Tool for Enhanced Operational Control of Energy Efficient Wastewater Treatment Systems" was conducted by Mr./Ms. Sidra Yasin under the supervision of Dr. Abeera Ayaz Ansari. No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the US-Pakistan Center of Advanced Studies in Energy (USPCASE) in partial fulfillment of the requirements for the degree of Master of Science in Field of Energy Systems Engineering, Department of USPCASE National University of Sciences and Technology, Islamabad.

Student Name: Sidra Yasin Signature: ____________

Examination Committee:

a) GEC Member 1: Engr. Abdul Kashif Janjua Signature: ____________

(Lecturer, National Skills University)

....

b) GEC Member 2: Dr. Muhammad Hassan Signature: ____________

(Associate Professor, ESE, USPCAS-E)

....

b) GEC Member 3: Dr. Mustafa Anwar Signature: ____________

(Assistant Professor, ESE, USPCAS-E)

....

Supervisor Name: Dr. Abeera Ayaz Ansari Signature: ____________

Name of HOD: Dr. Naseem Iqbal Signature: ____________

Name of Principal/Dean: Dr. Adeel Waqas Signature: ____________

AUTHOR'S DECLARATION

I Sidra Yasin hereby state that my MS thesis titled "Development of AI-Based Diagnostic Tool for Enhanced Operational Control of Energy Efficient Wastewater Treatment System" is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: SIDRA YASIN

Date: 27-09-2024

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled "Development of AI-Based Diagnostic Tool for Enhanced Operational Control of Energy Efficient Wastewater Treatment System" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and National University of Sciences and Technology (NUST), Islamabad towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the University reserves the rights to withdraw/revoke my MS degree and that HEC and NUST, Islamabad has the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized thesis.

Student Signature: ____________

Name: SIDRA YASIN

DEDICATION

This thesis is dedicated to my parents, for their endless support and encouragement throughout my academic journey.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to Allah for His guidance and blessings throughout this journey.

I would also like to thank my supervisor, whose guidance, patience, and insightful feedback have been invaluable. Your mentorship has significantly contributed to the successful completion of this thesis.

I am deeply grateful to my colleagues for their collaboration, support, and the stimulating discussions that enriched this research

Finally, I would like to acknowledge all those who have directly or indirectly supported me in this journey. Your encouragement has been greatly appreciated.

Thank you all for your support and contributions.

TABLE OF CONTENTS

LIST OF TABLES

Page No.

LIST OF FIGURES

Page No.

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

OPG	Oxygenated Photogranules
SVI_{30}	Sludge volume index at 30 minutes
XGBOOST	Extreme gradient boosting
KNN	K nearest neighbors
WWTPS	Wastewater treatment plants
AGS	Activated granules sludge
SVI	Sludge volume index
Scod	Soluble chemical oxygen demand
Tcod	Total chemical oxygen demand
HRT	Hydraulic retention time
SRT	Solid retention time
MLSS	Mixed liquor suspended solids
VSS	Volatile suspended solids
TSS	Total suspended solids
Svss	Settleable volatile suspended solids

Evss	Effluent volatile suspended solids
Cum Biomass	Cumulative biomass
Cum COD	Cumulative COD
Oper Day	Operation day
Inf	Influent
Eff1, eff2	Effluent 1 & 2
ML	Machine learning

ABSTRACT

Biological wastewater treatment is an established technique to treat industrial and municipal wastewater, which degrades pollutants through the actions of microorganisms. The primary challenge with current biological wastewater treatment is the need for external aeration or supply of O_2, which is required for the oxidation of organic matter and nitrification processes. Oxygenic photogranulation (OPG) is an aeration-free biological wastewater treatment in which dense photogranules are formed and characterized by high settling velocities. However, the scale-up of OPG-based wastewater treatment systems poses significant issues due to dynamic and complex system variables, which have non-linear interactions, making troubleshooting an expensive endeavour. To solve these issues, machine learning models are effective in simulating the wastewater treatment process, as mechanistic models are computationally expensive and interactions between input and output features are not well understood because of non-linearity. This study investigates the two-stage feature selection method to enhance the prediction performance of SVI30, an operational parameter that ensures the settleability of biomass and minimizes the loss of photogranules. The two-stage feature selection method identifies the relevant subset of input features to predict SVI30, thus enhancing the accuracy and performance of machine learning models. The optimal feature subsets generated by two-stage features are evaluated by four regression models: decision tree, random forest, gradient boosting, and XGBoost. The performance efficiency of all regression models is evaluated by an evaluation matrix. The regression models with optimal subsets of features identified by two-stage feature selection demonstrate a prediction efficiency of 85%. This research provides a comprehensive machine learning-based approach that can improve the predictability and control of operational parameters for an efficient OPG wastewater process. Advanced feature selection methods can significantly enhance the performance of machine learning models in OPG-based systems, leading to more sustainable wastewater management solutions.

Keywords: Machine Learning, Data-Driven Modelling, Feature Selection, Oxygenated Photogranules, Sludge Volume Index, Wastewater treatment

CHAPTER 1 : INTRODUCTION

1.1 Conventional Methods For Wastewater Treatment

The advancement in industrialization and urbanization has resulted in a substantial increase in wastewater quantities. The treatment of wastewater has become a viable solution at a global level to ensure public health and environmental protection as wastewater contains organic matter, pathogens, nutrients, and toxic substances that can contaminate receiving water bodies and may cause outbreaks of disease. Effective wastewater treatment helps mitigate these challenges by eliminating these pollutants. treating wastewater ensures the protection of ecosystems and public health by preventing the discharge of pollutants into natural water resources. Biological wastewater treatment is an established method to treat industrial and municipal wastewater, which degrades pollutants through the actions of microorganisms as it is designed to degrade pollutants in wastewater by the action of microorganisms and they utilize these pollutants as nutrients to live and reproduce.

Conventional activated sludge process has been around for decades to treat sewage or industrial wastewater. This widely used process faces sustainability issues due to some associated challenges such as substantial energy demand. This process requires an aeration tank where oxygen or air is injected into wastewater to remove carbonaceous pollutants. Notably this external aeration consumes up to 60% of total energy in WWTPs [1] The activated sludge itself does not involve direct GHG emissions but there is potential for indirect GHG emissions if energy being utilized in the process generated from the fossil fuel resources could be the contribution in the environment. Conventional AGS process exhibits a lower sludge retention time poor settleability and poor contractibility which is a significant drawback of traditional AGS systems [2] . Hence, there is a requirement for an efficient wastewater treatment system that can sustainably address these challenges.

1.2 Oxidation Photogranules for Wastewater Treatment

OPGs based wastewater treatment is a novel biotechnology, that emerged as an attractive alternative to the activated sludge process. OPGs are bio-aggregates that comprise of phototrophic microorganisms that surround heterotrophic bacteria in a dense, spherical structure. They are produced from the transformation of activated sludge under an illumination source during hydrostatic [3] or hydrodynamic cultivation environments [4]. OPGs can leverage the photosynthetically produced oxygen to treat wastewater [5], [6] which eliminates the need for mechanical energy required in aeration [6], [7]. The produced biomass is also three times higher than the conventional activated sludge (CAS) system because the photoautotrophic assimilation of CO_2 potentially reduces the emission of greenhouse gases than the activated sludge process [9]. Moreover, harvested OPG biomass can be utilized as an organic-rich feedstock for a renewable energy resource [7].

These phototrophic granules have higher density and settleability than other types of microbial biomass reducing the risk of biomass washout and enhancing the clarity of effluent[11][12]. The process of dewatering and harvesting biomass is an energy-intensive process, however, multiple factors aid in reduced energy requirements as increased density and higher settling velocity reduce the energy input of harvesting biomass. furthermore, it gives a high energy yield from biomass with a higher energy content [9] . Compact and granular dense biomass typically exhibit a favorable sludge volume index (SVI) indicative of good compaction, associated with parameters such as settling velocity, density, particle size distribution, porosity, and permeability.

Despite OPG-based wastewater treatment being promising at the laboratory scale, the attempts to scale up this technology have raised lots of issues including loss of granular biomass, decline in treatment performance, and subsequent loss of reactor functionality [9]. Photogranule, the core component of the OPG-based wastewater treatment process, need to maintain their structural integrity and settling properties for efficient wastewater treatment and biomass handling. Various studies have been conducted to determine the factors responsible for the promotion of photo granulation including mixing speed [13], hydraulic retention time (HRT) [14], EPS production [15], [16], seeding density [17], light intensity and Iron [18]. The sludge volume index (SVI) is a critical parameter determining biomass settleability and overall biological WWT system performance [19]. While SVI has been determined in early OPG studies, its detailed relationship with other operational and treatment parameters has not been extensively covered yet, presenting a gap in OPG literature.

An effective understanding of the dynamic relationship between SVI and OPG system parameters warrants a shift from traditional mechanistic models to data-driven techniques. This is because the OPG-based wastewater treatment process involves dynamic interactions between microorganisms and influent wastewater conditions and mechanistic models are often ineffective in predicting and optimizing complex processes [20], [21]. Moreover, the development of non-linear models to justify this level of interaction requires extensive experimental work which can be costly and time-consuming [22] In contrast, data-driven models leverage computational algorithms to identify patterns, correlations, and trends within datasets [23], [24]. These models can estimate the nonlinear dynamics between input and target variables without requiring a comprehensive understanding of the physical and chemical mechanisms of the process [25], [26].

In this thesis, the application of machine learning models to simulate an opg-based bioreactor to treat wastewater is investigated to predict the settling characteristic of biomass (SVI_{30}) by using different regression algorithms including decision trees, random forests, gradient boosting, and XGBoost. The advanced two-stage feature selection method is developed to find the subset of features which are important to predict SVI_{30}. For the OPG reactor modelling dataset of lab scale is used.

1.3 Research Objectives

This thesis aimed to simulate the OPG-based wastewater treatment system using Machine Learning algorithms and Machine Learning models developed using a small dataset.

- To gather, prepare, and select relevant features of secondary data of OPG wastewater treatment system.
- Apply the acquired dataset to develop AI-model for performance prediction and diagnosis of the OPG system.
- To develop a comprehensive predictive model by utilising the acquired dataset to accurately forecast biomass's settling characteristics, the Sludge Volume Index at 30 minutes (SVI30).
- Implement a two-stage feature selection (FS) method to identify the optimal combination of features for each regression model.

1.4 Problem Statement

- Fine-tuning of parameters through experiments is time-consuming and needs resources, modern AI-based diagnostic tool for the development of the OPG WWT system and performance prediction may solve the issue.

The arrangement of this thesis is as follows: Section 2 presents the literature review, In Section 3 explains the proposed methodology for data-driven modelling of OPG reactor for wastewater treatment. Section 4 presents results and related discussion for the implementation of machine learning models for OPG reactor. Section 5 is a concluding remark about this study and then future recommendations

CHAPTER 2 : LITERATURE REVIEW

2.1 Mechanistic Modelling of Wastewater Treatment

To meet waste treatment discharge criteria, it is critical to regulate and optimize wastewater treatment process variables [24]. The dynamics of wastewater reactors are non-linear and complex because of the interaction between the process variables and variation in terms of composition, concentration, and flow rate of influent wastewater treatment plants [25] . With the advent of Activated Sludge Models (ASM) models wastewater modelling gained significant importance in 1980. These kinetic models are based on the first principle, which is the differential equations that mathematically describe various biological processes occurring in wastewater treatment reactors. However, controlling the biological wastewater treatment process based on Kinetic models is difficult because of frequent calibration and validation. The mathematical modelling of wastewater treatment comprised of differential equations is challenging due to a large number of control parameters[26]. For reliable control and optimisation of the process, accurate modelling of key parameters is required. Modelling of complex and dynamic systems is computationally challenging as solving mathematical equations often requires optimisation algorithms and high-performance computing for the simulation and design of processes [27].

2.2 Data-driven models for wastewater treatment methods

In contrast to mechanistic models based on physical and chemical principles, data-driven models, particularly machine learning approaches, present a paradigm shift by leveraging computational algorithms to identify patterns, correlations, and trends within datasets[23], [24]. Machine learning models estimate the nonlinear dynamics between input and target variables without requiring a comprehensive understanding of the physical and chemical mechanism of the process.[31], [32]. Physical models are better for understanding factors that affect the process performance, kinetics, and conversion rates of the process but machine learning models are an excellent tool kit for predicting the reactor performance[33], [34], and optimization of a process[35]. These models overcome the need for time-consuming experiments and continuous re-calibration of physical models as they are adaptive enabling them to consistently acquire and extract information from newly collected data as the process continues [36].

ML-based models exceptionally the ability to learn the complex nonlinear relationships and dynamics of biological wastewater treatment. therefore, ML-based technologies can predict the water quality [28]. Ensemble machine learning models are employed to predict the level of COD, TDS and BOD5 of effluent wastewater using seven inputs. An Ensemble Machine learning model has been built to predict the performance of an activated granule sludge reactor. The developed model successfully predicts the SVI_{30}, SVI_5, effluent COD, granular size, MLSS, MLVSS, NH4-N, and PO43− and achieves performance accuracy with average MAE, RMSE, and R-square of 3.7%, 0.032% and 95.7% respectively [29]. The ANN-based model has been developed to forecast the removal of Chemical Oxygen Demand at three different temperatures (30, 40, and 50 degrees Celsius) by using experimental data and the model achieved the R-square values of 91%, 91%, and 89.6% at their respective temperatures [30]. ANN and SVR models are developed to compare to predict the ammonia and total nitrogen levels in effluent. The performance of the SVR model surpassed the ANN demonstrating R-squared for NH3 being 90% and 80.5%, respectively, and for (T-N)99.5% and 95.7%, respectively [31]. Aside from performance prediction, machine learning models are employed to adjust the operating parameters of wastewater treatment to enhance the effluent quality[32]. A hybrid machine learning model is developed for the identification of optimal setpoints of controllers to enhance the performance of wastewater treatment plants under fluctuating influent conditions[33]

2.3 Importance of Feature Selection in machine learning models

Addressing the significance of machine learning models in wastewater treatment, the emphasis shifts to the challenge provided by high dimensional data sets, highlighting the need for feature selection for optimal model performance. Conventionally, an extensive data set is required to train machine learning models. These high-dimensional data sets may contain non-informative, duplicate, or redundant features, posing a challenge for learning algorithms and increasing the model complexity. It is important to reduce the model complexity by removing irrelevant and redundant features. The process of feature selection is employed to reduce the dimensions of data which decreases the complexity and training time of the model. In the high dimensional data set, selecting the suitable subset is difficult as search space expands exponentially when the number of features increases [34], [35] . Selecting suitable features could enhance the prediction accuracy of total nitrogen (TN) in wastewater treatment processes by up to 20% [36]. Three different machine learning (ML) based models are developed to predict and identification of key factors that affect the performance of a ZVI-based anaerobic digestion reactor without time-consuming experiments and calculations [37]. A machine learning model is developed to predict the production of sewage sludge with increased Prediction accuracy of up to 40% by using selected features obtained via mutual information and a co-relation matrix [38] . Sludge bulking negatively affects the biomass settling and causes operational challenges in wastewater treatment. Feature importance methods help to identify process operating variables that control the sludge bulking[39]. Wastewater treatment is an energy-intensive process. Random forest and XGBoost Feature selection methods are applied to understand factors that affect energy consumption [40] . Feature selection significantly improves the performance of machine learning models by strategically identifying and selecting important features while removing the irrelevant features which makes the machine learning model interpretable and explainable. This process reduces the model complexity, leading to faster training and minimizing the use of computational resources.

2.4 Machine Learning Modelling of OPG Reactor

The Oxygenic photo granules-based wastewater treatment technology is relatively new, and the machine learning modelling approach is not readily available in the literature as compared to established methods using activated sludge technology and aerobic granular sludge. To understand the process dynamics of OPG reactors, Kinetic models are suitable to explain the complexity and behaviors of biological reactions however they lack generalization. Kinetic models need re-calibration for different types of reactors, environmental conditions and types of wastewater as these kinetic parameters are not transferable to another set of conditions that are obtained from specific conditions. This requires a new design of experiments and accurate estimation of parameters reflecting new conditions which are both time-consuming and costly. Therefore, this requires the need for another type of modelling approach.

This raises the need to develop a generalized and adaptive model that can stimulate the OPG-based wastewater treatment process. Data-driven models are good alternatives to the mathematical modelling approach as machine learning models learn and mimic the behaviour of a reactor or system by analyzing historical trends and using them to forecast future scenarios. Machine learning algorithms are adaptable and generalized because of their ability to update as new data is available continuously. This ongoing learning process of predictive algorithms allows them to maintain their relevance and accuracy as operational conditions change. Leveraging machine learning models reduces the need for time-consuming and costly experimentation while enhancing overall system performance.

CHAPTER 3 : METHODOLOGY

3.1 Secondary Data on Reactor Operation and Sample Analysis

To develop our ML models, the secondary dataset was obtained from a previously reported OPG-based wastewater treatment study by Gikonyo et al [50]. In brief, this study produced OPGs using 4x diluted activated sludge during eight days of inoculation under illuminated (150–210 µmol/m²-s), hydrodynamic (30 rpm) conditions. The produced biomass was harvested and sieved to obtain OPGs having a size greater than 200 µm. These OPGs were inoculated as seed in a 120 L reactor having primary wastewater effluent and overhead illumination of 413 ± 53 $\mu mol/m^2$-s at the water's surface. The authors analyzed influent and effluent samples for soluble chemical oxygen demand (sCOD), total chemical oxygen demand (tCOD), total suspended solids (TSS), and volatile suspended solids (VSS) at regular time intervals for over one year. SVI, solid retention time (SRT), and HRT were also calculated and monitored during this period. This study utilized the Reactor 1 (R1) dataset to develop an ML model for SVI prediction.

3.2 Implementation Of Machine Learning

Implementing machine learning is a multi-step systematic approach that includes data collection and preprocessing, learning, and evaluation of machine learning models. After data collection, there is preprocessing of data that provides for structuring and transforming raw data into a format suitable for machine learning models, including FS models, appropriate machine learning models are developed by using selected features, fine-tuning the models, and finally, the model performance is evaluated quantitatively.

3.3 Data Preprocessing and Imputation of Missing Values

The presence of missing values is an inevitable problem in real-time data collection due to sensors failing to record data or human error during data entry. Improper Handling of a missing value leads to inaccuracy in model performance and analysis. In this study, K-nearest neighbors (KNN) imputer is employed to handle the missing values as it preserves the original data structure because missing values are imputed by taking the weighted average of neighboring values which avoids distortion of data distribution [51], [52].

3.3.1 KNN imputer

It is a supervised machine learning algorithm to impute missing values and has parameter k which needs to be tuned to predict more accurate results. It improves the accuracy of the dataset as it fills the missing values based on the weighted average of neighbouring values. While imputation, a higher value of K, will assign more weight to the neighbours of data points and a lower value of K will give less weight to the neighbours of the data points. this means for the higher value of k, the greater number of nearby points have a high impact on the imputing values and for the lower value of K, few data points influence the imputed values. It preserves the structure of the dataset by maintaining the distribution and relationship of data. KNN imputation is sensitive to the value of K, as the inappropriate value of k either leads to too generalized or overfitting.

The presence of irrelevant features in datasets increases computational time and decreases the performance efficiency of the regressor or classifier. In this study, before imputation, redundant and less correlated features with the target variable SVI_{30} were removed by using the Pearson correlation coefficient. Originally, datasets contained 42 input features and one target variable. Eliminating features that are highly correlated with the input variable will reduce the multicollinearity issue and features that have less impact on the target variable are dropped. The remaining 32 features given in Table 3.1 have no redundant features and low-impact features with the target variable. The KNN imputer is then employed for imputation. A similar methodology has been applied in previous study [53]. This approach reduces biases towards the uncorrelated feature and selects only those features related to the target variable before estimating missing values.

Table 3.1 Statistical Properties of Dataset

Features	**Min**	**Mean**	**Max**
Oper. Day (operation day)			
INF VSS/TSS (influent)	0.47	0.85	1
Influent Tcod	73	190.07	460.7
HRT	0.25	0.71	1
Settler Volume	0.25	0.56	1
upflow velocity	10.07	16.61	40.28
Light	120.31	542.96	1483.13
Cycle	11.49	268.18	960
Waste	0	0.97	6.2
SRT	0.44	7.35	31
MLSS (Mixed Liquor Suspended Solids)	80	702.94	1756.67
VSS	80	525.4	1200
VSS/TSS	0.45	0.78	1
F/M (food to microorganism ratio)	0	0.14	1.17
Yield	0	0.18	2.42

Total Mass (g)	0	82.86	210.8
EFF VSS/TSS (EFF=effluent)	0.52	0.83	1
EFF2 TSS (EFF2=effluent 2)	0	30.35	93.3
EFF2 VSS	0	25.6	93.3
EFF2 VSS/TSS	0	0.7	1
Settled volume (ml) 5min	1	80.53	210
Settled volume (ml) 30min	1	58.11	130
SVI 5 mg	36.76	282.88	3683.54
Effluent sCOD	5.7	34.24	115.2
Removal	0	0.44	0.94
sVSS (g) (Settleable Volatile Suspended Solids)	-66.6	0.31	84.4
eVSS (g) (Effluent Volatile Suspended Solids)	0	47.36	248.6
Biomass Waste	0	1.58	14.32
Biomass Produced (g)	-66	49.49	192.27
Cum Biomass (Cum=cumulative)	0	2318.65	5346.79
tCOD (g) Consumed	-12.99	96.45	494.93
Cum COD	0	4625.53	10337.54

3.4 Feature Selection

Feature selection improves the regressor or classifier performance by selecting those attributes which are important to the target variable while leaving out the redundant attributes, it also reduces the computational cost of the learning algorithm. This is an essential goal of the FS process, retaining the groups of variables which adequately describe the target and avoiding the risk of overfitting. FS mainly consist of two key steps, first generating the subset of features and then evaluating the subset of features [54]. This process enhances model accuracy by identifying the important features to predict the target and which input variables contribute to the target variable based on ranking and quantifying feature importance.

FS is categorized into three groups (1) filter method (2) wrapper method (3) embedded method. The filter method is based on univariant feature selection, it ranks the feature and selects the top-ranked feature. It is a statistical method that does not depend on the learning model. Information Gain, Pearson correlation, and chi-square are the types of filter methods. Filter methods assume features are independent and do not consider interaction between features. It only captures the linear relationship and does not capture the nonlinearity between the input and target variables. The wrapper methods determine the importance of features and require an algorithm to evaluate the machine learning model performance towards all possible combinations of features. Backward selection, forward selection, and recursive FS are the types of wrapper methods. The Embedded method is a model-based FS and strikes a balance between computational efficiency and model base evaluation by combining filter and wrapper methods. In this method, FS is integrated into a Machine Learning algorithm. In the training step, the machine learning model determines the importance of each feature and selects the features that provide the best performance [43].

3.5 Model Development and Hyperparameter Selection

After the FS process, Decision Trees, Random Forest, Gradient Boosting, and XGBoost machine learning models were developed for each subset of features. The dataset was split into training and test sets using an 80:20 ratio. Hyperparameter tuning was then performed to optimize the regression models, aiming to strike a balance between model complexity and generalization while enhancing performance and accuracy on unseen data. This step was crucial in the development and optimization of model performance. In this study, hyperparameters were manually adjusted to find the optimal combination that improves the efficiency and effectiveness of the models. The suitable hyperparameters for each regression model are presented in Table 4.1.

3.6 Machine Learning Models

3.6.1 Decision tree

A supervised learning algorithm for both regression and classification tasks and prediction of target variables by learning simple decision rules deduced from the data attributes or features. It creates a hierarchical tree structure and the criteria for splitting datasets based on MSE and MAE. The tree structure comprises internal nodes, which represent a test on features; its branches depict the outcome of that test dataset; and at the leaf node, there is no further data splitting, and it gives the final decision. The entire dataset is recursively split, starting from root nodes and continuing until terminating specifications are achieved, including the maximum depth of the tree and the minimum samples per leaf. The decision tree provides an intuitive way to interpret the non-linearity between target and feature attributes.

3.6.2 Random Forest

It is an ensemble machine learning method as it developed multiple trees, each tree independently trains on a random subset of data and attributes and merges the output of each regression tree by taking the average. In contrast to the decision tree, a random forest built with multiple trees during the training of data, the prediction of each tree is aggregated by taking the mean output of individuals which improves the performance and accuracy of a regression model. Each tree in the forest is built from a random sample of data called bootstrapping. This randomness ensures the training of individual diverse trees on different subsets of data which reduce the correlation between trees and are less likely to overfit as decision trees. Ensemble learning works by combining the output of multiple trees thus leveraging the robustness, enhancing prediction accuracy and handling large amounts of non-linear datasets with high dimensions. However, it is computationally expensive because there are several parameters which need to be tuned carefully such as minimum samples per leaf, number of trees and their depth.

3.6.3 Extreme Gradient Boosting

An advanced machine learning algorithm used for solving the regression problem, renowned for its performance and speed is extreme gradient boosting regression. To increase overall prediction accuracy and speed performance, ensemble learning uses multiple base learners. This algorithm starts with the initial prediction and recursively adds trees for the residual prediction. The final prediction made by the algorithm based on the combined prediction of each subsequent tree focuses on the error made by its previous tree. The primary factor of XGBoost encompasses gradient boosting, which enhances the loss function through iteratively incorporating models that minimize residual errors, along with regularization methods that nullify overfitting by penalizing model complexity. This methodology utilizes decision trees and gradient descent optimization in a combined form to construct a robust and effective predictive model. XGBoost is distinguished by its scalability, ability to handle sparse data, and support for parallel processing, leading to superior speed and efficiency in comparison to other boosting algorithms. Nonetheless, achieving optimal performance with XGBoost necessitates meticulous tuning of hyperparameters including the learning rate, maximum tree depth, number of estimators, and regularization terms.

3.6.4 Gradient Boosting

Gradient Boosting Regression (GBR) is a robust machine learning technique utilized for regression purposes. It builds models by sequentially incorporating weak learners, usually, decision trees, to rectify the errors of the prior models. Each succeeding tree is adapted to the residuals of the combined earlier trees, focusing on sections where the model is performing inadequately. Key elements of GBR comprise weak learners, additive modelling, a learning rate controlling the contribution of each tree, a loss function for evaluating prediction precision, and regularization methods for combating overfitting. Although GBR exhibits elevated predictive accuracy and versatility in handling various data formats, fine-tuning of hyperparameters such as the number of trees, tree depth, and the learning rate is necessary to prevent overfitting.

3.7 Model Evaluation

Model evaluation quantifies the quality and performance of the Machine Learning model. The evaluation metrics are used to quantitatively measure the effectiveness of the machine learning Model and help to determine the model's ability to predict unseen data accurately [55]. The performance of machine learning models was assessed by comparing the actual and predicted values of SVI_{30}. The metrics used to evaluate the performance of the proposed machine learning models include root mean squared error (RMSE), mean absolute error (MAE), and R-squared error, which are commonly employed in regression analysis. Both RMSE and MAE were used to measure the proximity between the predicted and actual values of SVI30. R-squared was employed to assess the strength and goodness of fit for different regression models.

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{3.1}$$

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{3.2}$$

$$\text{R-squared}=1-\frac{\sum(y_i-\hat{y}_i)2}{\sum(y_i-\bar{y}))2} \tag{3.3}$$

where i = 1, 2, ...n = number of observations, y_i = actual value, $\hat{y}_i$= predicted value

To achieve the aim of this thesis to predict biomass settling characteristics, several machine learning models were employed, including decision trees, random forests, gradient boosting, and XGBoost. After the preprocessing of the dataset, a two-stage feature selection process was implemented to find the optimal combination of features for the regression models. The resulting models were able to predict biomass characteristics SVI_{30}, of oxygenated photo granules inside the reactor as it indicated the compactness and settleability of granules. ML regression models using decision trees, random forests, gradient boosting, and extreme gradient boosting, compared the effects of different combinations of features generated from two-stage feature selection to predict the SVI_{30} and find the best combination of feature and model to predict SVI_{30}. Figure 3.2 represents the steps which are followed in the development of the Machine Learning model for OPG wastewater treatment reactor

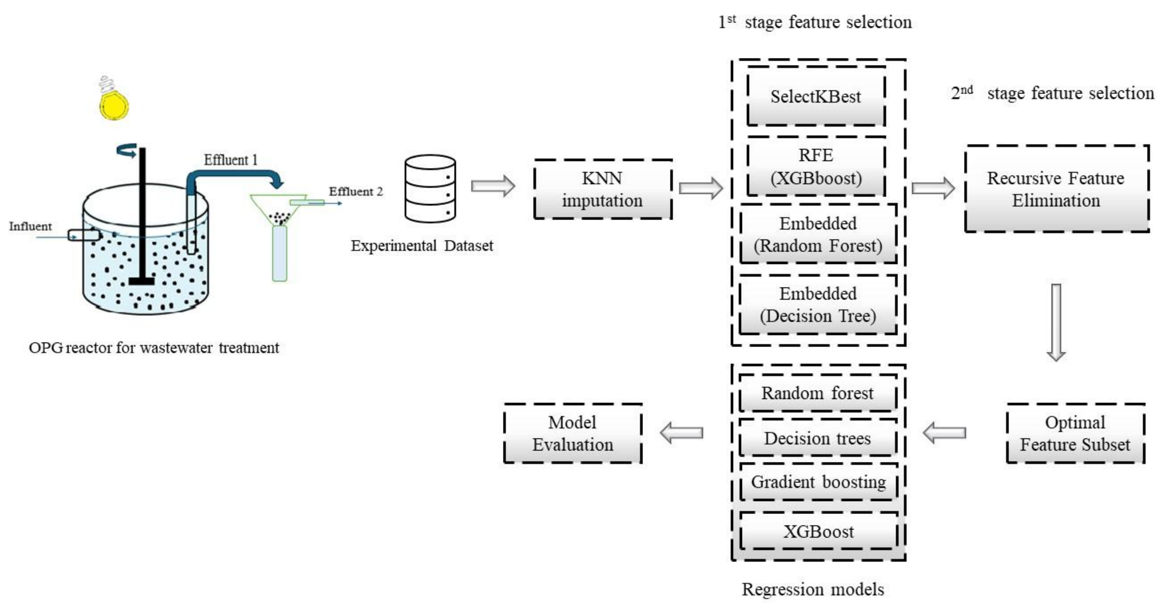

Figure 3.1: Machine learning model development steps for OPG

CHAPTER 4 : RESULTS AND DISCUSSION

4.1 KNN Imputation

After KNN imputation, imputed values are visualized by line graphs by comparing original and imputed values as ground truth is not present. The evaluation of the imputation method is challenging because there is no definite correct value present to which imputed values are compared. Figure 4.1 represents the line graphs of values before and after imputation. Visual analysis represents imputed values of variables aligned with the pattern and follows the trend of the dataset.

4.2 Two-stage Feature Selection

After handling the missing values using the KNN imputer, the two-stage FS approach was employed, to systematically reduce the dimensionality of the feature space. In the two-stage FS method, the initial subset of features is reduced in stage 1, and in stage 2 preselected feature subset is further refined by evaluating the remaining features [56]. Four FS methods independently preselected significant input features. These methods include embedded methods using Random Forest and Decision Tree, SelectKBest, and recursive FS using XGBoost. Each of these methods preselected the top 7 feature subsets. Following the preselection of features in stage 1, Recursive Feature Elimination was employed on each subset of first-stage features to retain the optimal combination of input variables for predicting the target variable.

4.3 1st Stage Feature Selection

FS served to reduce multicollinearity between input features, eliminate parameters that impaired the model performance, and identify relevant input parameters effective in predicting the SVI_{30}. 1st stage FS methods generate different subsets of features that predominantly influenced the target variable are given in Table 4.1.

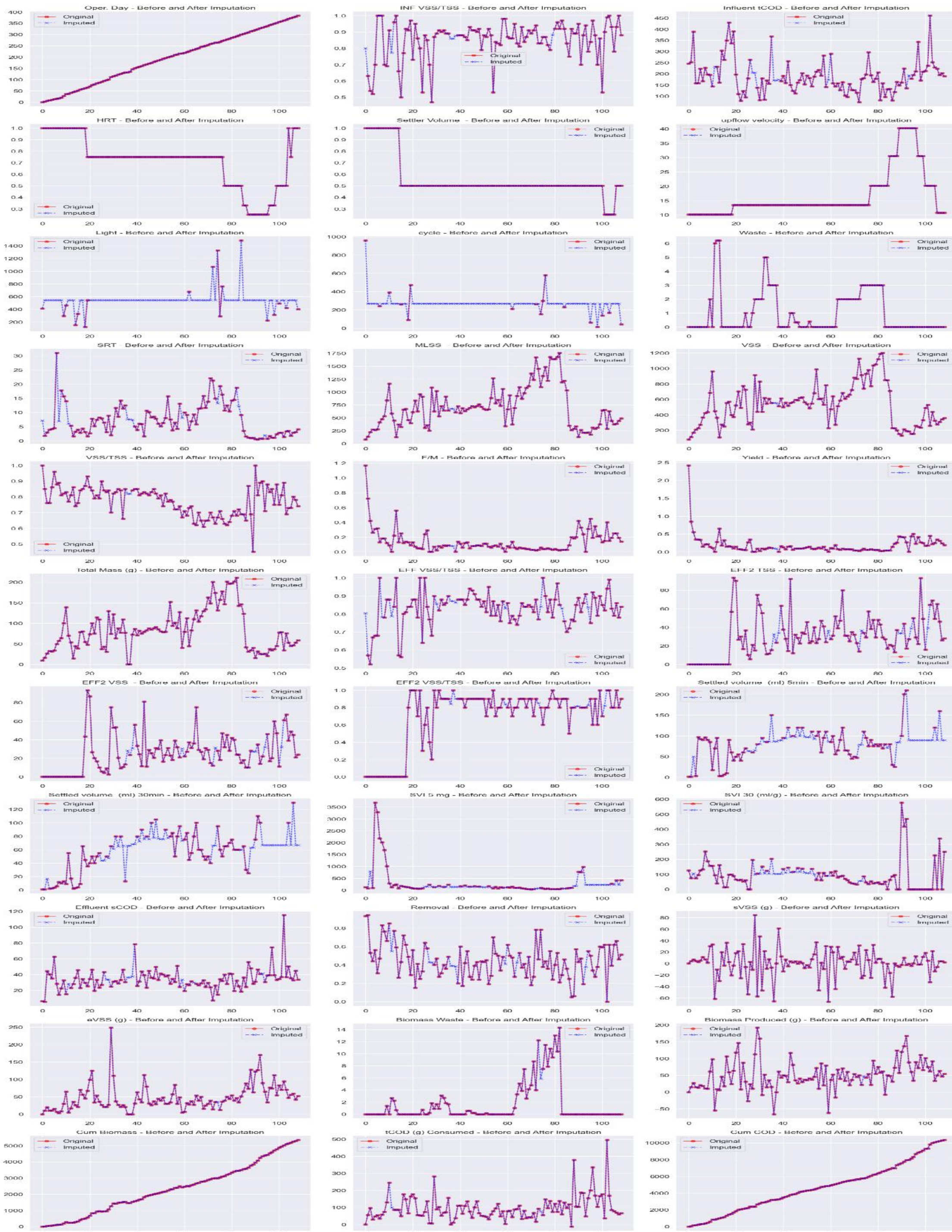

Figure 4.1: Imputation graph before and after

Table 4.1: Summary of Subsets Generated from 1st Stage and 2nd Stage Feature Selection

Feature selection method	Selected Subsets after 1st stage
Embedded method (random forest)	Oper. Day, Settled volume (ml) 5min, Settled volume (ml) 30min, SVI 5 mg, Cum Biomass, Cum COD
Embedded method (decision tree)	Oper. Day, HRT, Settled volume (ml) 30min, SVI 5 mg, Cum Biomass, Influent tCOD, Total Mass (g)
SelectKBest (f_regression)	MLSS, VSS, VSS/TSS, Total Mass (g), Settled volume (ml) 30min, SVI 5 mg, Biomass Waste
Recursive feature selection (xgboost)	Oper. Day, HRT, VSS/TSS, Settled volume (ml) 5min, Settled volume (ml) 30min, SVI 5 mg
Subsets selected by 2nd stage feature selection	
Embedded method (random forest)	Settled volume (ml) 5min, Settled volume (ml) 30min, SVI 5 mg, Cum COD
Embedded method (decision tree)	Oper. Day, HRT, Settled volume (ml) 30min, SVI 5 mg, Cum Biomass
selectKMethod (f_regression)	MLSS, VSS, Settled volume (ml) 30min, SVI 5 mg
Recursive feature selection (XGBoost)	Oper. Day, HRT, Settled volume (ml) 30min, SVI 5 mg

Figure 4.2 illustrates the important features identified through various FS methods. Feature importance analysis highlights the relationship between input features and SVI_{30}. Feature importance plots depicting the influence of each feature on the predictive performance of the Machine Learning model. Feature importance values of each feature were different for different FS methods, resulting in different feature importance rankings. Features with higher importance values have more impact in predicting the target variable and model performance than features with lower importance values. In both decision tree FS and random forest feature selection, SVI_5 exhibits the highest feature importance as shown in Figure 4.2 (a) &(b). Conversely, in the SelectKBest FS method settled volume (ml) 5 min has maximum feature importance as shown in Figure 4.2 (c), while HRT emerged as the most influential feature in recursive FS using XGBoost as shown in Figure 4.2 (d).

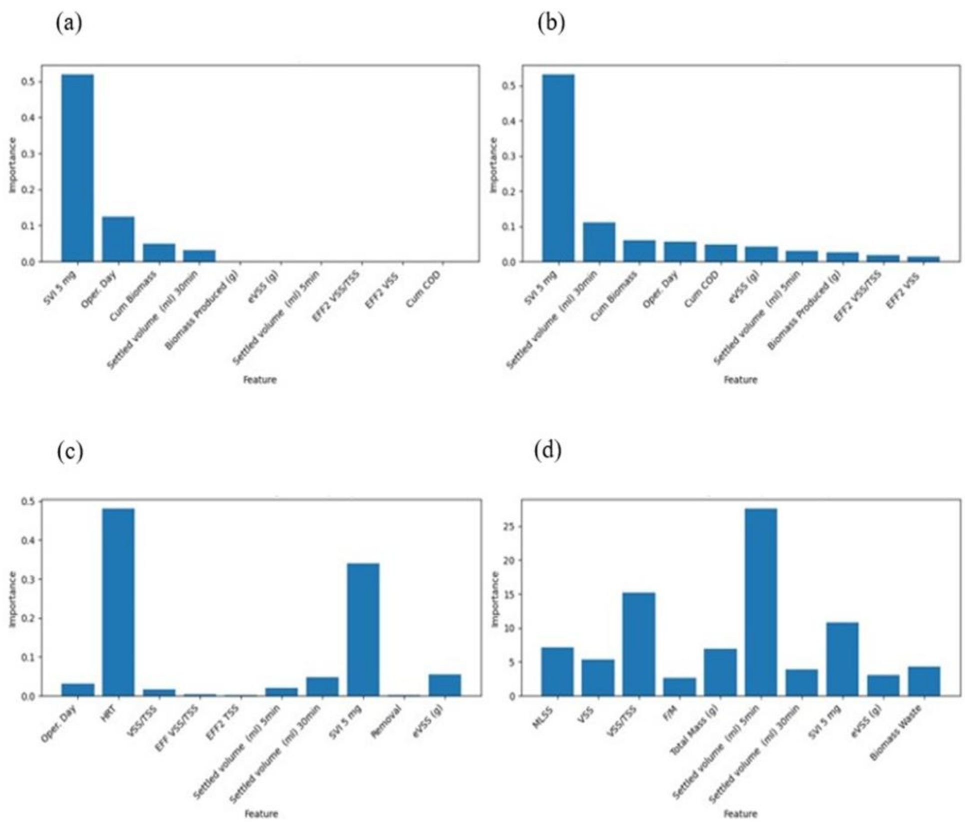

Figure 4.2: Feature Importance Plots of (a) Decision Tree, (b) Random Forest, (c) SelectKBest, and (d) XGBoost

4.4 Model Development

The decision tree, random forest, gradient boosting, and XGBoost regression models were employed to predict the $SVI_{30,}$ and all regression models were trained by using the 1st stage-selected input feature present in Table 4.1. Four models were developed for each subset of the features. Each model is trained using the training data set and adjusting the parameters of the model to optimize their performance as outlined in Table 4.2. For the SelectKBest method, the top seven features (k=7) were selected to predict SVI_{30}, as further increases or decreases in k values did not yield satisfactory performance. Similarly, for recursive FS using XGBoost, the top seven features were chosen for model development. Model performance was further evaluated by iteratively removing features to identify the best-performing set of input parameters for predicting SVI_{30}. Notably, the performance of models developed using recursive FS based on XGBoost improved upon the removal of the eVSS feature, suggesting its insignificance in predicting SVI_{30}. Models were also developed using input parameters selected by embedded methods based on decision trees and random forests, with a threshold of 0.03 and 0.05, respectively, chosen for identifying relevant features. The top seven features, as determined by random forest, were selected as important features. It was observed that adding eVSS decreased model performance, indicating its lack of significance in predicting SVI_{30}.

Table 4.2: Hyperparameters for The Regression Models

Feature Selection method	Model	Parameters
Embedded method (random forest)	Decision tree	max_depth=6, min_samples_split=4, min_samples_leaf=2, random_state=38
	Random forest	max_depth= 10,random_state=38
	Gradient boosting	n_estimators=160, learning_rate = 0.1,random_state=38
	XGBoost	learning_rate=0.08,random_state=38

Embedded method (decision tree)	Decision tree	max_depth=6, min_samples_split=4, min_samples_leaf=2, random_state=38
	Random forest	max_depth= 10,random_state=38
	Gradient boosting	learning_rate=0.1, max_depth=3, n_estimators=150,random_state=38
	XGBoost	learning_rate=0.07,random_state=38
selectKBest (f_regression)	Decision tree	max_depth=6, min_samples_split=3, min_samples_leaf=2, random_state=37
	Random forest	n_estimators=50,random_state=38
	Gradient boosting	random_state=37
	XGBoost	learning_rate=0.07,random_state=42
Recursive feature selection (XGBoost)	Decision tree	max_depth=8, min_samples_split=3, min_samples_leaf=2, random_state=37
	Random forest	max_depth=6, n_estimators=100 , random_state=38
	Gradient boosting	n_estimators=160, learning_rate = 0.1,random_state=37
	XGBoost	learning_rate=0.08,random_state=42

4.5 Model Performance

The evaluation of all four regression models for each subset of features identified by the first-stage FS method was conducted using various statistical indicators such as RMSE, MAE, and R-squared, as presented in Table 4.3.

Table 4.3: Evaluation Matrix for Each Regression Model

Feature Selection Method	Machine Learning Model	Train			Test		
		RMSE	MAE	R^2	RMSE	MAE	R^2
Embedded method (Random Forest)	Decision tree	**19.4382**	**9.3631**	**0.9607**	**17.3780**	**10.7579**	**0.8469**
	Random Forest	19.2832	7.1877	0.9614	24.2652	15.1888	0.7016
	Gradient boosting	1.1587	0.8952	0.9998	28.5139	17.6095	0.5879
	XGBoost	1.2662	0.6125	0.9999	27.8019	15.3011	0.6083
Embedded method (decision tree)	Decision tree	**17.2706**	**7.7974**	**0.9690**	**17.0632**	**10.1918**	**0.8524**
	Random Forest	19.7054	8.9734	0.9596	26.7277	16.7729	0.6379
	Gradient boosting	1.5947	1.2349	0.9997	19.6422	14.6249	0.8045
	XGBoost	2.0614	0.9692	0.9996	24.1310	14.3198	0.7049
SelectKBest (f_regression)	Decision tree	**0.9607**	**19.4363**	**0.8846**	**18.5870**	**12.3352**	**0.8249**
	Random Forest	25.4423	10.3421	0.9327	24.6873	15.2641	0.6911

	Gradient boosting	3.1736	2.5280	0.9990	19.7692	13.5815	0.8019
	XGBoost	2.0668	0.9459	0.9996	35.5989	19.0229	0.3577
Recursive feature selection (XGBoost)	Decision tree	**17.4213**	**7.7871**	**0.9685**	**17.0947**	**10.9581**	**0.8519**
	Random Forest	21.7703	9.7675	0.9507	21.1447	14.2935	0.7734
	Gradient boosting	1.3898	1.1115	0.9998	26.1534	15.6148	0.6533
	XGBoost	1.4993	0.9350	0.9998	33.5361	18.2037	0.4300

All the developed algorithms were validated using an evaluation dataset that was isolated from the training data before the development of regression models. It was found that for each feature subset generated by 1st stage FS method, the decision tree provides the best performance among all regression models in terms of R-squared. However, gradient boosting demonstrated superior performance in terms of R-squared for 1st stage FS obtained from an embedded method based on the decision tree and SelectKBest method. The performance of gradient boosting declined for feature selections obtained from the embedded method based on random forests and recursive FS based on XGBoost, indicating potential overfitting issues. Subsequently, random forest outperformed gradient boosting for feature subsets generated from the recursive FS (XGBoost) method but showed lower performance for features obtained from the embedded method (decision tree), suggesting overfitting. Furthermore, XGBoost was inefficient in predicting SVI_{30} for the subsets of 1st stage features obtained from the Embedded method (random forest), SelectKBest, and recursive FS (XGBoost) as their R-squared value indicates the inadequacy of the algorithm as it overly fits training data and performance decreased on the testing dataset.

4.6 2^{nd} Stage Feature Selection

After the preselection of input variables in the 1st stage of feature selection, Recursive Feature Elimination was employed to find the relevant features to predict the SVI_{30}. This 2^{nd} stage FS process yielded a reliable set of features for developing regression models to predict $SVI_{30.}$

The Recursive Feature Elimination method initiated with features selected in 1^{st} stage and iteratively deleted the feature that has less impact on predicting the SVI_{30} generating the subset of features that have maximum predictive accuracy. The best possible combination of features after Recursive Feature Elimination for each subset of 1^{st} stage FS is given in Table 4.1. Recursive Feature Elimination is a deterministic approach that systematically refines the feature set by removing less important features, ultimately generating an optimal subset that enhances model performance. Following the preselection of features in the first stage, four base models were developed for each preselected subset, incorporating all candidate features. This resulted in the development of 16 base models for the four preselect subsets. All base models were developed with all preselect subsets. Features were iteratively removed, and model performance was evaluated by comparing it with the base model.

4.6.1 Recursive feature elimination of features

In the case of the decision tree embedded FS method, the decision tree regression model exhibited the best performance among all regression models by utilizing all seven features listed in Table 4.4 and represented in Figure 4.3. However, when the feature Influent tCOD was removed, the R-squared value of XGBoost substantially declined to a negative number. Conversely, there was a slight improvement in the model performance of random forest, decision tree, and gradient boosting. Upon further removal of the Total Mass (g) feature, the performance of XGBoost improved to a positive R-squared value, while the performance of the decision tree remained the same. Gradient boosting performance also improved, while random forest performance slightly decreased. Removing the Cum Biomass feature resulted in a slight decrease in the performance of each model. Subsequent feature removal steps led to a decline in the performance of each model.

For the case of the random forest embedded FS method, the performance of the decision tree was best among all regression models by selecting all 6 features which were given in Table 4.4 is represented by Figure 4.3. Performance of the decision tree model for up to 3 feature elimination remains unchanged, for 3 subsets of features performance slightly decreased and with further elimination of features performance of the model decreased. For random forest, the performance of the model slightly decreased by up to 3 feature elimination. Further removal of features decreased the performance of the model. With feature SVI 5 mg, random forest performed well as compared to other regression models. The performance of the gradient boosting model improved by eliminating the two features. For the XGBoost, the performance of the model increased by eliminating up to 4 features. Further elimination of features degraded the performance.

For the case of SelectKBest, the performance of the decision tree and gradient boosting was relatively high in terms of R-squared valued by selecting all the 7 features which are given in Table 4.4 represented in Figure 4.3. The performance of both the decision tree model and gradient boosting remained approximately consistent even after feature elimination. However, the performance of both models degrades after 3 subsets of features. There was a slight change in the Performance of random forest by feature elimination. However, the XGBoost model didn’t perform well as compared to other regression models. There was no improvement in its performance with feature elimination, except for a slight improvement observed for the subset of two features.

For the case of embedded XGBoost feature selection, the decision tree demonstrated relatively high performance in terms of R-squared value by selecting all 6 features which were given in Table 4.4 and represented by Figure 4.3. Its performance remains approximately consistent even after feature elimination of up to 3 subsets of features. However, further elimination of features degraded its performance. Performance of random forest decline with feature elimination. On the other hand, The performance of both gradient boosting and XGBoost improves by feature elimination of up to 3 subsets of features. However, further elimination of features degraded the performance of both models.

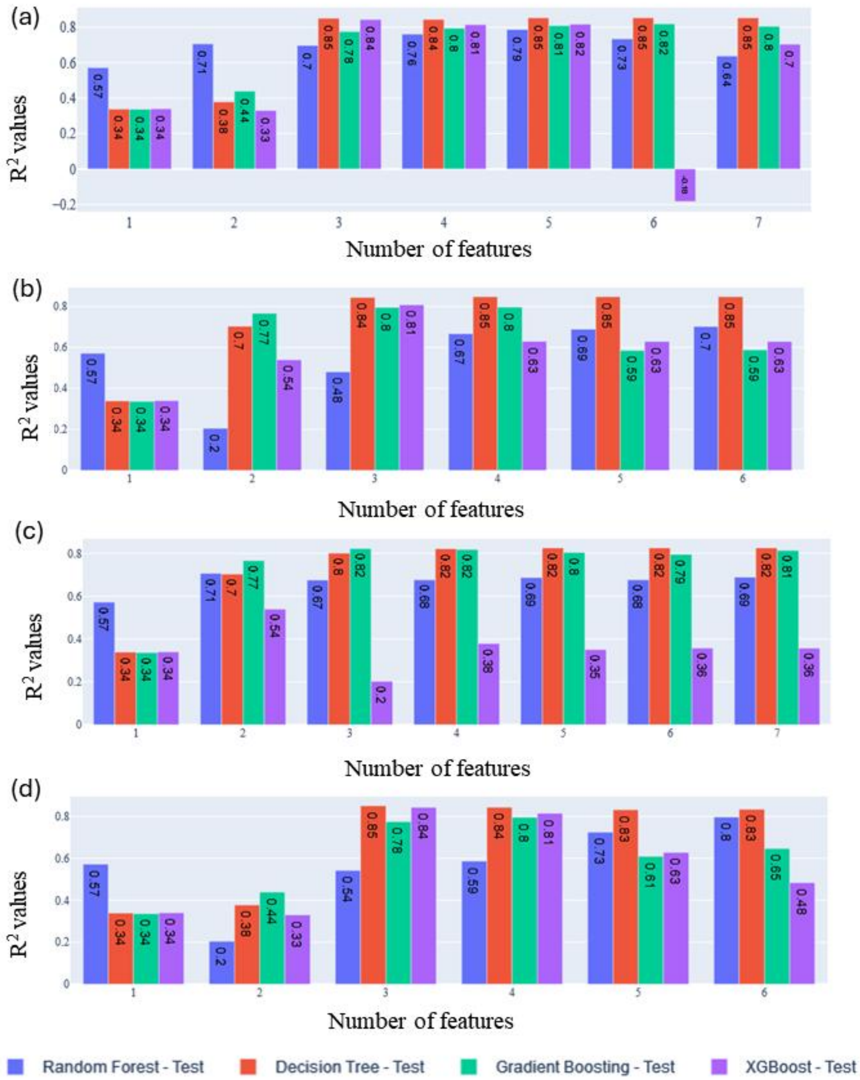

Figure 4.3: Effect on Performance of Machine Learning Models by Recursive Feature Elimination Approach is Presented through R2 Values for Different Methods (a) Decision Tree, (b) Random Forest, (c) SelectKBest, and (d) XGBoost.

Table 4.4: Feature Subsets after Recursive Feature Elimination

No of features	Embedded Method (Decision Tree)
7	Oper. Day, HRT, Settled volume (ml) 30min, SVI 5 mg, Cum Biomass, Influent tCOD, Total Mass (g)
6	Oper. Day, HRT, Settled volume (ml) 30min, SVI 5 mg, Cum Biomass, Total Mass (g)
5	Oper. Day, HRT, Settled volume (ml) 30min, SVI 5 mg, Cum Biomass,
4	HRT, Settled volume (ml) 30min, SVI 5 mg, Cum Biomass
3	HRT, Settled volume (ml) 30min, SVI 5 mg
2	Settled volume (ml) 30min, SVI 5 mg
1	SVI 5 mg
Embedded Method (Random Forest)	
6	Oper. Day, settled volume (ml) 5min, Settled volume (ml) 30min, SVI 5 mg, Cum Biomass, Cum COD
5	Settled volume (ml) 5min, Settled volume (ml) 30min, SVI 5 mg, Cum Biomass, Cum COD
4	Settled volume (ml) 5min, Settled volume (ml) 30min, SVI 5 mg, Cum COD
3	Settled volume (ml) 5min, SVI 5 mg, Cum COD
2	Settled volume (ml) 5min, SVI 5 mg
1	SVI 5 mg
SelectKBest (f_regression)	
7	MLSS, VSS, VSS/TSS, Total Mass (g), Settled volume (ml) 30min, SVI 5 mg, Biomass Waste

6	MLSS, VSS, Total Mass (g), Settled volume (ml) 30min, SVI 5 mg, Biomass Waste
5	MLSS, VSS, Settled volume (ml) 30min, SVI 5 mg, Biomass Waste
4	MLSS, VSS, Settled volume (ml) 30min, SVI 5 mg,
3	MLSS, Settled volume (ml) 30min, SVI 5 mg,
2	Settled volume (ml) 30min, SVI 5 mg,
1	SVI 5 mg
Recursive Feature Selection (Xgboost)	
6	Oper. Day, HRT, VSS/TSS, Settled volume (ml) 5min, Settled volume (ml) 30min, SVI 5 mg
5	Oper. Day, HRT, VSS/TSS, Settled volume (ml) 30min, SVI 5 mg
4	Oper. Day, HRT, Settled volume (ml) 30min, SVI 5 mg
3	Oper. Day, HRT, SVI 5 mg,
2	HRT, SVI 5 mg
1	SVI 5 mg

4.7 Summary of Key Features and Predictive Modelling of SVI_{30}

Table 4.4 summarizes the key features through two staged FSs by using various methods to demonstrate their relevance in the predictive modelling of SVI_{30}. Each 2nd stage FS method identifies SVI_5 as a significant feature. The settling volume index is utilized in conventional wastewater facility operations to provide indirect characterization of sludge physical parameters. Directly, it indicates the settleability potential of biomass providing early assessment of operational impacts. While normally undertaken at 30 min for CAS systems, granular systems such as OPG with higher settling velocities are evaluated at 5 min (SVI_5). OPG granular structure; density, size and porosity have a direct bearing on their function and settleability [50] making SVI_5 a potential robust predictor of granular moieties. These measurements are determined offline through conventional laboratory analytical methods, making them inadequate for real-time monitoring and control. [57]. Nevertheless, the rapid assessment with SVI_5 allows timely operational adjustments to prevent prolonged settling issues of biomass. Integrating significant features into ML models improves SVI_{30} prediction accuracy [58], enabling data-driven decisions for optimizing the settleability of photogranules. SVI_5 approximates SVI_{30} in well-settling granular systems with < 95 mL/g absolute SVI values. To facilitate enhanced separation of all particulate suspended biomass co-occurring with granules, a longer settling time e.g.30 min can be adopted. This optimization can reduce the risk of biomass washout, thereby enhancing the overall treatment performance and operational stability of the OPG reactor.

Oper. Day (Operational Day) indicates the stage of the treatment process and reflects sludge age and operational conditions, which influence the characteristics and settleability of the sludge. HRT (Hydraulic Retention Time) represents the duration wastewater remains in the treatment reactor, affecting treatment efficiency and sludge stabilization. The RReliefF ranking method indicates HRT is the main factor affecting SVI_{30} [59]. This study also indicates total volatile suspended solids TVSS as an important predictor for $SVI_{30.}$ Cum COD (Cumulative Chemical Oxygen Demand) consumed: Indicates the organic load consumed in the effluent of wastewater and better removal efficiency of COD is correlated with SVI [60]. MLSS (Mixed Liquor Suspended Solids) represents the concentration of suspended solids, directly impacting sludge settleability. A recent study highlights the significance of MLSS as a crucial factor for predicting the SVI_{30} in denitrifying granular sludge [61] . Cum Biomass measures the total biomass present in the WWT system. Biomass concentration in the WWT system declines when SVI has higher values as biogranules washed out with the effluent, primarily due to the poor settling velocity of the seed sludge during the initial stages of granule inoculation [62].

The identified features reflect the multifaceted nature of sludge settleability and the importance of considering both operational and chemical/biomass parameters. Operational parameters such as Oper. Day and HRT are temporal and process-related variables, influencing sludge characteristics. Chemical and biomass measures like Cum COD, Cum Biomass, MLSS, and VSS provide detailed insights into the organic and biological composition of the sludge, crucial for describing the mechanisms driving biogranules settleability.

4.8 Visualization of SVI_{30} Prediction Errors

Figure 4.4 (a) provides the visualization of error by comparing the predicted SVI30, and the true SVI30 values by selecting the optimal combination of features: Oper. Day, HRT, settled volume (ml) 30min, SVI 5 mg (Embedded Method using Decision Tree), and performance of decision tree to predict the SVI_{30} was evaluated in terms of r-squared which was 0.85. Figure 4.4 (b) presents the visualization of errors by comparing the predicted SVI_{30} and the true SVI_{30} values by selecting the optimal combination of features (embedded method using Random Forest) MLSS, VSS, Settled volume (ml) 30min, SVI 5 mg to predict the SVI_{30} and performance of decision tree was evaluated in terms of R-squared value, which was 0.84. Figure 4.4 (c) illustrates the visualization of errors by comparing the predicted SVI_{30} and the true SVI_{30} values by selecting the optimal combination of features (SelectKBest) Oper. Day, HRT, settled volume(ml) 30min, SVI 5 mg, and Cum Biomass to predict the SVI_{30} and performance of decision tree was evaluated in terms of R-squared which was 0.82. Figure 4.4 (d) presents the visualization of errors by comparing the predicted SVI_{30} and the true SVI_{30} values by selecting the optimal combination of features (Recursive feature elimination (XGBoost)) Settled volume (ml) 5min, Settled volume (ml) 30min, SVI 5 mg, Cum COD to predict the $SVI_{30,}$ the performance of Decision tree is evaluated in terms of R-squared, which was 0.84.

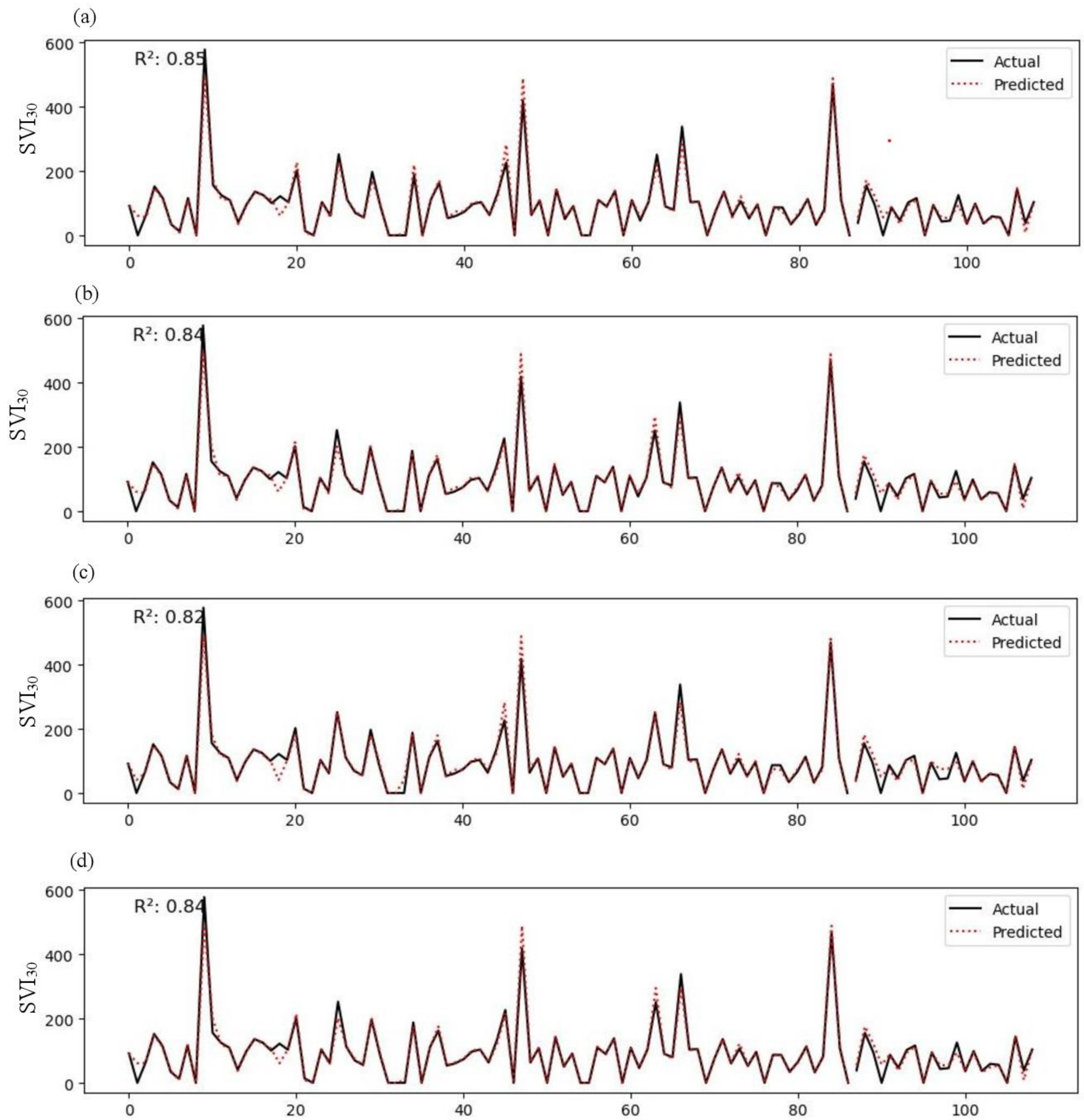

Figure 4.4: Visualization of Error by Comparing Predicted SVI30 and the true SVI30 Values by selecting the optimal combination of features after 2nd stage feature selection (a) Decision Tree (b) Random Forest (c) SelectKBest (d) XGBoost and evaluate using decision

CHAPTER 5 : CONCLUSION AND FUTURE RECOMMENDATION

This study employs the use of machine learning (ML) to predict the sludge volume index at 30 minutes (SVI_{30}) of an OPG-based wastewater treatment (WWT) system. Advance feature selection methods and two two-stage feature selection were utilized To identify the most influential features. Subsequently, these selected features were used to train four different regression models: decision tree, random forest, gradient boosting, and XGBoost. Each model's performance was evaluated using various metrics to determine their effectiveness in predicting SVI_{30}. Among these regression models, the decision tree and random forest outperformed the gradient-boosting and XGBoost models, demonstrating improved accuracy and performance

Recursive Feature Elimination (decision tree) yielded an optimal combination of features which were evaluated by a decision tree to give the best result in predicting SVI_{30}. The decision tree's ability to handle non-linear relationships and interactions between features made it particularly suitable for this application.

The study underscores the potential of data-driven modelling techniques in optimizing the performance of OPG-based wastewater treatment systems. By accurately predicting SVI30, these models can contribute to more efficient and sustainable wastewater management practices. The findings pave the way for future advancements in predictive modelling, suggesting that incorporating advanced ML techniques and comprehensive feature selection methods can significantly enhance the management and operational strategies of wastewater treatment facilities.

Future research on Sludge Volume Index (SVI) prediction of OPG-based wastewater treatment optimization could benefit in upscaling of wastewater treatment reactor. Incorporating hybrid models by combining the machine learning models with Kinetics-based models could provide more reliable and accurate predictions. Additionally, the development of soft sensors to estimate hard-to-measure variables in the wastewater treatment process, which are time-consuming and costly to measure by using easy-to-measure variables. These soft sensors are data-driven models that could significantly enhance real-time monitoring and control of the reactor by integrating easily measurable variables to infer critical information, such as dynamics of microbial community or sludge settling properties. This advancement would enable more precise and adaptive management of the OPG-based treatment process, leading to improved efficiency and stability of the wastewater treatment reactor.

CHAPTER 6 REFERENCES

[1] S. Longo *et al.*, "Monitoring and diagnosis of energy consumption in wastewater treatment plants. A state of the art and proposals for improvement," 2016. doi: 10.1016/j.apenergy.2016.07.043.

[2] A. Hussain, R. Kumari, S. G. Sachan, and A. Sachan, "Biological wastewater treatment technology: Advancement and drawbacks," in *Microbial Ecology of Wastewater Treatment Plants*, 2021. doi: 10.1016/B978-0-12-822503-5.00002-3.

[3] A. A. Ansari, A. A. Ansari, A. S. Abouhend, J. G. Gikonyo, and C. Park, "Photogranulation in a Hydrostatic Environment Occurs with Limitation of Iron," *Environ Sci Technol*, vol. 55, no. 15, 2021, doi: 10.1021/acs.est.0c07374.

[4] J. G. Gikonyo, A. A. Ansari, A. S. Abouhend, J. E. Tobiason, and C. Park, "Hydrodynamic granulation of oxygenic photogranules," *Environ Sci (Camb)*, vol. 7, no. 2, 2021, doi: 10.1039/d0ew00957a.

[5] K. Milferstedt *et al.*, "The importance of filamentous cyanobacteria in the development of oxygenic photogranules," *Sci Rep*, vol. 7, no. 1, 2017, doi: 10.1038/s41598-017-16614-9.

[6] O. Tiron, C. Bumbac, E. Manea, M. Stefanescu, and M. N. Lazar, "Overcoming Microalgae Harvesting Barrier by Activated Algae Granules," *Sci Rep*, vol. 7, no. 1, 2017, doi: 10.1038/s41598-017-05027-3.

[7] A. S. Abouhend *et al.*, "The Oxygenic Photogranule Process for Aeration-Free Wastewater Treatment," *Environ Sci Technol*, vol. 52, no. 6, 2018, doi: 10.1021/acs.est.8b00403.

[8] A. Chandran,] K Mophin Kani, and] Anu, "Treatment of Municipal Wastewater using Oxygenic Photo granules (OPGs)," 2018.

[9] A. M. McNair, "Pilot Reactor Operation of the Oxygenic Photogranule (OPG) Wastewater Treatment Process," *Environmental & Water Resources Engineering Masters Projects*, May 2017, doi: https://doi.org/10.7275/2rc1-qw06.

[10] D. Brockmann, Y. Gérand, C. Park, K. Milferstedt, A. Hélias, and J. Hamelin, "Wastewater treatment using oxygenic photogranule-based process has lower environmental impact than conventional activated sludge process," *Bioresour Technol*, vol. 319, 2021, doi: 10.1016/j.biortech.2020.124204.

[11] K. Milferstedt *et al.*, "Biogranules applied in environmental engineering," *Int J Hydrogen Energy*, vol. 42, no. 45, 2017, doi: 10.1016/j.ijhydene.2017.07.176.

[12] G. Smetana and A. Grosser, "The Oxygenic Photogranules—Current Progress on the Technology and Perspectives in Wastewater Treatment: A Review," 2023. doi: 10.3390/en16010523.

[13] A. S. Abouhend, J. G. Gikonyo, M. Patton, C. S. Butler, J. Tobiason, and C. Park, “Role of Hydrodynamic Shear in the Oxygenic Photogranule (OPG) Wastewater Treatment Process,” *ACS ES and T Water*, vol. 3, no. 3, pp. 659–668, 2023, doi: 10.1021/acsestwater.2c00317.

[14] L. M. Trebuch, B. O. Oyserman, M. Janssen, R. H. Wijffels, L. E. M. Vet, and T. V. Fernandes, “Impact of hydraulic retention time on community assembly and function of photogranules for wastewater treatment,” *Water Res*, vol. 173, p. 115506, 2020, doi: 10.1016/j.watres.2020.115506.

[15] W. C. Kuo-Dahab *et al.*, “Investigation of the Fate and Dynamics of Extracellular Polymeric Substances (EPS) during Sludge-Based Photogranulation under Hydrostatic Conditions,” *Environ Sci Technol*, vol. 52, no. 18, pp. 10462–10471, 2018.

[16] A. S. Abouhend *et al.*, “Growth Progression of Oxygenic Photogranules and Its Impact on Bioactivity for Aeration-Free Wastewater Treatment,” *Environ Sci Technol*, vol. 54, no. 1, pp. 486–496, 2020, doi: https://doi.org/10.1021/acs.est.9b04745.

[17] A. A. Ansari, A. S. Abouhend, and C. Park, “Effects of seeding density on photogranulation and the start-up of the oxygenic photogranule process for aeration-free wastewater treatment,” *Algal Res*, vol. 40, 2019, doi: 10.1016/j.algal.2019.101495.

[18] A. A. Ansari, A. A. Ansari, J. G. Gikonyo, A. S. Abouhend, and C. Park, “The Coupled Effect of Light and Iron on the Photogranulation Phenomenon,” *Environ Sci Technol*, vol. 57, no. 24, pp. 9086–9095, 2023, doi: 10.1021/acs.est.3c00432.

[19] P. Wongburi and J. K. Park, “Prediction of Sludge Volume Index in a Wastewater Treatment Plant Using Recurrent Neural Network,” *Sustainability (Switzerland)*, vol. 14, no. 10, 2022, doi: 10.3390/su14106276.

[20] Y. Yao, T. Chen, and F. Gao, “Multivariate statistical monitoring of two-dimensional dynamic batch processes utilizing non-Gaussian information,” *J Process Control*, vol. 20, no. 10, 2010, doi: 10.1016/j.jprocont.2010.07.002.

[21] M. Da Ma, D. S. H. Wong, S. S. Jang, and S. T. Tseng, “Fault detection based on statistical multivariate analysis and microarray visualization,” *IEEE Trans Industr Inform*, vol. 6, no. 1, 2010, doi: 10.1109/TII.2009.2030793.

[22] M. Bahramian, R. K. Dereli, W. Zhao, M. Giberti, and E. Casey, “Data to intelligence: The role of data-driven models in wastewater treatment,” May 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.eswa.2022.119453.

[23] N. D. Viet and A. Jang, “Comparative mathematical and data-driven models for simulating the performance of forward osmosis membrane under different draw solutions,” *Desalination*, vol. 549, 2023, doi: 10.1016/j.desal.2022.116346.

[24] S. Hiemer and S. Zapperi, “From mechanism-based to data-driven approaches in materials science,” *Materials Theory*, vol. 5, no. 1, 2021, doi: 10.1186/s41313-021-00027-3.

[25] H. Su, T. Zhu, J. Lv, H. Wang, J. Zhao, and J. Xu, “Leveraging machine learning for prediction of antibiotic resistance genes post thermal hydrolysis-anaerobic digestion in dairy waste,” *Bioresour Technol*, vol. 399, p. 130536, May 2024, doi: 10.1016/J.BIORTECH.2024.130536.

[26] H. Bao *et al.*, “Automated machine learning-based models for predicting and evaluating antibiotic removal in constructed wetlands,” *Bioresour Technol*, vol. 385, Oct. 2023, doi: 10.1016/J.BIORTECH.2023.129436.

[27] M. J. K. Bashir *et al.*, “Wastewater treatment processes optimization using response surface methodology (RSM) compared with conventional methods: Review and comparative study,” *Middle-East Journal of Scientific Research*, vol. 23, no. 2, 2015.

[28] N. K. Singh *et al.*, “Artificial intelligence and machine learning-based monitoring and design of biological wastewater treatment systems,” 2023. doi: 10.1016/j.biortech.2022.128486.

[29] A. F. Del Castillo, M. V. Garibay, C. Senés-Guerrero, C. Yebra-Montes, J. de Anda, and M. S. Gradilla-Hernández, “Mathematical Modeling of a Domestic Wastewater Treatment System Combining a Septic Tank, an Up Flow Anaerobic Filter, and a Constructed Wetland,” *Water 2020, Vol. 12, Page 3019*, vol. 12, no. 11, p. 3019, Oct. 2020, doi: 10.3390/W12113019.

[30] F. Harrou, A. Dairi, Y. Sun, and M. Senouci, “Statistical monitoring of a wastewater treatment plant: A case study,” *J Environ Manage*, vol. 223, 2018, doi: 10.1016/j.jenvman.2018.06.087.

[31] Z. Ge, Z. Song, and F. Gao, “Review of recent research on data-based process monitoring,” 2013. doi: 10.1021/ie302069q.

[32] E. A. del Rio-Chanona, J. L. Wagner, H. Ali, F. Fiorelli, D. Zhang, and K. Hellgardt, “Deep learning-based surrogate modeling and optimization for microalgal biofuel production and photobioreactor design,” *AIChE Journal*, vol. 65, no. 3, 2019, doi: 10.1002/aic.16473.

[33] Y. Zeng, J. Liu, K. Sun, and L. wen Hu, “Machine learning based system performance prediction model for reactor control,” *Ann Nucl Energy*, vol. 113, 2018, doi: 10.1016/j.anucene.2017.11.014.

[34] W. Xu *et al.*, “Performance prediction of ZVI-based anaerobic digestion reactor using machine learning algorithms,” *Waste Management*, vol. 121, 2021, doi: 10.1016/j.wasman.2020.12.003.

[35] C. Dong and J. Chen, “Optimization of process parameters for anaerobic fermentation of corn stalk based on least squares support vector machine,” *Bioresour Technol*, vol. 271, 2019, doi: 10.1016/j.biortech.2018.09.085.

[36] M. Y. Arshad, M. A. Saeed, M. W. Tahir, H. Pawlak-Kruczek, A. S. Ahmad, and L. Niedzwiecki, “Advancing Sustainable Decomposition of Biomass Tar Model Compound: Machine Learning, Kinetic Modeling, and Experimental Investigation in a Non-Thermal Plasma Dielectric Barrier Discharge Reactor,” *Energies (Basel)*, vol. 16, no. 15, 2023, doi: 10.3390/en16155835.

[37] H. Y. Shyu, C. J. Castro, R. A. Bair, Q. Lu, and D. H. Yeh, "Development of a Soft Sensor Using Machine Learning Algorithms for Predicting the Water Quality of an Onsite Wastewater Treatment System," *ACS Environmental Au*, 2023, doi: 10.1021/acsenvironau.2c00072.

[38] M. S. Zaghloul, O. T. Iorhemen, R. A. Hamza, J. H. Tay, and G. Achari, "Development of an ensemble of machine learning algorithms to model aerobic granular sludge reactors," *Water Res*, vol. 189, 2021, doi: 10.1016/j.watres.2020.116657.

[39] N. Mahmod and N. A. Wahab, "Dynamic modelling of aerobic granular sludge artificial neural networks," *International Journal of Electrical and Computer Engineering*, vol. 7, no. 3, 2017, doi: 10.11591/ijece.v7i3.pp1568-1573.

[40] N. S. A. Yasmin, N. A. Wahab, A. N. Anuar, and M. Bob, "Performance comparison of SVM and ANN for aerobic granular sludge," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, 2019, doi: 10.11591/eei.v8i4.1605.

[41] L. Wang, F. Long, W. Liao, and H. Liu, "Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning algorithms," *Bioresour Technol*, vol. 298, 2020, doi: 10.1016/j.biortech.2019.122495.

[42] S. K. Heo, K. J. Nam, S. Tariq, J. Y. Lim, J. Park, and C. K. Yoo, "A hybrid machine learning–based multi-objective supervisory control strategy of a full-scale wastewater treatment for cost-effective and sustainable operation under varying influent conditions," *J Clean Prod*, vol. 291, 2021, doi: 10.1016/j.jclepro.2021.125853.

[43] B. Venkatesh and J. Anuradha, "A review of Feature Selection and its methods," *Cybernetics and Information Technologies*, vol. 19, no. 1, 2019, doi: 10.2478/CAIT-2019-0001.

[44] I. M. El-Hasnony, S. I. Barakat, M. Elhoseny, and R. R. Mostafa, "Improved Feature Selection Model for Big Data Analytics," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2986232.

[45] F. Bagherzadeh, M. J. Mehrani, M. Basirifard, and J. Roostaei, "Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance," *Journal of Water Process Engineering*, vol. 41, 2021, doi: 10.1016/j.jwpe.2021.102033.

[46] W. Xu *et al.*, "Performance prediction of ZVI-based anaerobic digestion reactor using machine learning algorithms," *Waste Management*, vol. 121, pp. 59–66, Feb. 2021, doi: 10.1016/j.wasman.2020.12.003.

[47] E. Ekinci, B. Özbay, S. İ. Omurca, F. E. Sayın, and İ. Özbay, "Application of machine learning algorithms and feature selection methods for better prediction of sludge production in a real advanced biological wastewater treatment plant," *J Environ Manage*, vol. 348, Dec. 2023, doi: 10.1016/J.JENVMAN.2023.119448.

[48] N. Hvala and J. Kocijan, "Input variable selection using machine learning and global sensitivity methods for the control of sludge bulking in a wastewater treatment plant," *Comput Chem Eng*, vol. 154, 2021, doi: 10.1016/j.compchemeng.2021.107493.

[49] Y. Alali, F. Harrou, and Y. Sun, "Unlocking the Potential of Wastewater Treatment: Machine Learning Based Energy Consumption Prediction," *Water (Switzerland)*, vol. 15, no. 13, 2023, doi: 10.3390/w15132349.

[50] J. G. Gikonyo, A. S. Abouhend, A. Keyser, Y. Li, and C. Park, "Scaling-up of oxygenic photogranular system in selective-CSTR," *Bioresour Technol Rep*, vol. 23, 2023, doi: 10.1016/j.biteb.2023.101523.

[51] A. Juna *et al.*, "Water Quality Prediction Using KNN Imputer and Multilayer Perceptron," *Water (Switzerland)*, vol. 14, no. 17, 2022, doi: 10.3390/w14172592.

[52] A. M. Anter, D. Gupta, and O. Castillo, "A novel parameter estimation in dynamic model via fuzzy swarm intelligence and chaos theory for faults in wastewater treatment plant," *Soft comput*, vol. 24, no. 1, 2020, doi: 10.1007/s00500-019-04225-7.

[53] P. Meesad and K. Hengpraprohm, "Combination of KNN-based feature selection and KNN-based missing-value imputation of microarray data," *3rd International Conference on Innovative Computing Information and Control, ICICIC'08*, 2008, doi: 10.1109/ICICIC.2008.635.

[54] A. M. Taha, A. Mustapha, and S. Der Chen, "Naive Bayes-guided bat algorithm for feature selection," *The Scientific World Journal*, vol. 2013, 2013, doi: 10.1155/2013/325973.

[55] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, 2021, doi: 10.7717/PEERJ-CS.623.

[56] J. Meng, H. Lin, and Y. Yu, "A two-stage feature selection method for text categorization," in *Computers and Mathematics with Applications*, 2011. doi: 10.1016/j.camwa.2011.07.045.

[57] B. van den Akker, H. Beard, U. Kaeding, S. Giglio, and M. D. Short, "Exploring the relationship between viscous bulking and ammonia-oxidiser abundance in activated sludge: A comparison of conventional and IFAS systems," *Water Res*, vol. 44, no. 9, pp. 2919–2929, May 2010, doi: 10.1016/J.WATRES.2010.02.016.

[58] P. Wongburi and J. K. Park, "Prediction of Sludge Volume Index in a Wastewater Treatment Plant Using Recurrent Neural Network," *Sustainability 2022, Vol. 14, Page 6276*, vol. 14, no. 10, p. 6276, May 2022, doi: 10.3390/SU14106276.

[59] M. S. Zaghloul and G. Achari, "Application of machine learning techniques to model a full-scale wastewater treatment plant with biological nutrient removal," *J Environ Chem Eng*, vol. 10, no. 3, 2022, doi: 10.1016/j.jece.2022.107430.

[60] M. Bubalo, I. Šumelj, K. Herceg, and N. V. Medvidović, “Assessment in treatment efficiency of a small-scale municipal wastewater treatment plant with activated sludge,” *Ekologia Bratislava*, vol. 41, no. 3, pp. 272–282, Sep. 2022, doi: 10.2478/EKO-2022-0028.

[61] J. Jeon, M. Choi, S. Park, and H. Bae, “Management strategy of granular sludge settleability in saline denitrification: Insights from machine learning,” *Chemical Engineering Journal*, vol. 493, Aug. 2024, doi: 10.1016/J.CEJ.2024.152747.

[62] H. Harun and A. Nor-Anuar, “Development and utilization of aerobic granules for soy sauce wastewater treatment: Optimization by response surface methodology,” *Jurnal Teknologi (Sciences and Engineering)*, vol. 69, no. 5, pp. 31–37, 2014, doi: 10.11113/JT.V69.3201.

LIST OF PUBLICATIONS

Journal of Water Process Engineering 66 (2024) 106064

Contents lists available at ScienceDirect

Journal of Water Process Engineering

journal homepage: www.elsevier.com/locate/jwpe

Performance prediction of sludge volume index of oxygenic photogranule based wastewater treatment system using machine learning algorithms

Sidra Yasin [a], Abeera Ayaz Ansari [a,*], Abdul Kashif Janjua [b,c,**], Joseph Gitau Gikonyo [d], Ghayoor Abbas [a]

[a] *Department of Energy Systems Engineering, U.S.-Pakistan Center for Advanced Studies in Energy (USPCASE), National University of Sciences and Technology (NUST), Sector H-12, Islamabad 44000, Pakistan*
[b] *School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Sector H-12, Islamabad 44000, Pakistan*
[c] *National Skills University, Faiz Ahmed Faiz Rd, Sector H-8/1, Islamabad 44000, Pakistan*
[d] *Department of Civil and Environmental Engineering, University of Massachusetts, Amherst, MA 01003, USA*

A R T I C L E I N F O

Editor: Guangming Jiang

Keywords:
Oxygenic Photogranules
Wastewater treatment
Machine learning
Feature selection
Data-driven modelling

A B S T R A C T

Oxygenic Photogranulation is a novel biotechnology that treats wastewater without external aeration and produces biomass in dense photogranules with high settling velocities. Oxygenic photogranules (OPG) based wastewater treatment (WWT) faces challenges during scaleup due to its dynamic and complex system variables, making troubleshooting costly. The Machine Learning (ML) approach can address this issue by creating a WWT process simulation. Moreover, traditional mechanistic models do not capture the interaction between input and output features due to their high dimensionality and the non-linear relationship, making them computationally expensive. In this study, the two-stage feature selection (FS) method is studied to enhance the prediction performance of SVI_{30}, a critical operational parameter, to ensure optimal settleability and minimal loss of active photogranules. The optimal feature subsets generated by the two-stage selection method were evaluated using four regression models: decision tree, random forest, gradient boosting, and extreme gradient boosting. Results indicate that, among all regression models, the decision tree performs well having a prediction efficiency of 85 % with the subset of features obtained after Recursive Feature Elimination (RFE) of decision tree features in the second stage. This indicates the effectiveness of the two-stage FS method in identifying the most relevant features for predicting SVI_{30}. The structured approach of FS and model evaluation highlights the potential of ML in addressing complex operational challenges in OPG WWT operations.

1. Introduction

OPG-based WWT is a novel biotechnology, which has emerged as an attractive alternative to the energy-intensive activated sludge process [1,2]. OPGs are bio-aggregates comprised of phototrophic microorganisms that surround heterotrophic bacteria in a dense, spherical structure [3]. They are produced from transforming activated sludge under illumination sources during hydrostatic [4] or hydrodynamic cultivation environments [5]. The presence of a phototrophic community promotes aeration-free WWT while assimilating CO_2, which encourages a reduction in WWT-associated GHG emissions [6,7]. The produced OPGs also have higher density and settleability, hence reducing the risk of biomass washout and enhancing the effluent quality [8,9].

Despite OPG-based WWT being promising at the laboratory scale, the attempts to scale up this technology have encountered numerous setbacks, including loss of granular biomass, decline in treatment performance, and subsequent loss of reactor functionality [10]. Photogranules, the core component of the OPG-based WWT process, need to maintain their structural integrity and settling properties for efficient WWT and biomass handling. Various studies have been investigated to determine the factors responsible for the promotion of photo granulation including mixing speed [11], hydraulic retention time (HRT) [12], extracellular polymeric substances (EPS) production [13,14], seeding density [7], light intensity and Iron [15]. The sludge volume index (SVI) is a critical

* Corresponding author.
** Correspondence to: A. K. Janjua, School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Sector H-12, Islamabad 44000, Pakistan.
E-mail addresses: abeera@uspcase.nust.edu.pk (A.A. Ansari), kashif@nsu.edu.pk (A.K. Janjua).

https://doi.org/10.1016/j.jwpe.2024.106064
Received 31 May 2024; Received in revised form 31 July 2024; Accepted 24 August 2024