Improvement of Operational Performance of Submerged Membrane Photocatalytic Ultrafiltration for Industrial Oily Wastewater Treatment using AL/ML Technique and Statistical Optimization Methodology

by

Alireza KHADEMI

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR A MASTER'S DEGREE
WITH THESIS IN ENGINEERING WITH A PERSONALIZED
CONCENTRATION
M.A.SC.

MONTREAL, JANUARY 31, 2023

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS (THESIS M.A.SC.)

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Prof. Claudiane Ouellet-Plamondon, Thesis Supervisor
Département de génie de la construction at École de technologie supérieure

Prof. Benoit Barbeau, Thesis Co-supervisor
Department of Civil, Geological, and Mining Engineering at Polytechnique Montréal

Prof. Mohammad Jahazi, President of the Board of Examiners
Département de génie mécanique at École de technologie supérieure

Prof. Mathieu Lapointe, Member of the Jury
Département de génie de la construction at École de technologie supérieure

THIS THESIS WAS PRENSENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND PUBLIC

ON JANUARY 20, 2023

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGMENT

First and foremost, I would like to thank my research supervisors, Prof. Claudiane Ouellet-Plamondon and Prof. Benoit Barbeau, for their unwavering support throughout my studies at École de Technologie Supérieure (ETS). They were a huge inspiration for me to work on this project. Their willingness to motivate me was extremely beneficial to the project. They taught me a lot, and I consider myself fortunate to have them as supervisors. In addition, I would like to thank the Advanced Membrane Technology Research Center (AMTEC), Prof. Dr. Noordin bin Mohd. Yusof, Prof. Datuk Dr. Ahmad Fauzi Bin Ismail, and a special thanks to Prof. Dr. Takeshi Matsuura for their support and for providing the necessary information to carry out this project.

AMELIORATION DES PERFORMANCES OPERATIONNELLES DE L'ULTRAFILTRATION PHOTOCATALYTIQUE A MEMBRANE IMMERGEE POUR LE TRAITEMENT DES EAUX USEES HUILEUSES INDUSTRIELLES A L'AIDE DE LA TECHNIQUE AL/ML ET DE LA METHODOLOGIE D'OPTIMISATION STATISTIQUE

Alireza Khademi

RESUME

La pénurie d'eau est le prochain grand problème auquel le monde est sur le point d'être confronté, car 70 % de la surface de la Terre est recouverte d'eau, dont seulement 2.5 % d'eau douce. Les raffineries de pétrole sont une source majeure de pollution de l'eau. Elles produisent une quantité importante d'eaux usées pétrolières. L'élimination de ces eaux contaminées représente un défi important pour l'ensemble de l'industrie pétrolière. Le traitement des eaux usées municipales et industrielles est un autre enjeu crucial. De 2018 à 2025, le marché mondial de la filtration membranaire devrait atteindre 19.6 milliards USD, avec un taux de croissance annuel composé de 6.4 %. L'augmentation de la population, la sensibilisation croissante à la réutilisation des eaux usées, l'industrialisation rapide, les produits haut de gamme et l'efficacité offerts par les technologies de filtration membranaire, le passage du traitement chimique de l'eau au traitement physique et les réglementations strictes concernant les eaux de traitement et les rejets d'eau sont les principaux moteurs du marché de la filtration membranaire. Cette recherche se concentre sur la méthode de traitement des eaux usées de l'ultrafiltration photocatalytique à membrane immergée. Dans cette méthode, la membrane est utilisée pour nettoyer les eaux usées huileuses.

Les projets d'évaluation de la performance des membranes impliquent des méthodes d'essai et d'erreur dans un large éventail de conditions d'exploitation du procédé. La majorité des projets de recherche précédents sur le système d'ultrafiltration à membrane immergée (SMUF) ont utilisé une approche expérimentale avec une seule variable de procédé à la fois. Cette méthode prend beaucoup de temps et est coûteuse. L'analyse en composantes principales (ACP), les plans d'expériences (DOE) et l'intelligence artificielle/apprentissage machine (IA/ML) semblent être capables de contourner cette limitation. Grâce à une stratégie expérimentale systématique basée sur l'ACP, le DOE et l'AI/ML, cette étude a permis d'améliorer les performances opérationnelles de la technologie des opérations d'ultrafiltration photocatalytique à membrane immergée dans le traitement des eaux usées huileuses industrielles. Cette stratégie se concentre sur le réglage et l'optimisation simultanée des variables de fonctionnement de la membrane dans diverses conditions contrôlables. Parallèlement, l'absence d'un modèle de simulation numérique rapide capable de reconnaître et de prédire les effets des diverses conditions de fonctionnement du processus sur les performances finales du système de traitement des eaux usées a incité à utiliser des méthodologies statistiques et d'intelligence artificielle (IA). L'objectif ultime de cette étude est d'optimiser et d'améliorer les performances opérationnelles du système d'ultrafiltration photocatalytique à membrane immergée (SMPUF) dans le traitement des eaux usées huileuses industrielles. Pour atteindre cet objectif, des méthodes statistiques, l'analyse en composantes principales (ACP), des plans d'expériences (DOE), et des techniques d'IA/ML sont utilisées pour étudier et améliorer les performances du système.

Mots-Clés: Analyse en Composantes Principales (ACP), Plan d'Expériences (DOE), Intelligence Artificielle (IA), Apprentissage Machine (ML), Traitement des Eaux usées Huileuses Industrielles, Membrane à Fibres Creuses Asymétriques en PVDF, Système d'Ultrafiltration à Membrane Immergée (SMUF), Réacteur Photocatalytique à Membrane Immergée (SMPR)

IMPROVEMENT OF OPERATIONAL PERFORMANCE OF SUBMERGED MEMBRANE PHOTOCATALYTIC ULTRAFILTRATION FOR INDUSTRIAL OILY WASTEWATER TREATMENT USING AL/ML TECHNIQUE AND STATISTICAL OPTIMIZATION METHODOLOGY

Alireza Khademi

ABSTRACT

Water scarcity is the next big problem the world is about to face as 70% of the Earth's surface is covered in water of which only 2.5% is fresh water. Oil refineries are a major source of water pollution. They produce a significant amount of oil wastewater. The disposal of this contaminated water represents a significant challenge for the entire petroleum industry. The treatment of municipal and industrial wastewater is another critical issue. From 2018 to 2025, the global membrane filtration market is expected to reach USD 19.6 billion, growing at a CAGR of 6.4%. The increasing population, growing awareness of wastewater reuse, rapid industrialization, high-end products and efficiency offered by membrane filtration technologies, shift from chemical water treatment to physical treatment, and stringent regulations regarding treatment water and water discharge are the key drivers of the membrane filtration market. This research focuses on the wastewater treatment method of submerged membrane photocatalytic ultrafiltration. The membrane is used to clean oily wastewater in this method.

Membrane performance evaluation projects entail trial-and-error methods under a wide range of process operating conditions. The majority of previous research projects on the submerged membrane ultrafiltration (SMUF) system used an experimental approach with one process variable at a time. This method is time-consuming and expensive. Principal component analysis (PCA), design of experiments (DOE), and Artificial Intelligence/Machine Learning (AI/ML) appear to be capable of circumventing the limitation. Through a systematic experimental strategy based on PCA, DOE, and AI/ML, this study improved the operational performance of submerged membrane photocatalytic ultrafiltration operations technology in industrial oily wastewater treatment. This strategy focuses on tuning and simultaneous optimization of membrane operating variables under various controllable conditions. Meanwhile, the lack of a quick numerical simulation model capable of recognizing and predicting the effects of various process operating conditions on the final performance of the wastewater treatment system has prompted the utilization of statistical and artificial intelligence (AI) methodologies. This study's ultimate goal is to optimize and improve the operational performances of the Submerged Membrane Photocatalytic Ultrafiltration (SMPUF) system in industrial oily wastewater treatment. To accomplish this goal, statistical methods, principal components analysis (PCA), design of experiments (DOE), and AI/ML techniques are employed to investigate and improve the system's performance.

Keywords: Principal Components Analysis (PCA), Design of Experiments (DOE), Artificial Intelligence (AI), Machine Learning (ML), Industrial Oily Wastewater Treatment, PVDF

Asymmetric Hollow Fiber Membrane, Submerged Membrane Ultrafiltration (SMUF) System,
Submerged Membrane Photocatalytic Reactor (SMPR)

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABREVIATIONS

ABFR	Air Bubbling Flow Rate
ABS	Absorbance
AI	Artificial Intelligence
AMTEC	Advanced Membrane Technology Research Center
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
BOD	Biological Oxygen Demand
BPNN	Backpropagation Neural Network
COD	Chemical Oxygen Demand
DFSS	Design for Six Sigma
DL	Deep Learning
DOE	Design of Experiments
GA	Genetic Algorithm
HRT	Hydraulic Retention Time
LM	Levenberg-Marquardt
MAE	Mean Absolute Error
MEUF	Micellar-Enhanced Ultrafiltration
ML	Machine learning
MLP	Multi-Layer Perceptron
MLSS	Mixed Liquor Suspended Solids
MPD	Module Packing Density
MR	Multiple Regression

MSE	Mean Squared Error
NH_3-N	Sulfide Ammonia Nitrogen
OC	Oil Concentration
PCA	Principal Components Analysis
PSO	Particle Swarm Optimization
PVDF	Hydrophobic Polyvinylidene Fluoride
SDS	Sodium dodecylbenzenesulfonate
SI	Swarm Intelligence
SMPR	Submerged Membrane Photocatalytic Reactor
SMPUF	Submerged Membrane Photocatalytic Ultrafiltration
SMUF	Submerged Membrane Ultrafiltration
SNN	Spiking Neural Network
TDS	Total Dissolved Solid
TL	Transfer Learning
TOC	Total Organic Carbon
TSS	Total Suspended Solid
UVIT	UV Irradiation Time

INTRODUCTION

Water scarcity is the world’s next major issue, with 70% of the Earth’s surface covered in water, only 2.5% of which is fresh water. Oil refineries are a major source of water pollution. The disposal of this contaminated water presents a significant challenge. The treatment of municipal and industrial wastewater is another critical issue. From 2018 to 2025, the global membrane filtration market is expected to reach USD 19.6 billion, growing at a CAGR of 6.4%. The increasing population, growing awareness of wastewater reuse, rapid industrialization, high-end products, and efficiency offered by membrane filtration technologies, shift from chemical water treatment to physical treatment, and stringent regulations regarding treatment water and water discharge are the key drivers of the membrane filtration market.

Some conventional oily wastewater treatment methods are incapable of treating nanoscale oil particles. These technologies are constrained by high operating costs, low efficiency, a large footprint, and high energy consumption. The use of ultrafiltration (UF) membrane technology for oily wastewater treatment has several advantages, including high efficiency in oil droplet removal, low energy consumption, minimal chemical use (only for cleaning), and no production of harmful by-products.

In general, membrane performance evaluation models use trial-and-error methods to determine optimum material formulation, critical operating variable selection, and proper fiber spinning conditions. Despite extensive research in the past, evaluating membrane performance for specific applications remains difficult.

As previously reported, most previous research projects on SMUF wastewater treatment systems used a one-process variable at a time experimental approach, i.e., the influence of variables was investigated separately, which is not only time-consuming but also costly. Furthermore, the use of conventional experimentation methods ignored the effect of factor interaction, resulting in low efficiency in process optimization. The use of principal components analysis (PCA), design of experiments (DOE), and AI/ML appears to have the potential to overcome the limitations of the one-variable-at-a-time approach. It is widely acknowledged that PCA, DOE, and AI/ML are effective tools for analyzing, modeling, and optimizing the operating conditions of various processes.

AI/ML techniques have successfully been applied to a wide range of problems. However, determining an appropriate set of structural and learning parameter values for an AI/ML remains a challenging task. Considered structures for AI/ML models are highly integrated with the structure of the dataset. Therefore, the structure of the dataset and the correlation between variables makes it more difficult to find a model that performs well in all output variables. Meanwhile, experts are unable to recognize and predict the effects of various process operating conditions on membrane performances due to the lack of a quick numerical simulation model.

The main goal of this work is to utilize AI/ML techniques and statistical optimization methodologies to optimize and improve the operational performance of membrane operations technology for industrial/refinery-produced oily wastewater treatment.

In this study, the focus is on PVDF asymmetric hollow fiber membrane, which is applied in industrial oily wastewater treatment using the SMPUF system. There are various categories of operating variables and membrane performance factors. However, in this work, the focus is on the available and deliverable input and output factors offered by the Advanced Membrane Technology Research Center (AMTEC).

Using PCA, DOE, and AI/ML methodologies, this study presents a systematic experimental strategy for analyzing, modeling, and optimizing the SMPUF oily wastewater treatment system. The models optimized and improved in this study enable early recognition of membrane performances based on operating process conditions and identify the optimum setting for the process operating conditions to obtain optimum performance values.

Organization of the Project

This project is divided into four phases. The first phase provides an overview of membrane separation technology, the submerged membrane ultrafiltration (SMUF) wastewater treatment system, and the research context. It provides a general overview of submerged ultrafiltration for the treatment of oily wastewater generated by refineries, as well as brief information on principal components analysis (PCA), design of experiments (DOE), and artificial intelligence (AI) methodology. Furthermore, previous research projects related to the current study and their directions are critically discussed. The problem statements, which identify the research direction, are then presented. The objectives and scopes of the study are elaborated in detail based on the problem statement. The experimental details are presented during the measurement phase. It describes where and how measurements were taken, what measurement tools were used to record input and output data, and what units these factors are measured in. Finally, the section explains the logic that goes into calculating the output results. The way the experiments were carried out dictated the merging of the analysis and improvement phases in the study. The project bounced back and forth between analysis and implementation-related activities due to multiple experiment iterations, factor level changes, and output results recording. This study's largest section contains collections of measured and calculated data. The models are developed here, and their veracity is empirically validated. We examine whether the models' predicted values and factor levels correspond to reality. The confirmation test aggregates, and sorts all obtained input and output results. The conclusion about how good the proposed models are reached here. The section includes model recommendations as well as a list of their benefits and drawbacks. Finally, a general conclusion and goals for future research round out the study.

To improve content consistency, the project structure was designed and ordered according to the **DMAIC** problem-solving structure (i.e., Define, Measure, Analyze, Improve, and Control). The structure and order of the project steps are shown below:

Chapter 1 (Define Phase)

- Introduction and background of the study
- Statement of the problem
- Purpose and objectives of the study
- Scope of the study
- Significance of the study

Chapter 2 (Literature Review): Previous research on AL/ML and statistical optimization methodology were reviewed.

Chapter 3 (Measurement Phase)

- Experimental details: Membrane and SMPR experimental details
- Membrane and module preparation
- Synthetic wastewater preparation
- Studied factors
- Description of equipment used

Chapter 4 (Analysis and Improvement Phase): The chapter also includes the control phase after DOE and AI/ML methodology to keep the context consistent.

- Operational Framework (outlines the steps involved in the analysis, optimization, and improvement)
- **1st Methodology: PCA**
  - PCA experimental design and analysis
  - PCA conclusion
- **2nd Methodology: DOE**
  - DOE experimental design and analysis
  - DOE verification or confirmation
  - DOE conclusion
- **3rd Methodology: AI/ML**
  - Data visualization and preprocessing
  - How to evaluate the performance of the multi-output regression models

- Description and comparison between 5 multi-output regression algorithms used in the study
- **Develop** and **train** the MLP model without PCA
    - **Test** and **comparison** between the actual data and the predicted data between the MLP model without PCA and other multiple-output models compared in this study
- **Develop** and **train** the MLP model with PCA
    - **Test** and **comparison** between the actual data and the predicted data between the MLP model with PCA and other multiple-output models compared in this study
- The best-performed multi-output AI/ML model **verification** or **confirmation** (i.e., MLP with and without PCA)
- Comparison between the MLP models and the DOE model
- An optimization approach to finding the best input settings to maximize responses (***Forward analysis***)
- What are the input settings that can produce the maximum output values? (***Backward analysis***)
- AI/ML conclusion

Finally, the conclusion, highlights, and recommendations round out the study.

CHAPITRE 1

INTRODUCTION

1.1 Introduction

The globe will soon have to deal with a water crisis in addition to fuel and energy shortages. Only 2.5 percent of the world's water is fresh, and two-thirds of that amount is frozen in ice caps and glaciers. A thousand times more water is needed for food production than is required for regular access to potable water. By 2025, 3.4 billion people will reside in nations classified as water-scarce zones, according to the UN (Peinemann, & Pereira Nunes, 2010).

1.2 Background of the Study

Traditional biological solids wastewater treatment methods are cost-effective, but the increasing demand for treated water reuse is forcing researchers to develop alternative treatment methods. Membrane development technologies are currently attracting a lot of attention in the field of wastewater treatment. Ultrafiltration membrane is an appealing option for treating wastewater generated by refineries. It has the advantage of consistently producing an almost pollutant-free effluent with fewer operational issues (Cao, Ma, Shi, & Ren, 2006; Yuliwati, & Ismail, 2011; Yuliwati, Ismail, Matsuura, Kassim, & Abdullah, 2011). One well-known application of the porous membrane in wastewater treatment is the production of emulsions of various sizes (Busch, Cruse, & Marquardt, 2007). It works with both oil-in-water (O/W) and water-in-oil (W/O) emulsions and is based on the ability of low pressure to force the dispersed phase to permeate through a membrane into a continuous phase. These membranes have the benefit of being simple and efficient separating devices for oil, grease, metals, biological oxygen demand (BOD), and chemical oxygen demand (COD). They can supply clear permeate that can be reused (Peinemann, & Pereira Nunes, 2010).

1.3 Wastewater Treatment

Because of its small footprint and ease of module manufacturing and operation, the approach of submerged membranes in refinery-produced wastewater treatment is advantageous. The assembly of hollow fiber membranes in the feed reservoir with liquid withdrawal through fibers enables direct immersion. The benefits of submerged hollow fiber membrane ultrafiltration paved the way for its use in oily wastewater treatment as well as various process industries such as petrochemical, metallurgical, and transportation (Peinemann, & Pereira Nunes, 2010; Yuliwati, 2012).

This study focuses on the performance of hydrophobic Polyvinylidene Fluoride (PVDF) asymmetric hollow fiber membranes used in the treatment of refinery-produced wastewater using submerged membrane ultrafiltration (SMUF). A dry-jet wet-spinning process was used to fabricate the hollow fiber in this study.

The global membrane filtration market is expected to be worth USD 13.5 billion in 2019 and USD 19.6 billion by 2025, growing at a 6.4% CAGR from 2018 to 2025. The primary drivers of the membrane filtration market are the rising population, increased awareness of wastewater reuse, and rapid industrialization. Furthermore, the increasing demand for premium products and the efficiencies provided by membrane filtration technologies have significantly fueled the market for membrane filtration for a variety of end uses. The transition from chemical to physical water treatment, as well as strict water treatment and discharge regulations, are driving the membranes market (MarketsandMarkets, FB 6183).

1.4 Principal Components Analysis Methodology

Principal components analysis (PCA) is a multivariate analysis method that is commonly used to reduce the dimensionality of large datasets. It is the most widely used statistical technique for detecting patterns in high-dimensional data and expressing the data in a way that highlights similarities and differences. PCA makes it easier to identify the variables that cause correlations and anticorrelations between samples. PCA data are organized in matrices, with rows representing samples and columns representing variables. The original data matrix is transformed into a set of new variables called loadings using singular value decomposition. Thus, PCA converts several correlated or potentially correlated variables into several uncorrelated variables known as principal components. The first principal component has the greatest variance, and the second, orthogonal to the first, has the greatest inertia (Babanova, Artyushkova, Ulyanova, Singhal, & Atanassov, 2014).

PCA is a projection method that can help you visualize all of the information in a datasheet. PCA can help determine how one sample differs from another, which variables contribute the most to this difference, and whether those variables contribute in the same way (i.e., correlated) or independently. It also allows for the detection of sample patterns, such as any particular grouping. Finally, it quantifies the amount of useful information contained in the data, as opposed to noise or meaningless variation (Babanova et al., 2014).

1.5 Design of Experiments Methodology

The Design of Experiments (DOE) is a statistical method used to determine unknown effects and evaluate factor effects under controlled conditions. Controlling input factors enables rapid analysis and optimization of system performance (Uy, & Telford, 2009). This method has been used to develop hypotheses about a specific process. The DOE method, according to Hambli et al. (2003), is an efficient and cost-effective way to model and analyze process variations. The method is applied to process improvement in such a way that processes can be defined using a number of controllable variables (Hambli, Richir, Crubleau, & Taravel, 2003). Using DOE engineers can determine which process variables have the greatest impact on its performance (Antony, & Capon,1998).

The factorial experimental design is an efficient method for experiments involving two or more variables (Uy, & Telford, 2009; Montgomery, 2008). When several independent sources of variation are presented, the highly effective technique of analysis of variance (ANOVA) is used to analyze experimental output (Montgomery, & Runger, 2011).

A designed experiment, according to Montgomery (2008, & 2011), is a test or series of tests in which deliberate modifications are made to the input variables of a process to see and track matching changes in the output response. The classification, as depicted in Figure 1.1, can be seen as one or more quality features or responses that can be found in the output product. Some process input factors can be changed, whereas others cannot, although they might not be controllable during the test. These uncontrollable factors are called noise factors. The objectives of the experiment may include identifying the input variables that have a significant effect on the responses and figuring out how to set those significant input variables so that the responses are close to the nominal requirements, responses variabilities are low, and the effects of the uncontrollable factors are reduced.

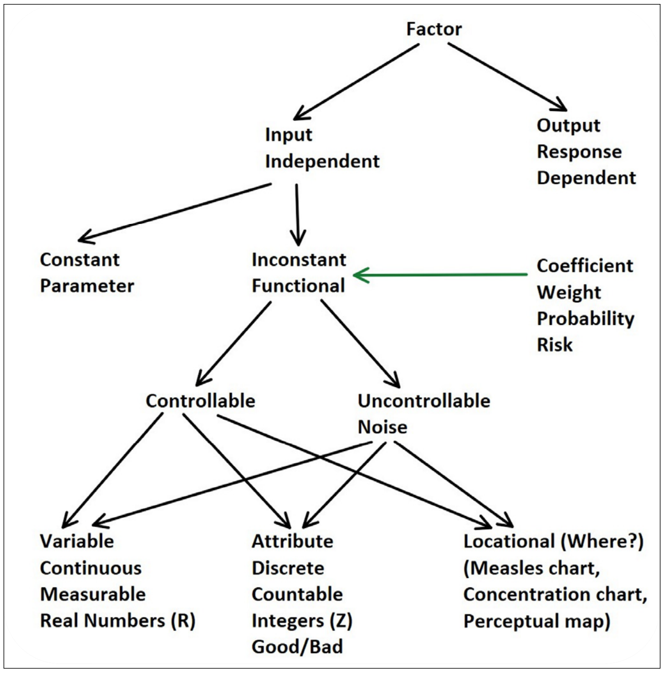


Figure 1.1 Factors Classification

Montgomery (2008, & 2011) stated that experimental design methods can be used in process development or troubleshooting to improve performance or achieve a robust or insensitive process to external sources of variability. Experimental design methods can also play an important role in engineering design activities, such as the development of new products and the improvement of existing ones. Designed experiments are frequently used in Design for Six Sigma (DFSS) activities.

1.6 Artificial Neural Networks Methodology

There are numerous types of Artificial Neural Networks (ANN) and their applications. The functions, accepted values, topology, learning algorithms, and so on may differ between them. There are also many hybrid models in which each neuron has more properties (Kantardzic, 2011). Because the function of ANN is to process information, it is primarily used in fields related to that. Pattern recognition, forecasting, and data compression are just a few of the engineering applications for ANN (Kantardzic, 2011).

Artificial neural networks (ANN) have been used successfully across a wide range of problem domains, and their widespread success may be attributed to their ability to approximate or learn complex relationships by utilizing massively interconnected and parallelly distributed nonlinear processors (Kantardzic, 2011).

Extensive research has been conducted to investigate the capability of an ANN as a tool for modeling membrane systems. As a result, the application of an ANN to membrane research is an appealing topic. For example, in the drinking water treatment system, ANN has been used to predict steady-state contaminant removal efficiency during nanofiltration. The ability of ANN to dynamically simulate membrane fouling during crossflow micro-filtration has also been investigated. Other applications include predicting permeate flux, hydraulic resistance, and rejection for various feed solutions (Chen, & Kim, 2006; Peterson, 2007; Wang, & Fu, 2007).

Among the various neural networks, the multilayer feed-forward neural network with a backpropagation training algorithm, commonly referred to as the backpropagation neural network (BPNN), is one of the most widely used ANN models in membrane research, with introductory overviews available in several studies (Chen, & Kim, 2006; Peterson, 2007; Wang, & Fu, 2007).

Rahmanian et al. (2011) used micellar-enhanced ultrafiltration (MEUF) to efficiently remove zinc ions from wastewater. They used design of experiments (DOE) and ANN models to test the prediction of permeate flux and metal ion rejection by MEUF. The fractional factorial design was used to determine all of the influential factors and their mutual effect on overall performance. The results demonstrated that, due to the difficulty in generalizing the MEUF process by any mathematical model, the neural network proves to be a very promising method for process simulation when compared to fractional factorial design. According to statistical analysis, a multilayer neural network was used in this work because it is effective in finding complex non-linear relationships for zinc removal using the MEUF process.

1.7 Statement of the Problem

Conventional oily wastewater treatment methods cannot handle nanoscale oil particles. These technologies are constrained by high operating costs, low efficiency, a large footprint, and high energy consumption. The use of ultrafiltration (UF) membrane technology for oily wastewater treatment has several advantages, including high efficiency in oil droplet removal, low energy consumption, minimal chemical use (only for cleaning), and no production of harmful by-products.

In general, membrane performance evaluation models use trial-and-error methods to determine optimum material formulation, critical operating variable selection, and proper fiber spinning conditions. Despite extensive research in the past, evaluating membrane performance for specific applications remains difficult.

As previously reported, most previous research projects on SMUF wastewater treatment systems used a one-process variable at a time experimental approach, i.e., the influence of variables was investigated separately, which is not only time-consuming but also costly. Furthermore, the use of traditional experimentation methods ignored the effect of factor interaction, resulting in low efficiency in process optimization. The use of principal components analysis (PCA), design of experiments (DOE), and AI/ML appears to have the potential to overcome the limitations of the one-variable-at-a-time approach. It is widely acknowledged that PCA, DOE, and AI/ML are effective tools for analyzing, modeling, and optimizing the operating conditions of various processes.

AI/ML techniques have been used successfully to solve a wide range of problems. However, determining a suitable set of structural and learning parameter values for an AI/ML, on the other hand, remains a difficult task. And the considered structures for AI/ML models are highly integrated with the structure of the dataset. In addition, the structure of the dataset and the correlation between variables makes it more difficult to find a model that performs well in all output variables. Furthermore, the lack of a quick numerical simulation model means that experts are unable to recognize and predict the effects of different process operating conditions on membrane performance.

1.8 Purpose and Objectives of the Study

The overall goal of this project is to utilize AI/ML techniques and statistical optimization methodologies to optimize and improve the operational performance of Submerged Membrane Photocatalytic Ultrafiltration (SMPUF) for industrial/refinery-produced oily wastewater treatment, which will be covered by the following objectives:

- To find patterns in a high dimensional dataset and express the dataset in such a way as to highlight their similarities and differences, and to reduce the dimensionality of the dataset while retaining as much information as possible about the inherent variability in the dataset using principal components analysis (PCA) methodology to facilitate visualization of all information contained in the dataset.

- To obtain the SMPUF system model based on the operating variables and performances to predict the membrane performances, and to investigate the influences of the simultaneous SMPUF oily wastewater treatment process operating conditions on the membrane performances using Design of Experiments (DOE) and AI/ML methodologies to optimize the performances of the SMPUF system.

- To identify the optimum setting of the SMPUF oily wastewater treatment process operating conditions to obtain optimum values of the membrane performances using DOE and AI/ML methodologies, and to do a comparison between the AL/ML models and the statistical model to determine which model is most efficient, effective, and sensitive to the operating variables.

1.9 Scope of the Study

Any study that wants to accomplish its objectives needs to have certain sufficiently narrow scopes. This study is also not an exception to the rule in this regard. The following are the research's scopes:

- There are various types of membranes with various kinds of applications. However, in this study, the focus is on PVDF asymmetric hollow fiber membrane, which is applied in the refinery-produced oily wastewater treatment using the Submerged Membrane Photocatalytic Ultrafiltration (SMPUF) system.
- There are various categories of operating variables, which affect the membrane performances when treating refinery-produced oily wastewater. However, in this work, the focus is on the available and deliverable factors offered by the Advanced Membrane Technology Research Center (AMTEC).
- Along the same line, this work just focuses on available membrane performance factors in the treatment of refinery-produced wastewater offered by the Advanced Membrane Technology Research Center (AMTEC).

The performances of the UF membrane are influenced by numerous variables. characteristics of the polymer, orientation, feed properties, operational conditions, and others. For ease of understanding, these variables are divided into three groups: feed properties, membrane composition, and membrane spinning conditions. The experiments only examined how feed properties affected membrane performance, all other variables remained constant (unchanged) throughout the experiments. The constant factors from Table 3.1 for membrane composition and Table 3.2 for process conditions are presented for reference.

Five of the most significant feed factors influencing membrane performance were the subject of research aimed at analyzing and optimizing. Hydrogen activity (pH), oil concentration (OC), ultraviolet irradiation time (UVIT), air bubbling flow rate (ABFR), and module packing density (MPD) are the variables investigated during the experiments. Table 3.3 provides an overview of the factors' composition and quality levels.

Figure 3.4 depicts membrane performance characteristics that are relevant to the measure. These are **flux** or permeate volume per unit area and unit time, oil rejection or absorbance (**ABS**), and total organic carbon (**TOC**) degradation. AMTEC specialists proposed the three responses under consideration. The permeation flux (***F***) represents SMUF productivity. And the rejection rates ($\boldsymbol{R_1}$ and $\boldsymbol{R_2}$) of PVDF hollow fibers are related to permeate water quality, which consists of absorbance (ABS) and total organic carbon (TOC) degradation. ***F*** represents SMUF productivity, and $\boldsymbol{R_1}$ and $\boldsymbol{R_2}$ represent, respectively, the filtration efficiency of hollow fiber UF membrane and photocatalytic efficiency of TiO_2 catalysts. TOC degradation samples were collected from feed, whereas oil rejection samples were collected from permeate. The effect of the feed factors i.e., ABFR, OC, UVIT, pH, and MPD on membrane performance characteristics i.e., Flux, ABS, and TOC was studied to obtain optimum operating conditions for the SMUF system.

1.10 Significance of the Study

This study strongly contributes to the body of knowledge from academic and professional perspectives.

- *In terms of the scientific aspects,* this study presents a systematic experimental strategy to analyze, model, and optimize the Submerged Membrane Photocatalytic Ultrafiltration (SMPUF) oily wastewater treatment system using principal components analysis (PCA), design of experiments (DOE), and AI/ML methodologies.
- *In terms of the applicability aspects,* the models optimized and improved through this study facilitate early recognition of membrane performances based on operating process conditions used and identify the optimum setting for the process operating conditions to obtain optimum values of the performances.

Since statistical approaches are not able to find the best input setting to optimize all responses together. And they are just able to find the best input setting to optimize each response individually. Therefore, a desirability analysis and AI/ML techniques were utilized to overcome this problem to find the most desirable local optimum point by considering all responses together.

In the AI/ML models, the output variables are typically dependent upon the input variables and each other. This means that the output variables are not independent and may require a model that predicts output variables together. That is why AI/ML models (multi-output regression models) were utilized.

The nonlinearity and complicated process dynamics make it challenging to model and predict membrane performances. As a result, several research fields perceive the application of improved AI/ML models as a practical approach for analyzing historical data. The effectiveness of treatment and the sustainability of water use for end users can both be enhanced by models and predictions of water quality variables.

AI/ML approaches perform better when simulating complicated and nonlinear engineering problems. Different methods of water treatment demonstrated that AI/ML is a cutting-edge tool in the field of water and wastewater treatment that can overcome significant drawbacks of traditional modeling techniques and lower the likelihood of human mistakes.

CHAPITRE 2

LITERATURE REVIEW

2.1 Introduction

Faced with water scarcity, the world is looking into all available options to reduce the overuse of limited freshwater resources. Wastewater is one of the most abundant sources of water. To meet human needs, industrial, agricultural, and domestic activities expand in tandem with population growth. These activities generate large amounts of wastewater, which can be recovered and used for a variety of purposes. Conventional wastewater treatment processes have had some success in treating effluent discharges over the years. However, improved wastewater treatment processes are required to make treated wastewater reusable for industrial, agricultural, and domestic purposes. Membrane technology has emerged as a reliable option for recovering water from various wastewater streams for reuse (Obotey Ezugbe, & Rathilal, 2020).

The global membrane filtration market is expected to be worth USD 13.5 billion in 2019 and USD 19.6 billion by 2025, growing at a 6.4% CAGR from 2018 to 2025. The primary drivers of the membrane filtration market are the rising population, increased awareness of wastewater reuse, and rapid industrialization. Furthermore, the increasing demand for premium products and the efficiencies provided by membrane filtration technologies have significantly fueled the market for membrane filtration for a variety of end uses. The transition from chemical to physical water treatment, as well as strict water treatment and discharge regulations, are driving the membranes market. (MarketsandMarkets, FB 6183).

2.2 Wastewater Treatment

Due to the combination of photocatalytic degradation and membrane separation, the submerged membrane photocatalytic reactor (SMPR) has been used for many years. A significant amount of research effort has gone into developing a long-lasting, high-efficiency, low-energy-consumption sustainable photocatalytic membrane reactor (PMR). In general, there are two types of SMPR reactors: **i)** reactors with catalysts suspended in the feed solution and **ii)** reactors with catalysts immobilized in or on the membrane. In practice, the reactor with suspended catalysts has some drawbacks, such as inefficient photocatalyst recovery and additional operating costs. The second type is preferable because the photocatalyst recovery process can be greatly simplified, reducing operational complexity and cost in practical applications (Ong, Lau, Goh, Ng, & Ismail, 2014). As a result, the immobilization of photocatalysts in a membrane matrix with vacuum pressure driving through permeate side under UV irradiation is presented as a hybrid submerged membrane photocatalytic reactor (SMPR) (Ong et al., 2014).

C.S. Ong et al. (2014) evaluated the performance of a submerged membrane photocatalytic reactor (SMPR) composed of PVDF hollow fiber membranes and TiO_2 for the separation and degradation of synthetic oily wastewater under UV irradiation. The effects of operating variables like TiO_2 catalyst content in the membrane, membrane module packing density, feed oil concentration, and air bubble flow rates on permeate flux, oil rejection, and TOC degradation were investigated. TOC degradation was compared using direct photolysis, a neat PVDF membrane, and a PVDF-TiO_2 membrane. The TOC degradation rate of a PVDF-TiO_2 membrane was significantly higher than that of a pure PVDF membrane. The oil components in the wastewater could be effectively degraded in the presence of TiO_2 under UV irradiation, according to gas chromatography-mass spectrometry (GC-MS) analyses. Using a PVDF membrane embedded with 2% TiO_2 at 250 ppm oil concentration, a module packing density of 35.3%, and air bubble flow rates of 5 L/min, the average flux of the membrane was reported to be around 73.04 L/m^2 h. Under these optimized conditions, remarkable TOC degradation and oil rejection rates of up to 80% and over 90%, respectively, could be achieved. The research findings presented in this paper are useful for the study of simultaneous separation and degradation of oily wastewater, as well as for the development of a hybrid SMPR (Ong et al., 2014).

The significant benefits of using ultrafiltration (UF) membrane technology for the treatment of oily wastewater include high efficiency in removing oil droplets, low energy consumption, minimal chemical use (only for cleaning), and no by-product production.

Yu et al. (2020) tested the performance of ultrafiltration (UF) as a process upgrade in a large-scale drinking water treatment plant over 7 years. The results showed that the UF system improved the rejection and organic removal rates of contaminants. The UF's average annual permeate flux decreased, necessitating the use of various cleaning approaches with increasing frequency to maintain stable operation. After 7 years of operation, the total cost per cubic meter of water for the UF system increased by 55%. Membrane replacement costs, energy consumption, wastewater, and chemicals all contributed significantly to the UF system's life cycle operating costs. During long-term UF filtration, membrane fouling worsened, and the organic composition of the drinking water was altered. During the 7 years of operation, the excellent integrity of the UF membrane was sufficient to prevent organics from the metabolic activity of microorganisms and to maintain the biological safety of the permeate.

2.3 Artificial Intelligence (AI)

Artificial intelligence (AI) is a rapidly developing, innovative technology that can simulate complex real-world issues. The automation of these facilities produced simple, inexpensive operations as well as a notable decrease in the incidence of human mistakes, making the modeling capabilities of AI techniques highly helpful in the processes of water purification and wastewater treatment. It is typically impossible to model multi- or non-linear interactions and process dynamics using traditional modeling techniques (Safeer et al., 2022). Safeer et al. (2022) present a comprehensive summary of recent developments and discoveries in various AI technologies applied to source water quality determination, coagulation/flocculation, disinfection, membrane filtration, desalination, modeling wastewater treatment plants, membrane fouling prediction, removal of heavy metals, and monitoring of biological oxygen demand (BOD) and chemical oxygen demand (COD) levels. This review's investigation of the effectiveness of various AI technologies shows that these technologies have been successfully integrated into applications connected to water treatment. It also draws attention to the drawbacks that prevent their use in actual water treatment systems (Safeer et al., 2022).

According to Safeer et al. (2022), non-linearity and complicated process dynamics make it challenging to model and predict water quality. As a result, several research fields view the application of improved AI models as a workable strategy for analyzing historical data. The effectiveness of treatment and the security of water use for end users can both be enhanced by models and forecasts of water quality variables. The protection of living organisms and environmental pollution depends on wastewater treatment. The removal of heavy metals and dyes from waste effluents is done using the effective and inexpensive techniques of membrane filtering and adsorption. Artificial intelligence techniques perform better when simulating complicated and nonlinear engineering problems. The removal efficiency was employed as the output of ANN, with the typical input variables being pH, concentration, treatment duration, the effective area of membrane and adsorbent, and temperature. This review's extensive presentation of the use of ANNs, SVM, GA, and DL in various water treatment processes, including source water quality determination, coagulation/flocculation, disinfection, desalination, and BOD and COD determination, demonstrated that AI is a rapidly emerging tool in the field of water and wastewater treatment that can overcome significant drawbacks of traditional modeling techniques and lessen the likelihood of human error. It also demonstrates how hybrid AI models that incorporate several AI algorithms perform better and are more predictable (Safeer et al., 2022).

Natural systems monitoring and management, water, and wastewater treatment applications, and water-based agriculture such as hydroponics and aquaponics have all been optimized, modeled, and automated using artificial intelligence techniques and machine learning models. Artificial intelligence and machine learning (AI/ML) technologies are projected to better improve water-based applications and reduce capital costs in addition to giving computer-assisted assistance to complicated problems surrounding water chemistry and physical/biological processes (Lowe, Qin, & Mao, 2022). A cross-section of critically important, peer-reviewed water-based applications that have been combined with AI or ML, such as automation and monitoring for aquaponics and hydroponics, adsorption, membrane filtration, water-quality index monitoring, water-quality parameter modeling, and river level monitoring (Lowe et al., 2022). With the AI/ML models, and smart technologies, including the Internet of Things (IoT), sensors, and systems based on these technologies that are reviewed by Lowe et al. (2022), success in control, optimization, and modeling has been achieved; however, significant challenges and limitations were prevalent and common throughout. Poor model reproducibility and standardization, poor data management, limited explain ability, and a lack of academic transparency are all significant obstacles that must be overcome to successfully execute these intelligent applications. To get past these obstacles and keep implementing these potent tools successfully, recommendations are made to help with explain ability, data management, reproducibility, and model causality (Lowe et al., 2022).

The classification of water pollution and the three conventional treatment approaches of precipitation or encapsulation, adsorption, and membrane technologies, including electrodialysis, nanofiltration, reverse osmosis, and other artificial intelligence technologies, were covered by Altowayti et al. in 2022. The application and varied performance of the treatment have been properly handled. The management and treatment of wastewater must be done correctly. The environment can benefit from both centralized and decentralized water management schemes. To effectively manage wastewater, authorities must design the system in accordance with the environment. Overall, the public must take part in bettering water and wastewater management (Altowayti et al., 2022).

Efforts are being made to commercialize membrane reactor technologies that have already been developed in the laboratory and on a pilot scale. To achieve this goal, membrane reactors must be highly efficient, as well as long-lasting, dependable, and cost-effective. Artificial intelligence (AI) promotion is expected to have a positive impact on these criteria. Kamali et al. (2021) discussed the advantages and disadvantages of AI-based models used in wastewater treatment from various sources. Existing gaps in membrane reactor technology commercialization are discussed to provide an overview of the state of the art for future research. The findings and discussions presented in this review demonstrate that artificial intelligence models can be used to predict the performance of membrane reactor technologies used to recover clean water from polluted sources. However, more work is needed to achieve an excellent match between AI-based predictions and experimental results when treating very strong and highly polluted effluents. This could be accomplished by modifying and integrating existing AI-based methods. Additionally, one of the top priorities for promoting the use of membrane reactor technologies at full scale is the development of suitable variables to optimize the performance of membrane reactors and improve their efficiency in treating recalcitrant pollutants like contaminants of emerging concern (CECs) (Kamali et al., 2021).

Because of its ability to solve real-world problems for which deterministic solutions are difficult to obtain, AI is a capable tool that is commonly used in multidisciplinary engineering. Recently, there has been a revolution in the automation of water treatment and desalination processes. The water sector presents several challenges related to data structuring and smart water services, which AI could greatly benefit from once these issues are resolved. Artificial Neural Networks (ANNs) as a regression model and Genetic Algorithm (GA) as one of the global optimization techniques have been widely used in desalination and water treatment for a variety of applications. Among the many applications are the modeling of desalination and water treatment processes, as well as the optimization of operating conditions. Al Aani et al. (2019) highlighted the comparison of ANNs with conventional modeling approaches, as well as the shortcomings and challenges that should be associated with these common tools in some practical applications of complex nature. It was determined that the use of AI tools will undoubtedly pave the way in the water sector toward better operation, process automation, and water resource management in an increasingly volatile environment (Al Aani et al., 2019).

In comparison to the traditional approach, the literature review revealed the advantages of AI in extracting nonlinear models for the various operating conditions of the wastewater treatment process and the relationship between them. Al Aani et al. (2019) also provided a brief comparison of traditional approaches and AI, demonstrating that artificial intelligence is an effective tool in wastewater treatment engineering applications. Furthermore, research has shown that benchmarking ANNs against other classical modeling approaches and AI tools may reveal a greater potential to be used to create optimal operating conditions, particularly in complex operational circumstances.

Artificial neural networks (ANNs) are black-box models that are becoming more popular in prediction due to their high reliability and short computational time. Jawad et al. (2020) create a multilayer neural network model to predict permeate flux in forward osmosis. The developed model's generalizability is tested by incorporating laboratory-scale experimental data from several published studies. There are nine input variables to consider. The development of the optimal network architecture is aided by research into the effect of the number of neurons and hidden layers on neural network performance. The best-trained network had a coefficient determination value of 97.3%. Furthermore, the model's generalized validation and predictive ability are tested against untrained published data. For comparison to the ANN model, a simple machine learning technique such as the multiple linear regression (MLR) model is used. The ANN model outperformed the MLR model in terms of forming a complex relationship between input and output variables (Jawad et al., 2020).

Because of its robust autonomous learning and ability to solve complex problems, artificial intelligence (AI) has increasingly demonstrated its potential to solve the challenges faced in water treatment. AI technology assists in the management and operation of water treatment processes, which is more efficient than relying solely on human operations. AI-based data analysis and scalable learning mechanisms have the potential to establish a universal platform for process analysis and predictive models, as well as achieve water quality diagnosis, autonomous decision-making, and operation process optimization (Li et al., 2021).

The challenges that AI technologies face, as well as the issues that require further investigation, can be summarized as follows:

- With the help of AI technologies, obtain more effective characterization data to screen and identify targeted contaminants in the complex background.
- And establish a macro-intelligence model and decision scheme for the entire water treatment plants to support water supply system management (Li et al., 2021).

2.4 Machine Learning (ML)

The diagram below depicts AI and ML technologies that are commonly used in water treatment. Classification, regression, dimensionality reduction, and clustering are four common applications of machine learning. Based on the process data, ML methods used to solve regression and classification problems construct predictive models. In other words, applying an ML method entails **1)** estimating the relationship between the system's input variables and the target output from a given dataset, i.e., the training process, and **2)** using the estimated nonlinear relationship to predict the new system output. Because the goal of ML is to predict rather than estimate, it is well suited for water quality prediction (Li et al., 2021).

Nonlinear classification and regression analysis are handled by ANN, deep learning (DL), support vector machine (SVM), and random forest (RF), while principal component analysis (PCA) is used to handle high-dimensional data. ANN and DL both mimic the behavioral characteristics of animal neural networks and perform distributed parallel information processing. The distinction is that DL has a more complex structure and, in general, better prediction accuracy than ANN. However, DL requires more data to train the network and is more prone to overfitting. Unlike classical ANN, SVM is based on a more robust mathematical theory, which improves the model's interpretability. Furthermore, SVM can be transformed into a convex optimization problem, ensuring the algorithm's global optimality and avoiding the difficult-to-solve local minimum problem for neural networks. SVM models, on the other hand, are difficult to train on a large-scale sample of data and require a long training time. RF has a significant advantage over ANN, DL, and SVM in that it can evaluate the importance of variables while performing classification or regression analysis. Although the weight of the ANNs can be used to assess the importance of the input variables, a more complex statistical analysis is required. The four ML methods discussed above have a strong ability to handle nonlinear relationships, making them suitable for developing predictive models in water treatment when the background substances are complex and there are unknown interactions between these substances, such as predicting specific contaminants in source water, membrane fouling, and analyzing disinfection by-product precursors. Furthermore, the ML method is chosen and applied based on the prediction model's target demand, the size of the existing dataset, and the characteristics of various ML methods (Li et al., 2021).

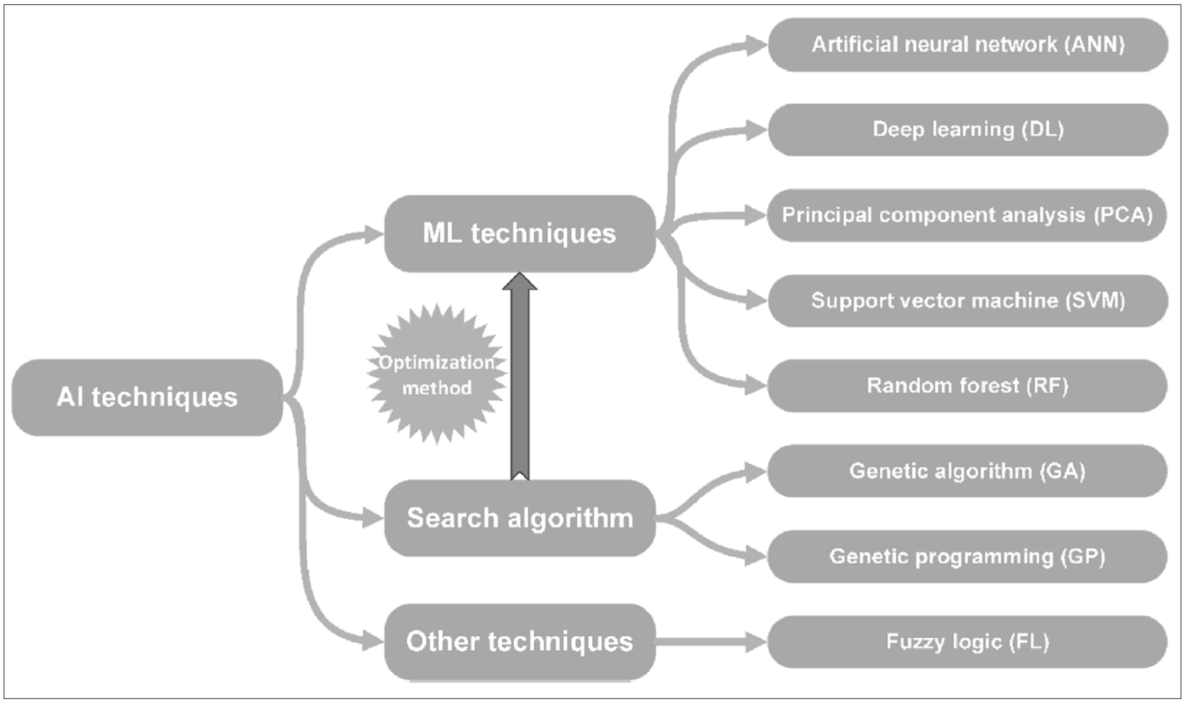


Figure 2.1 Classification of AI and ML technologies used in water treatment taken from Li et al. (2021, p. 4)

2.5 Search Algorithms

Search algorithms are a technique for determining the best solution to multiple-solution problems given certain constraints. Genetic algorithms (GA) and genetic programming are two search algorithms that are commonly used in water treatment (GP). Both are global optimization methods that simulate the process of biological evolution in nature, with the main distinction being the algorithm representation. GA solutions are typically represented by a fixed-length string, whereas GP solutions are typically represented by a tree structure of nested data structures. Search algorithms are used in water treatment and supply to solve optimization problems such as determining the best feed ratio when adding multiple coagulants and locating pollution in the water distribution system (Li et al., 2021).

2.6 Fuzzy Logic (FL)

Fuzzy logic (FL), a multivalued logic-based method, studies fuzzy judgment using fuzzy sets, allowing FL-based fuzzy inference systems to simulate the human brain and implement natural inference. When compared to the common ANN, the adaptive neural fuzzy inference system (ANFIS) composed of FL and ANN with an inference mechanism has a high interpretability. Coagulant dosing systems have been controlled using the combined model (Li et al., 2021).

2.7 Transfer Learning (TL)

Transfer learning is an effective strategy for dealing with small amounts of data and increasing the power of models for specific tasks. As a result, it has advantages for wastewater treatment research. However, there is no comparison between with and without transfer learning used in some applications because models without transfer learning cannot be obtained for such small data. Furthermore, rather than just computational models, transfer learning methods benefit real-world discovery use cases. The use of transfer learning in wastewater treatment discovery is still in its early stages, with much more research needed in terms of related theoretical studies. For example, there is no standardized metric for assessing the effectiveness of transfer learning methods. In practice, transfer learning is typically evaluated based on the model's performance on specific tasks such as accuracy improvement or error reduction. As a result, making comparisons across applications is difficult. The observed performance improvements in some cases of feature-based TLs may be due in part to increased network complexity. Furthermore, there is currently no appropriate benchmark dataset for evaluating TL applications in wastewater treatment, and due to the small data sizes in TL scenarios, overfitting issues are expected to be a major concern. Furthermore, publications that include in-depth discussions of transfer learning methodologies as they apply to wastewater treatment discovery are scarce (Cai et al., 2020).

There are currently several obstacles to the practical application of transfer learning for wastewater treatment discovery. One of these difficulties is determining how to properly implement transfer learning methods. Although the fine-tuning strategy is useful for many types of models, the careful network structure design is a non-trivial decision. Negative transfer, or when transfer learning degrades model performance, can be caused by poor method choice. For example, when pre-trained at both the node and graph levels, the GNN performed well but gave a negative transfer when only pre-trained at the graph level. Finally, there are currently no common criteria for selecting transfer learning methods because evaluating transfer learning methods without theoretical support is difficult. However, there are some general guidelines for fine-tuning. Fixing some layers, for example, is a good way to avoid overfitting when the target data is too small. When the target data is larger, fine-tuning all layers is still the best option. Feature-based methods have demonstrated their utility as strategies for resolving wastewater treatment discovery issues (Cai et al., 2020).

2.8 Principal Components Analysis (PCA)

There is an urgent need to develop predictive methodologies that will accelerate membrane industrialization. Predicting membrane performance, on the other hand, has proven difficult due to a large number of potential variables and the complex interaction relationship between the functional variables and the membrane. As a result, rather than developing traditional mathematical equations, Hu et al. (2021) compiled a large dataset and developed AI-based predictive models for rejection and permeate flux from a collected dataset containing 38,430 data points with more than 18 dimensions (variables). The researchers conducted an extensive principal component analysis (PCA) to highlight the important variables affecting membrane performance, which revealed that the factors affecting permeate flux and rejection are surprisingly similar. Three different AI models (ANN, SVM, and RF) were trained to predict membrane performance with unprecedented accuracy, up to 98% for permeate flux and 91% for rejection rate. The findings pave the way for proper data normalization, which will allow for better performance prediction as well as better membrane design and development (Hu et al., 2021).

Naessens et al. (2017) validated a fouling monitoring methodology based on principal component analysis (PCA) using data from a pilot-scale pressurized ultrafiltration (UF) system operating with seawater. The evolution of membrane fouling was investigated in order to understand its relationship with the cleaning strategy used on the one hand and the quality of raw seawater on the other. Furthermore, it was demonstrated that using PCA as a monitoring technique to detect abnormal fouling behavior is a reliable tool. Instead of running the system with fixed cycles, decisions on cleaning sequences or frequencies could be made dynamically using PCA. PCA can be used to visually represent the current process state, easily detect outliers, and compare UF performance under different operational conditions (Naessens et al., 2017).

Fouling is a significant barrier to maintaining efficient membrane water treatment processes. Peiris et al. (2010) used a fluorescence excitation-emission matrix (FEEM) approach to characterize major membrane fouling. This method is capable of producing fast and consistent analyses with high instrumental sensitivity. Principal component analysis (PCA) was used to extract principal components containing information relevant to membrane fouling from fluorescence EEM measurements collected during crossflow ultrafiltration of river water. These key elements were linked to the major membrane foulants. The extracted membrane foulants' fluorescence EEMs also revealed different rejection characteristics for two different membranes. The proposed method was found to be appropriate for identifying the major fouling components and their contributions to reversible and irreversible membrane fouling, demonstrating its potential for monitoring and controlling membrane fouling in water treatment applications (Peiris et al., 2010).

2.9 Design Of Experiments (DOE)

In recent years, there has been an increase in research on filtration processes for removing contaminating compounds from water and wastewater. To optimize the performance of nanofiltration in the treatment of antibiotic-containing wastewater, Souza et al. (2020) used a 2k factorial design with five control factors. To forecast discharge rates and permeate fluxes, a multiple linear regression model was used. To validate the developed model, additional factorial designs were performed using the same membrane and contaminants under different conditions. Because of the high agreement between predicted and experimental values, the developed model can predict the performance of nanofiltration when treating antibiotic-containing wastewater (Souza et al., 2020).

Dopar et al. (2011) described using the photo-Fenton process to treat simulated industrial wastewater after pretreatment with the dark-Fenton process. The efficiency of the process was investigated, which is dependent on several important process variables such as initial pH, iron catalyst and oxidant concentration, and type of UV irradiation. A 3-factor, 3-level Box-Behnken experimental design combined with response surface modeling was used to consider the combined effects of the studied process variables. Quadratic models were developed to predict the mineralization of simulated industrial wastewater by the applied photo-Fenton processes. The significance of the models and model components was assessed using ANOVA. The results showed that both models were highly accurate and predictive for the mineralization of simulated industrial wastewater. The results showed that both photo-Fenton processes could successfully treat the simulated industrial wastewater studied (Dopar et al., 2011).

In terms of spinning conditions, Xuezhong He (2017) investigated the spinning variables of cellulose acetate (CA) hollow fiber membranes under various conditions. Based on the experimental design method and multivariate analysis, the spinning variables of the air gap, bore fluid composition, bore fluid flow rate, and quench bath temperature were optimized. The optimal spinning condition for producing high polymer content, symmetrical cross-section, and highly cross-linked CA hollow fibers was identified. The findings can be used to direct the production of defect-free CA hollow fiber membranes with the desired structures and properties (Xuezhong He, 2017).

2.10 Conclusion

The nonlinearity and complicated process dynamics make it challenging to model and predict water quality. As a result, several research fields perceive the application of improved AI/ML models as a practical approach for analyzing historical data. The effectiveness of treatment and the sustainability of water use for end users can both be enhanced by models and forecasts of water quality variables. The protection of organisms and environmental pollution depend on wastewater treatment.

AI/ML approaches perform better when simulating complicated and nonlinear engineering problems. Different methods of water treatment demonstrated that AI/ML is a cutting-edge tool in the field of water and wastewater treatment that can overcome significant drawbacks of traditional modeling techniques and lower the likelihood of human mistakes. It also emphasizes how hybrid AI/ML models that incorporate several AI/ML techniques perform better and are more predictable.

Smart technology and AI/ML can be utilized to clarify and comprehend some of the most complicated problems affecting the water-based sectors. Some of the most important applications in water-based organizations and operations, including as those of water-treatment and wastewater-treatment facilities, natural systems, and water-based agriculture, have been successfully optimized, predicted, modeled, and controlled using AI/ML approaches.

Even if many of the research have been successfully published and reviewed, there are still a number of difficulties and restrictions with them. To further these intelligent applications, significant hurdles such as data management, public or governmental perspectives, predictability, and research transparency must be overcome. Although these difficulties and restrictions are undoubtedly evident, they do not negate the present research and advancements that indicate that AI/ML approaches and smart technologies have significant implications and possibilities for one of the most valuable resources on our planet.

CHAPITRE 3

MEASUREMENT PHASE

3.1 Experimental Details

To properly plan and perform a full-scale commercial process, small-scale research is carried out to assess the feasibility, time, scaling factors, unexpected results, further improve a process, etc.

For the most part, evaluating the performance of the membrane still involves trial and error. The selection of important operating factors, determining fiber spinning settings, and material optimization are all covered. This work makes use of extensive experimental data on a membrane assessment. But evaluation for a particular application can still be quite challenging (Yuliwati, 2012; Chen, & Kim, 2006; Yuliwati et al., 2012). Asymmetric hollow fiber membranes made of hydrophobic Polyvinylidene Fluoride (PVDF) were used in this research (Figure 3.1). Refineries for wastewater treatment use these membranes. For this study, dry-jet wet-spinning was utilized to fabricate the membranes.

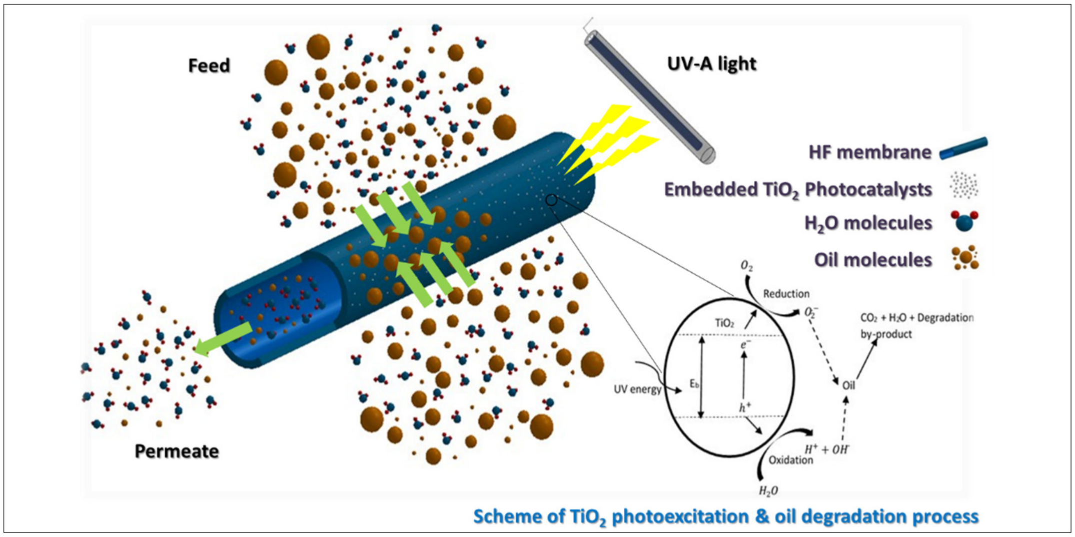


Figure 3.1 Schematic diagram of oily wastewater separation and degradation taken from Ong et al. (2014, p. 48)

A submerged membrane reactor was constructed manually. The measurements are $18cm\ (W) \times 20cm\ (L) \times 40cm\ (H)$ (Figure 3.2). Figure 3.3 depicts a schematic diagram of the laboratory pilot submerged membrane photocatalytic reactor system with submerged UV-A lamp, for semi-batch wastewater treatment operation. To reduce fouling at room temperature, two hollow fiber modules were immersed in the feed tank (14 L) and air bubbles were generated from the bottom. The oily wastewater will be filtered using hollow fibers. The cleaned permeate was sucked through a peristaltic pump and collected in a permeate tank.

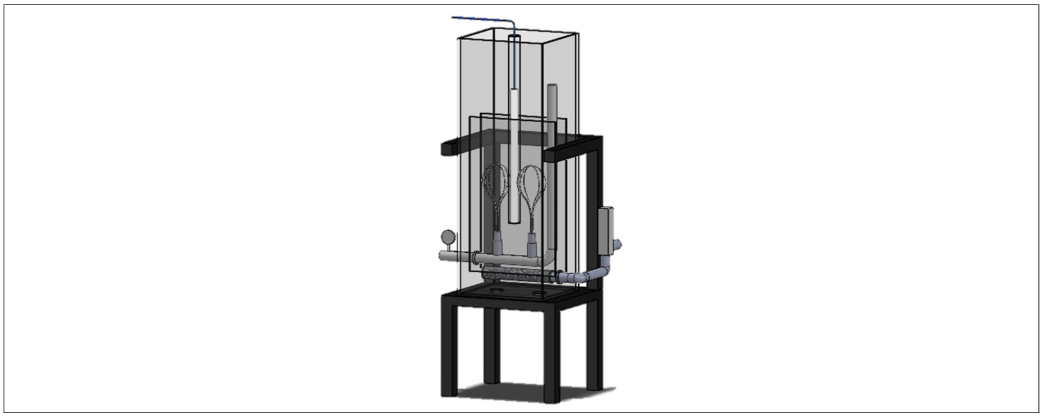


Figure 3.2 Handmade Submerged Membrane Photocatalytic Reactor (SMPR) taken from Ong et al. (2014, p. 50)

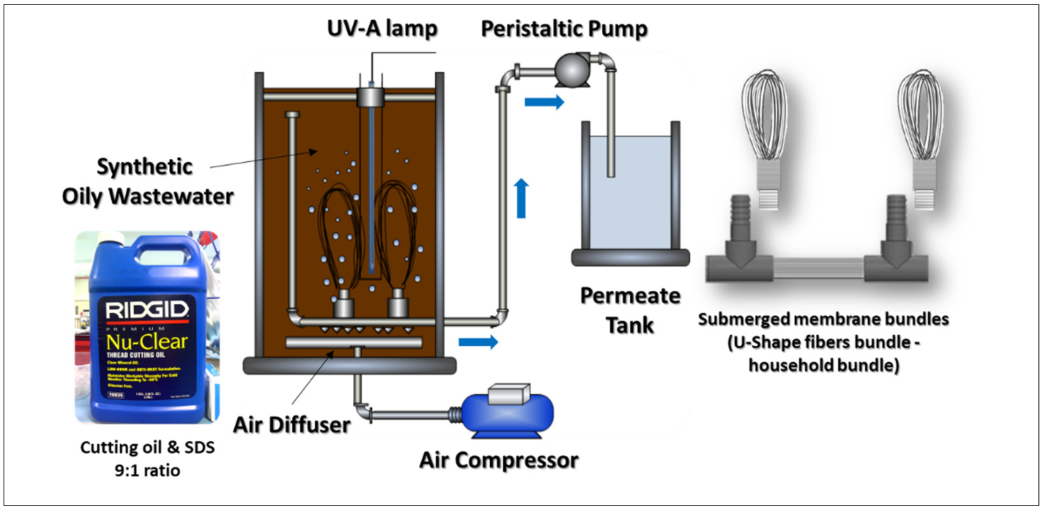

UV-A lamp
Peristaltic Pump
Synthetic Oily Wastewater
RIDGID
PREMIUM
Nu-Clear
THREAD CUTTING OIL
Cutting oil & SDS 9:1 ratio
Air Diffuser
Air Compressor
Permeate Tank
Submerged membrane bundles (U-Shape fibers bundle - household bundle)

Figure 3.3 Schematic diagram of the SMPUF system

There is a large number of factors that affect UF membrane performances (Busch, Cruse, & Marquardt, 2007). Polymer characteristics, orientation, feed properties, operating conditions, and others. For simplicity, these factors are categorized into three groups: membrane composition, membrane spinning conditions, and **feed properties**. The experiments covered only feed properties' effect on membrane performances, leaving the other factors fixed (unchanged) throughout all experiments. For reference, the constant factors in Table 3.1 for membrane composition, and Table 3.2 for membrane spinning condition are listed below.

Table 3.1 Membrane composition

Chemical Types	**Wt%**
PVDF	18
TiO_2	2
PVP	5
DMAc	75
	100

Table 3.2 Spinning condition of the hollow fiber membrane

Spinning Parameters	**Value**
[a] O.D. / I.D. of spinneret (mm/mm)	1.15/0.55
Dope flow rate (cm^3/min)	10.5
Bore fluid rate (cm^3/min)	3.5
Bore fluid temperature (°C)	27
Air gap distance (cm)	3
External coagulant	Tap water
Coagulant temperature (°C)	27
Wind-up drum speed (Hz)	3.8

[a] O.D. / I.D. = Outer diameter/Inner diameter

3.2 Membrane and Module Preparation

After being dried for 24 hours at 50°C in the oven, 18 weight percent of PVDF was initially added to pre-weighed DMAc solvent. After that, mechanical stirring was used to swirl the mixture at 600 rpm until all the polymeric pellets had been dissolved. To produce dope solutions with TiO_2 loading, 5 weight percent of PVP and 2 weight percent of TiO_2

nanoparticles were added after that. Finally, before spinning, the dope solution was ultrasonically deflated to release any trapped air bubbles.

The dry-jet wet-spinning technique was used to spin PVDF hollow fiber membranes using the produced solution. Table 3.2 shows the specifics of the spinning procedure. To eliminate any remaining solvent, the as-spun hollow fibers were submerged in the water bath for two days. To reduce fiber shrinkage and pore collapse, the fibers were post-treated with a 10 weight percent glycerol aqueous solution for a day before air drying. Before module fabrication, hollow fibers were lastly dried at room temperature for three days.

A bundle of 30, 60, and 90 hollow fibers, each measuring around 28 cm in length (total effective membrane area is 607 cm^2), were potted into a PVC tube using epoxy resin (E-30CL Loctite® Corporation, USA) to create a membrane module. The module preparation was finished by cutting the protruding portions and fixing them into a PVC adaptor after the module had been left to solidify at room temperature.

3.3 Synthetic Wastewater Preparation

By combining distilled water and commercial cutting oil (RIDGID Nu-Clear Cutting Oil, #70835, Ridge Tool Company), synthetic cutting oil wastewater was produced. Cutting oil in the range of 250 to 10000 ppm and sodium dodecylbenzenesulfonate (SDS) were combined in a 9:1 ratio to produce the emulsion. The solution was then mixed for two minutes at room temperature with a high-speed mixer (Model: BL 310AW, Khind).

3.4 Studied Factors

The research concentrated on studying and optimizing five of the most important feed factors affecting membrane performances. The factors studied during experiments are **pH** - hydrogen activity, **OC** - oil concentration, **UVIT** - ultraviolet irradiation time, **ABFR** - air bubbling flow rate, and **MPD** - module packing density. The factors' composition and quality levels are summarized in Table 3.3.

Turbulence of feed solution near the hollow fibers was promoted by air bubbling generated by a diffuser located beneath the submerged membrane bundles for mechanical cleaning. To investigate the effect of ABFR, an airflow meter was used to keep the flow rate within a predetermined range. Table 3.3 shows how the feed factors ABFR, OC, UVIT, pH, and MPD were adjusted to predetermined values. A vacuum pump provided the filtration pressure, and a flow meter was used to continuously record the permeate flow rate, as shown in Figure 3.3. As a result, the immobilization of photocatalysts in a membrane matrix with vacuum pressure driving through permeate side under UV irradiation was presented as a hybrid submerged membrane photocatalytic reactor (SMPR).

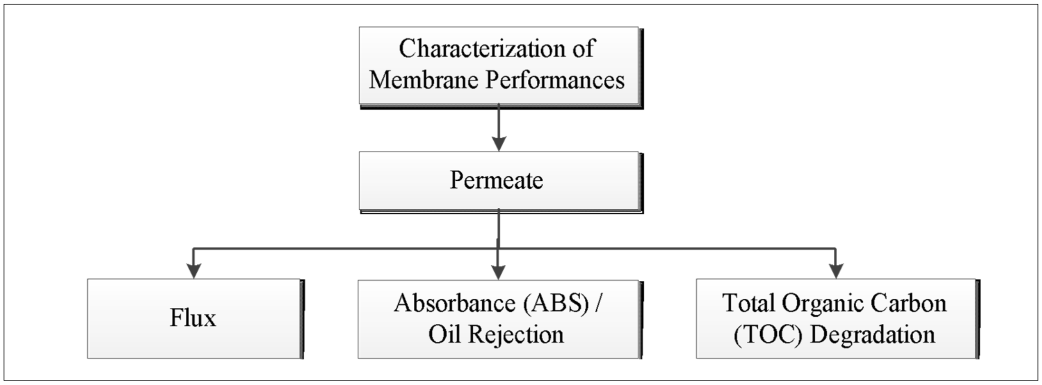
Between two bundles, a UV-A blacklight blue filter lamp (8W) with a maximum light intensity of 365 nm was jacketed in a quartz tube and placed in the center of the feed container. The ultraviolet irradiation time (UVIT) ranged from 30 to 480 minutes. The filtration process continued under controlled conditions depending on the time selected between the predetermined range. To degrade the oil attached to the membrane surface, ultraviolet (UV) light (8W) was irradiated on it. By exposing the membrane to UV-A light for up to 8 hours, the effect of the UV irradiation period on the intrinsic and separation properties of the composite membrane composed of organic Polyvinylidene Fluoride and inorganic Titanium Dioxide (TiO_2) nanoparticles was evaluated.

Meanwhile, since the experiments lasted for a long time, to ensure that the characteristics of the feed solution remained unchanged over time, feed solution was added from another reservoir to the photocatalytic reaction vessel to maintain the volume of the synthetic cutting oil solution to about 14 L.

Table 3.3 Minimum, maximum, and optimum value of the SMUF system variables

Input & Output Factor	Objective	Actual Value Range
x_1 (ABFR - L/min)	In Range	$0 < x_1 < 5$
x_2 (OC - ppm or mg/L)	In Range	$250 < x_2 < 10{,}000$
x_3 (UVIT - min)	In Range	$30 < x_3 < 480$
x_4 (pH)	In Range	$4 < x_4 < 10$
x_5 (MPD - n or number)	In Range	$30 < x_5 < 90$
y_1 (Permeate Flux)	**Maximize** (L/m^{2}h)	**100** (L/m^{2}h)
y_2 (ABS / Oil Rejection)	Minimize (mg/L) or **Maximize** (%)	0 (mg/L) or **100** (%)
y_3 (TOC Degradation)	Minimize (mg/L) or **Maximize** (%)	0 (mg/L) or **100** (%)

Figure 3.4 depicts membrane performance characteristics that are relevant to the measure. These are **flux** or permeate volume per unit area and unit time, oil rejection or absorbance (**ABS**), and total organic carbon (**TOC**) degradation. The permeation flux (***F***) represents SMUF productivity. And the rejection rates (***R₁*** and ***R₂***) of PVDF hollow fibers are related to permeate water quality, which consists of absorbance (ABS) and total organic carbon (TOC) degradation. The effect of the feed factors (i.e., ABFR, OC, UVIT, pH, and MPD) on membrane performance characteristics (i.e., Flux, ABS, and TOC) was studied to obtain optimum operating conditions for the SMPUF system (Figure 3.5).



Characterization of Membrane Performances

Permeate

Flux

Absorbance (ABS) / Oil Rejection

Total Organic Carbon (TOC) Degradation

Figure 3.4 Characterization of membrane performances

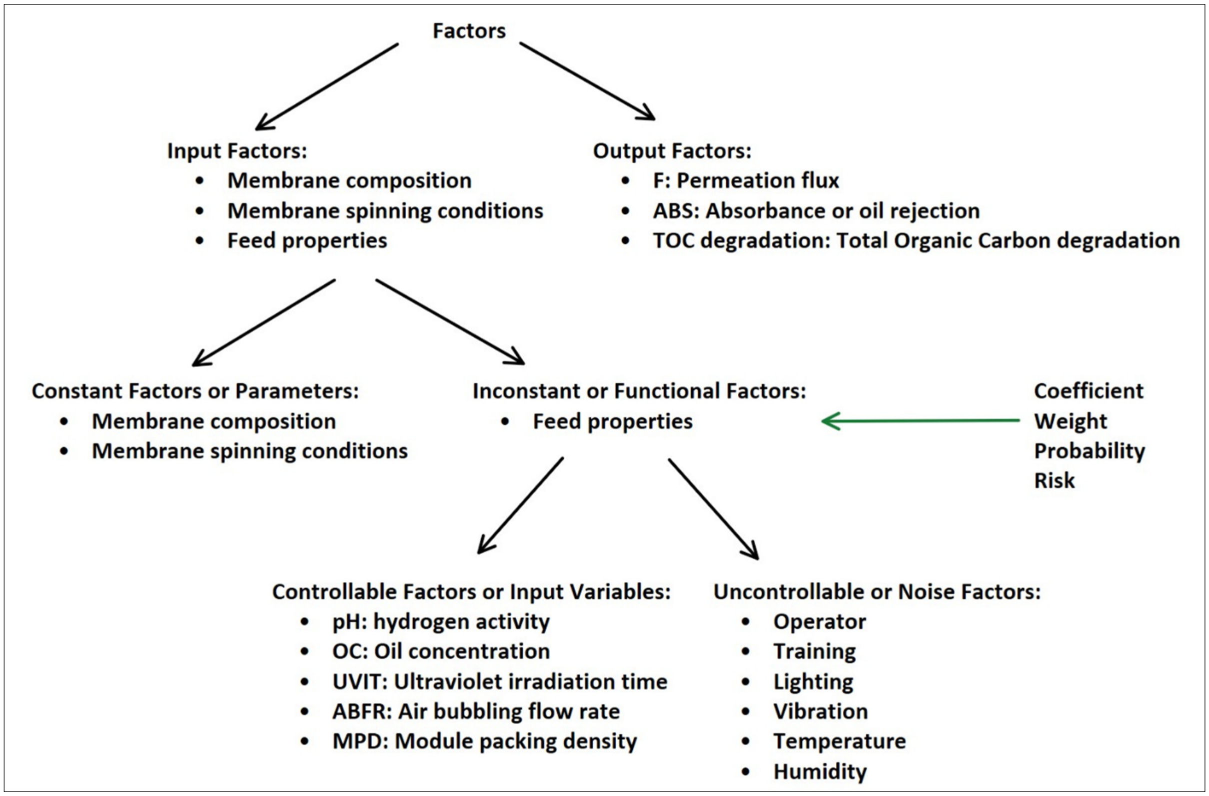

Factors
Input Factors:
Membrane composition
Membrane spinning conditions
Feed properties
Output Factors:
F: Permeation flux
ABS: Absorbance or oil rejection
TOC degradation: Total Organic Carbon degradation
Constant Factors or Parameters:
Membrane composition
Membrane spinning conditions
Inconstant or Functional Factors:
Feed properties
Coefficient
Weight
Probability
Risk
Controllable Factors or Input Variables:
pH: hydrogen activity
OC: Oil concentration
UVIT: Ultraviolet irradiation time
ABFR: Air bubbling flow rate
MPD: Module packing density
Uncontrollable or Noise Factors:
Operator
Training
Lighting
Vibration
Temperature
Humidity

Figure 3.5 SMPUF Factors Classification

The characteristics of oily wastewater produced by a synthetic refinery for the experiments and used as a feed solution in submerged ultrafiltration were very similar to those of typical refinery wastewater shown in Table 3.4. Equations 3.1, 3.2, and 3.3 are used to calculate the output permeate flux (***F***) and rejection (***R₁*** and ***R₂***) (Ong et al., 2014).

$$(Permeate\ Flux)\ F = \frac{V}{At} \tag{3.1}$$

Where ***F*** is the permeate water flux (L/m^2 h), V is the permeate volume (quantity) (L), A is the membrane surface area (m^2), and t is the time to obtain the permeate volume (h) (Equation 3.1).

$$(Absorbance\ /\ Oil\ Rejection)\ R_1 = (1 - \frac{c_p}{c_f}) \times 100 \tag{3.2}$$

Where $\boldsymbol{R_1}$ is the oil rejection (ABS) in the ultrafiltration process (%), c_p is the concentration of oil in the permeate (ppm), and c_f is the concentration of oil in the feed solution (ppm) (Equation 3.2).

$$(TOC\ Degradation)\ R_2 = (1 - \frac{TOC_t}{TOC_o}) \times 100 \qquad (3.3)$$

Equation 3.3 demonstrates how to determine the efficiency of the photocatalytic degradation of TOC (%) (Ong et al., 2014). Since photocatalytic reaction takes place only in the feed chamber by the TiO_2 particles embedded in the hollow fiber membranes, the samples were collected from the feed solution and subjected to TOC analysis. Thus, in equation (3.3) TOC_0 and TOC_t are the TOC (ppm) at time zero and TOC (ppm) at time t, respectively, in the feed. TOC was measured by a TOC analyzer.

F represents SMUF **productivity**, and $\boldsymbol{R_1}$ and $\boldsymbol{R_2}$ represent, respectively, the **filtration efficiency** of hollow fiber UF membrane and **photocatalytic degradation efficiency** of TiO_2 catalysts.

TOC degradation samples were collected from feed, whereas oil rejection samples were collected from permeate. Because most oil particles are unable to pass through the membrane, if samples are only collected from permeate for TOC degradation and oil rejection measurements, the TOC degradation will be very high, but this is not due to the photocatalytic degradation being very efficient, but rather because the oil particles are blocked by the pores on the membrane surface.

The value ranges of the feed factors are listed in Table 3.3. Based on the ranges, treatment combination records were obtained. Experimentally obtained values of the five independent feed factors combined with the three responses are tabulated in APPENDIX and Table 4.2.

Table 3.4 Composition of synthetic refinery wastewater which was used as feed solution and the national discharge standard for refinery wastewater

Constituent, unit	Feed Composition	Environment Quality (Industrial Effluent) Regulation 2009 - Standard B
pH	6.7	5.5 - 9.0
Oil and Grease, mg/L	200	10
TOC, mg/L	243	75
COD, mg/L	398	200
TSS, mg/L	543	100
NH_3-N, mg/L	27.6	20
Sulfide, mg/L	0.08	0.5

3.5 Description of Equipment Used

To fabricate hollow fiber membranes, the spinning machine was utilized. Membrane performance was tested by the submerged membrane ultrafiltration (SMUF) system. To mix the synthetic feed solution, a high-speed mixer (Model: BL 310AW, Khind) was used. Stopwatch, pH meter, UV-A blacklight blue filter lamp (8W, Model: FL8BLB, Sankyo Denki Co., Ltd., Japan), air compressor (Model: 2 HP single cylinder 24 L tank, Orimas), the airflow meter to control the flow rate, and vacuum pump (Model: 77200-60, Masterflex L/S, Cole Parmer), were used when measuring variables of feed factors. The permeate flow rate was recorded using a flow meter. Using a UV-vis spectrophotometer (Model: DR5000, Hach) with absorbance measured at 294 nm, where the maximum absorption happens, the oil concentrations in permeate and feed were measured. It is observed that the relationship between absorbance and oil concentration is linear (Ong et al., 2014). For calculating unknown oil concentrations in permeate, the same relationship has been utilized. And a Total Organic Carbon Analyzer (Model: TOC-LCPN, Shimadzu Co.) were used to measure the TOC values.

In the principal component analysis (PCA) phase, to analyze the SMUF system observations, MATLAB software version R2016b was used. In the design of experiments (DOE) phase, to analyze the SMUF system orthogonal arrays (OA) based on the input variables and responses, Design-Expert software version 9 (Stat-Ease Inc., USA) was used. Furthermore, in the AI/ML phase, to identify the optimum setting of the SMUF process operating conditions to obtain optimum values of the performances using AI/ML methodology, Python programming language and its libraries were used.

CHAPITRE 4

ANALYSIS AND IMPROVEMENT PHASE

4.1 Introduction

The DMAIC problem-solving structure was utilized to design and order the project structure (i.e., Define, Measure, Analyze, Improve, and Control). The analysis and improvement phase comes after the measurement phase (Chapter 3). To keep the context consistent, this chapter also includes the control phase (verification or confirmation) following DOE and AI/ML methodologies.

This phase describes the methodologies utilized to develop, validate, and justify the project's objectives, as well as a systematic experimental strategy for analyzing, modeling, and optimizing the Submerged Membrane Photocatalytic Ultrafiltration (SMPUF) oily wastewater treatment system through principal components analysis (PCA), design of experiments (DOE), and AI/ML methodologies. The models optimized and improved in this study enable early recognition of membrane performances based on operating process conditions and identify the optimum setting for the process operating conditions to obtain optimum performance values.

4.2 Operational Framework

This phase outlines seven major steps involved in the analysis, optimization, and improvement of the SMPUF operating process conditions for the treatment of oily wastewater produced by refineries, as listed below:

i. Prepare an adequate number of observations (dataset) based on the specified SMPUF oily wastewater treatment system inputs and outputs to organize experimental observations of PCA.

ii. Reduce the dimensionality of the dataset while retaining as much information as possible about the inherent variability in the dataset based on the PCA methodology using

MATLAB and Python to facilitate visualization of all information contained in the dataset.

iii. Prepare an adequate number of records based on the specified SMPUF oily wastewater treatment system inputs and outputs to organize experimental design orthogonal arrays (OA) of DOE.

iv. Obtain an empirical model based on the design of experiments (DOE) methodology using Design-Expert software to provide statistical analysis to determine which SMPUF process inputs have a significant effect on the membrane performances and to optimize the performance of the system.

v. Prepare an adequate number of records (dataset) based on the specified SMPUF oily wastewater treatment system inputs and outputs to train, test, and verify the multi-output models of the AI/ML phase.

vi. Obtain multi-output models based on AI/ML methodology using Python to investigate the influences of simultaneous SMPUF process inputs on the membrane performances, find the best input settings to optimize responses (***forward analysis***), and recognize which input settings can produce the optimum output values (***backward analysis***).

vii. Compare MLP models to other multi-output AI/ML models, as well as MLP models to the DOE model, to determine which is more efficient, effective, and sensitive to operating variables.

4.3 Principal Components Analysis Phase

PCA provides statistical analysis in this phase to find patterns in a high-dimensional dataset, reduce its dimensionality, and facilitate visualization of all information contained in the dataset.

To perform this statistical observation, five variables of the Submerged Membrane Photocatalytic Ultrafiltration (SMPUF) system were considered (i.e., UVIT, MPD, Flux, ABS, and TOC). Hence, 44 observations were collected randomly (Table 4.1). To evaluate and analyze the observations, version R2016b of the MATLAB software was used.

When the variables are in different units or the variance of the different columns of the dataset is significant, standardizing the dataset is often preferable (Babanova et al., 2014). In other words, PCA is performed on the correlation matrix rather than the covariance matrix.

Table 4.1 The Experimental Observations of the Five Variables

Observation	**X_3: UVIT (min)**	**X_5: MPD (n)**	**Y_1: Flux (L/m^2hr)**	**Y_2: ABS (%)**	**Y_3: TOC (%)**
obs 1	30	30	17.73	87.18	34.32
obs 2	75	60	35.28	95.24	52.36
obs 3	30	90	69.99	90.64	50.00
obs 4	30	30	24.17	85.79	35.05
obs 5	30	90	77.24	90.6	30.13
obs 6	120	30	26.96	81.15	67.38
obs 7	120	30	13.48	87.38	100
obs 8	75	60	45.28	95.24	52.36
obs 9	30	90	63.18	87.97	40.05
obs 10	120	30	27.41	79.8	54.94
obs 11	120	90	80.75	87.38	23.63
obs 12	90	80	64.51	85.10	39.04
obs 13	30	30	16.13	100	20.91
obs 14	120	90	47.7	77.39	28.35
obs 15	30	30	63.58	96.1	100
obs 16	120	30	12.40	98.54	28.60
obs 17	75	60	46.42	92.89	62.72
obs 18	120	60	56.63	90.36	91.34
obs 19	120	30	9.72	86.2	69.97
obs 20	30	30	6.43	100	3.61
obs 21	75	60	51.97	95.05	66.91
obs 22	120	90	100	83.45	85.11
obs 23	120	90	58.45	90.34	34.24
obs 24	120	90	95.25	85.72	22.63
obs 25	120	30	73.09	100	86.14
obs 26	120	30	19.31	99.94	6.94
obs 27	30	90	67.14	91.28	22.72
obs 28	75	30	45.36	96.39	66.52
obs 29	120	90	34.1	77.99	18.81
obs 30	30	30	21.91	84.14	31.59
obs 31	120	30	2.4	98.54	28.6
obs 32	120	90	71.56	83.63	26.25
obs 33	30	30	57.95	98.78	91.63
obs 34	30	90	45.34	71.68	64.74
obs 35	120	90	93.33	81.12	73.16

Observation	X_3: UVIT (min)	X_5: MPD (n)	Y_1: Flux (L/m^{2}hr)	Y_2: ABS (%)	Y_3: TOC (%)
obs 36	30	90	59.06	92.12	39.80
obs 37	30	90	92.11	73.4	93.08
obs 38	120	30	65.15	99.25	97.41
obs 39	45	40	33.31	95.85	49.09
obs 40	30	90	59.06	92.12	42.41
obs 41	75	60	40.96	93.22	42.21
obs 42	30	90	100	76.82	99.26
obs 43	30	90	59.28	72.82	79.02
obs 44	30	30	1.01	86.83	61.35

4.3.1 Boxplots

Side-by-side boxplots for the SMPUF system dataset are shown in Figure 4.1. It can be observed that the individual boxplots do not contain outliers. In addition, the median and variance of the variables are almost the same.

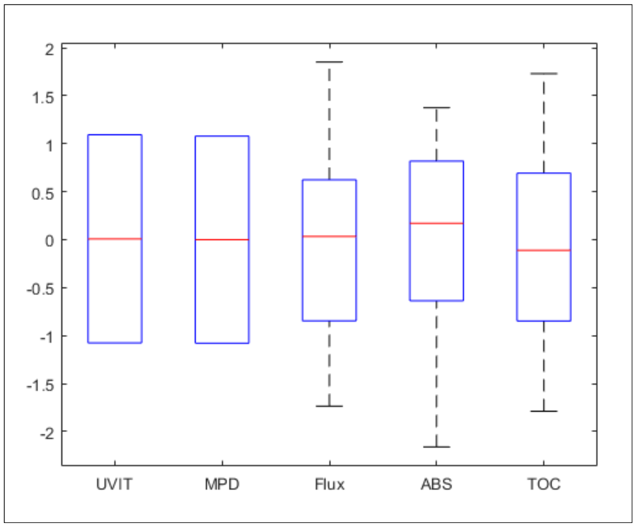


Figure 4.1 SMPUF System Boxplot

4.3.2 Scatterplot Matrix

When interpreting correlations, it is important to visualize the bivariate relationships between all pairs of variables. SMPUF System scatterplot matrix is shown in Figure 4.2. However, based on the structure of the dataset, a strong bivariate linear relationship for some variables in the scatterplot matrix was not recognized.

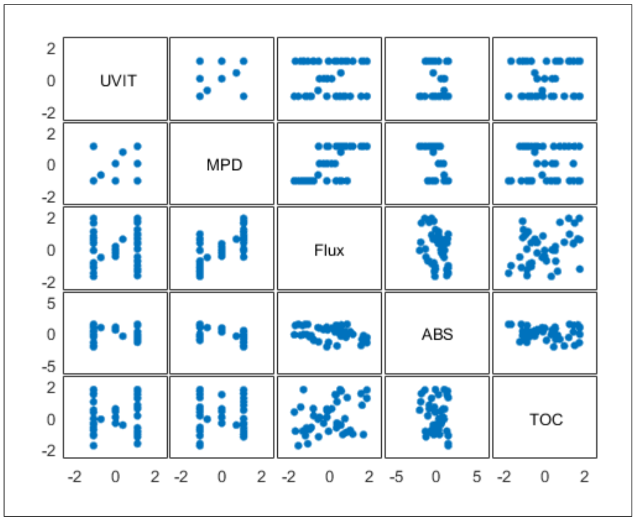


Figure 4.2 SMPUF System Scatterplot Matrix

4.3.3 Correlation Matrix

The bivariate correlation coefficient of -1 and +1 means that all data points fall on a straight line and have strong linearity. The correlation coefficient around zero means weak linearity. And the correlation coefficient of zero means there is no linear correlation between variables and the curvilinearity of the relationship should be evaluated. Meanwhile, If the variables are independent, Pearson's correlation coefficient is zero, but the converse is not true because the correlation coefficient only detects linear dependencies between variables. (Croxton, Cowden, & Klein, 1968)

The correlation matrix of the dataset is shown in Figure 4.3. From this correlation matrix plot, MPD is highly correlated with Flux and moderately correlated with ABS, but MPD is positively correlated with Flux and negatively correlated with ABS.

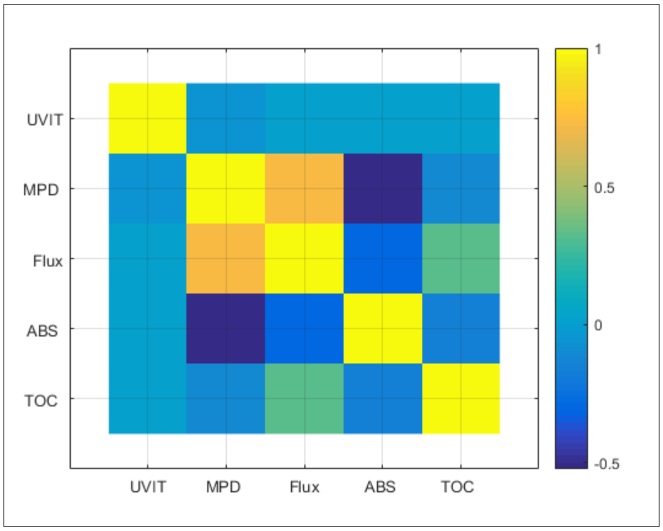


Figure 4.3 SMPUF System Correlation Matrix

4.3.4 First, Second, and Third Principal Components

Using the columns of the eigenvector matrix:

$$A = \begin{pmatrix} -0.04 & +0.28 & +0.96 & -0.01 & +0.05 \\ +0.61 & -0.36 & +0.10 & +0.16 & +0.68 \\ +0.59 & +0.14 & +0.01 & +0.53 & -0.59 \\ -0.48 & +0.04 & -0.04 & +0.83 & +0.27 \\ +0.19 & +0.88 & -0.26 & -0.06 & +0.34 \end{pmatrix} \tag{4.1}$$

The first PC is then given by:

$$Z1 = +0.61\, X2 + 0.59\, X3 - 0.48\, X4 + 0.19\, X5 \tag{4.2}$$

$$Z1 = +0.61\, MPD + 0.59\, Flux - 0.48\, ABS + 0.19\, TOC \tag{4.3}$$

MPD plays a major role in the first PC. This is evidenced by the high positive coefficient for MPD. The second PC is given by:

$$Z2 = +0.28\, X1 - 0.36\, X2 + 0.14\, X3 + 0.88\, X5 \tag{4.4}$$

$$Z2 = +0.28\, UVIT - 0.36\, MPD + 0.14\, Flux + 0.88\, TOC \tag{4.5}$$

TOC plays a major role in the second PC. This is evidenced by the high positive coefficient for TOC. And the third PC is then given by:

$$Z3 = +0.96\,X1 + 0.10\,X2 - 0.26\,X5 \tag{4.6}$$

$$Z3 = +0.96\,UVIT + 0.10\,MPD - 0.26\,TOC \tag{4.7}$$

UVIT is important in the third PC. This is evidenced by the high positive coefficient for UVIT.

4.3.5 Principal Components Coefficients

The scatter plot of PC2 coefficients Vs. PC1 coefficients is shown in Figure 4.4. As can be seen, MPD and Flux play a major role in the first PC. This is evidenced by the high positive coefficients of these variables. In addition, TOC is important in the second PC due to the high positive coefficient.

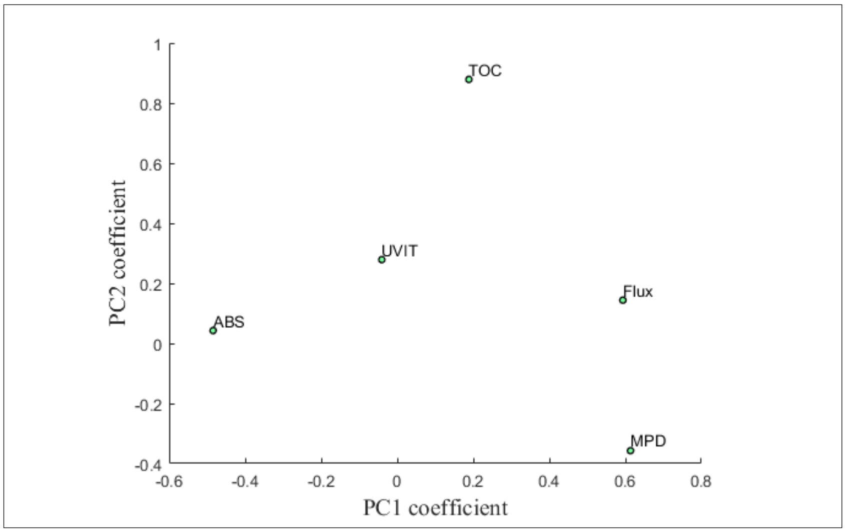


Figure 4.4 Scatterplot of PC$_2$ Coefficient Vs. PC$_1$ Coefficient

The scatter plot of PC3 coefficients Vs. PC2 coefficients is shown in Figure 4.5. As can be seen, TOC plays a major role in the second PC. This is evidenced by the high positive coefficients of this variable. In addition, UVIT is important in the third PC due to the high positive coefficient.

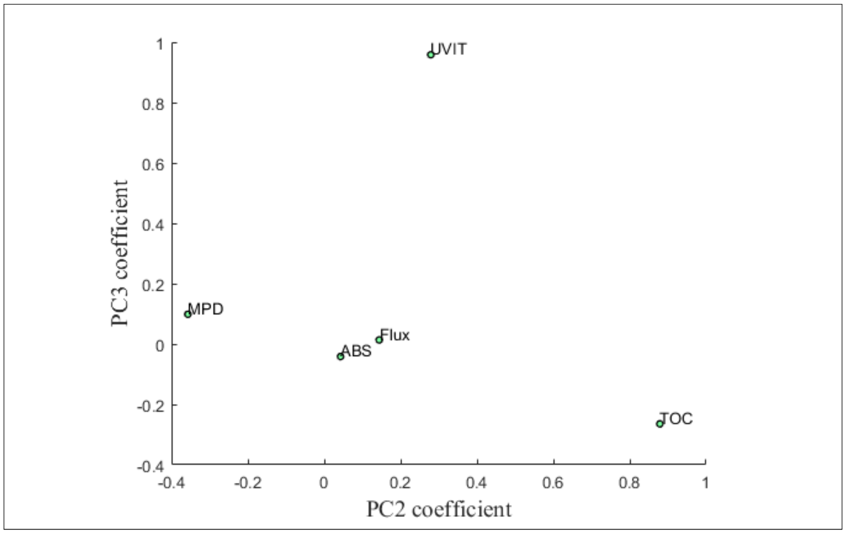


Figure 4.5 Scatterplot of PC_3 Coefficient Vs. PC_2 Coefficient

4.3.6 Biplot of Principal Components

The 2D biplot of PC2 Vs. PC1 is shown in Figure 4.6. The axes in the biplot represent the principal components, and the observed variables are represented as vectors. Each observation is represented as a red point in the biplot. From Figures 4.6 and 4.7, we can see that the first principal component has 3 positive coefficients for the variables X2 (MPD), X3 (Flux), and X5 (TOC), and 2 negative coefficients for the variables X1 (UVIT) and X4 (ABS). That corresponds to 3 and 2 vectors directed into the right and left halves of the plot, respectively. The second principal component, represented by the vertical axis, has 4 positive coefficients for the variables X1 (UVIT), X3 (Flux), X4 (ABS), and X5 (TOC), and 1 negative coefficients for the variables X2 (MPD). That corresponds to 4 and 1 vectors directed into the top and bottom halves of the plot, respectively. The 3D biplot of PC1, PC2, and PC3 is shown in Figure 4.7.

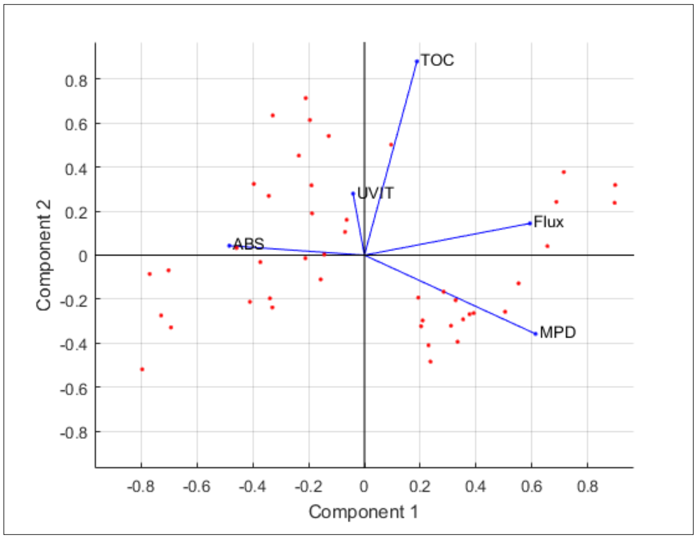


Figure 4.6 2D Biplot

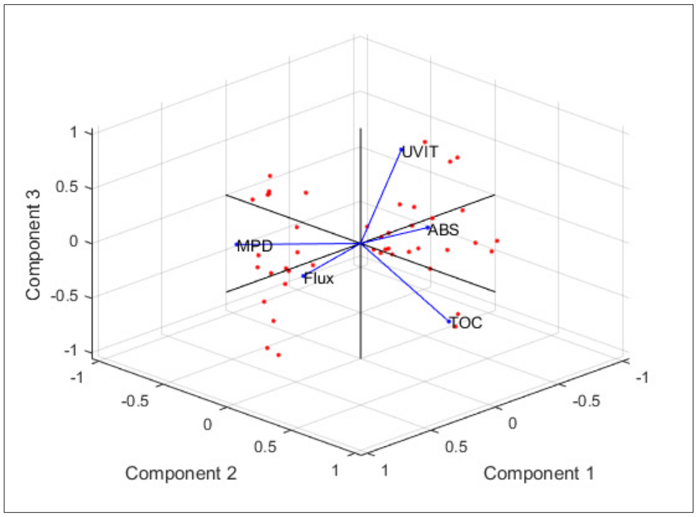


Figure 4.7 3D Biplot

4.3.7 Explained Variance and Lowest-Dimensional Space

The explained variance by the three PCs is $\ell 1 = 41.58\%$, $\ell 2 = 21.78\%$, and $\ell 3 = 19.85\%$. Notice that PC1, PC2, and PC3 combined account for 83.21% of the variance in the dataset. The Scree and Pareto plots of the explained variance Vs. the number of PCs are shown in Figures 4.8 and 4.9. Based on the explained variance by the first three PCs and also from the scree and Pareto plots, it can be deduced that the lowest-dimensional space to represent the SMPUF system dataset corresponds to $d = 3$.

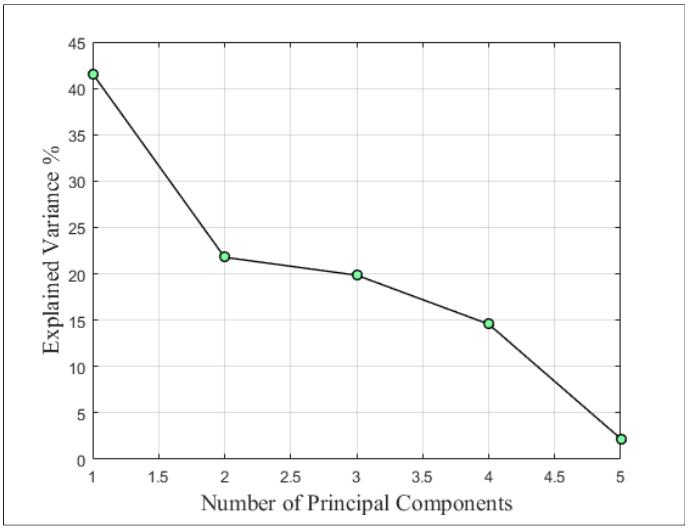


Figure 4.8 Scree Plot

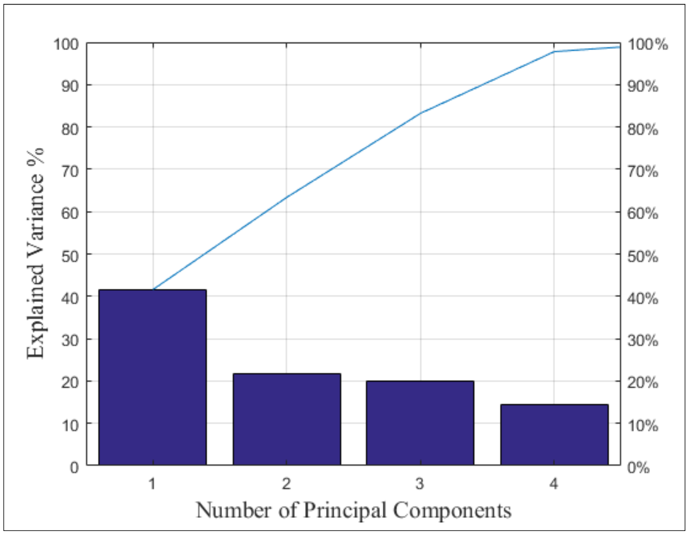


Figure 4.9 Pareto Plot

4.3.8 Principal Components Scores

The plot of the 2^{nd} PC score Vs. the 1^{st} PC score is shown in Figure 4.10. The labels displayed in the scatterplot represent the observation numbers. Note that none of the sample numbers appear to be outliers.

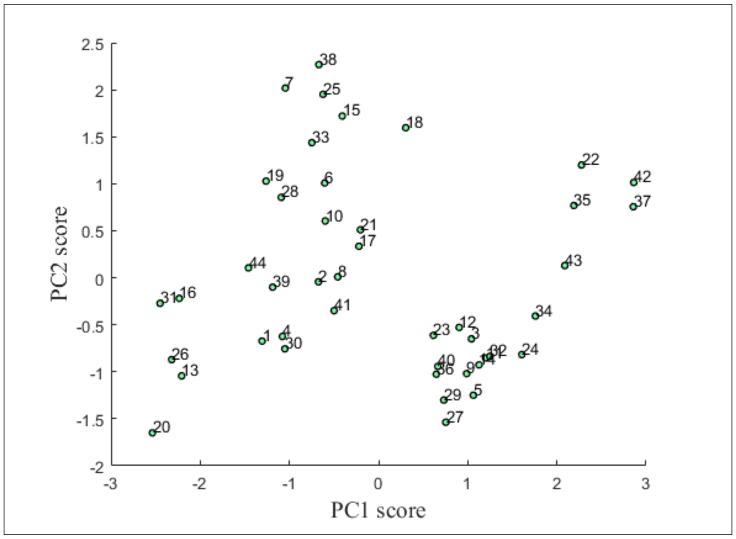


Figure 4.10 Scatterplot of PC_2 Score Vs. PC_1 Score

4.3.9 Hotelling and Control Chart

The Hotelling control chart and first PC control chart are shown in Figure 4.11 and Figure 4.12, respectively. The control charts indicate that all the plotted points on the control charts are within the control limits.

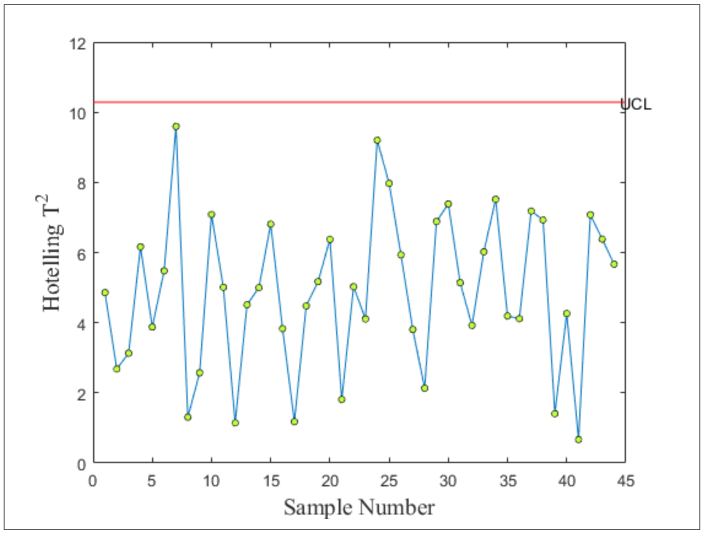


Figure 4.11 Hotelling Control Chart

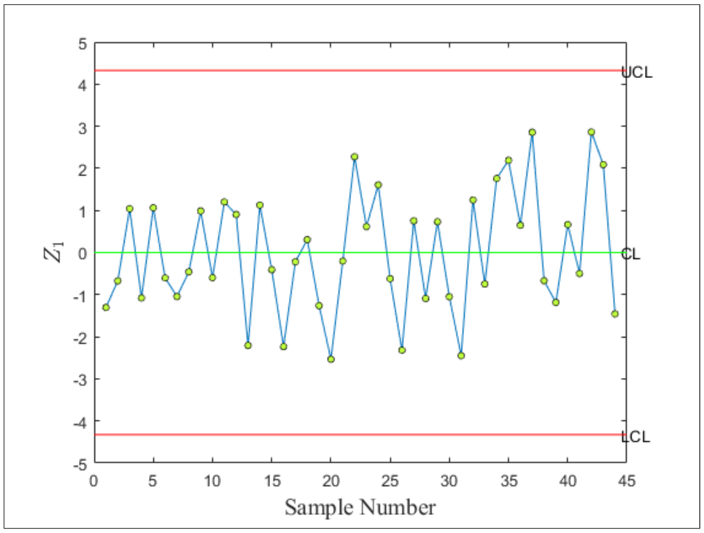


Figure 4.12 1st PC Control Chart

4.3.10 Conclusion

According to the PCA phase of the project, we can conclude that by implementing principal components analysis (PCA) on the SMPUF system, we can analyze either the principal components or just latent factors to facilitate visualization of all information contained in the SMPUF system observations, rather than analyzing the entire system's variables, which can be a time-consuming task. Meanwhile, it can be demonstrated that the principal components contain more than 80% of the system variability and that the latent factors are sufficient to predict the system’s future behavior; thus, it is best to focus on the principal components or the latent factors rather than analyzing all system variables.

4.4 Design of Experiments Overview

A systematic experimental strategy based on the design of experiments (DOE) was developed to determine the appropriate set of the submerged membrane photocatalytic ultrafiltration (SMPUF) system. This strategy emphasizes investigating the influences of the simultaneous SMPUF oily wastewater treatment process conditions on the performances of the membrane for the treatment of refinery-produced oily wastewater under various controllable conditions.

In this project, the DOE method included a full factorial design and an environmental noise-blocking technique. The procedure was examined on three facets: input factors to the process that can be classified as either controllable or uncontrollable variables, levels or settings of each factor, and response or output factor of the experiment (Montgomery, 2008).

The SMPUF system variables orthogonal arrays (OA) used in this study are shown in Table 4.2. Following data collection, all data are entered into the Design-Expert software version 9 (Stat-Ease Inc., USA) for analysis.

4.5 Design of Experiments Phase

Using Design-Expert software, empirical models based on the design of experiments (DOE) methodology were obtained during this phase. The experimental design provides statistical analysis to determine which SMPUF process inputs have a significant effect on membrane performances to optimize system performance.

To perform this experiment, a two-level full factorial design was considered. Hence, 4 center points were defined to examine the existence of curvature in the experiment's response surface. Based on the minimum and maximum range of input variables in the actual form, $2^5 + 4$ treatment combinations were collected (Table 4.2).

To evaluate and analyze the experimental design, version 9 of the Design-Expert® software (Stat-Ease Inc., USA) was used. The design summary (Table 4.3) shows a summary of the design, factors, and response information.

Table 4.2 The experimental values of the five independent variables together with the responses

Standard	Input Variable					Response		
	X_1: ABFR (L/min)	X_2: OC (ppm)	X_3: UVIT (min)	X_4: pH (pH)	X_5: MPD (n)	Y_1: Flux (L/m^{2}hr)	Y_2: ABS (%)	Y_3: TOC (%)
						Experimental	Experimental	Experimental
1	1	250	30	4	30	17.73	87.18	34.32
2	5	250	30	4	30	16.13	100	20.91
3	1	1000	30	4	30	1.01	86.83	61.35
4	5	1000	30	4	30	6.43	100	3.61
5	1	250	120	4	30	13.48	87.38	100
6	5	250	120	4	30	19.31	99.94	6.94
7	1	1000	120	4	30	9.72	86.2	69.97
8	5	1000	120	4	30	2.4	98.54	28.6
9	1	250	30	10	30	24.17	85.79	35.05
10	5	250	30	10	30	57.95	98.78	91.63
11	1	1000	30	10	30	21.91	84.14	31.59
12	5	1000	30	10	30	63.58	96.1	100
13	1	250	120	10	30	26.96	81.15	67.38
14	5	250	120	10	30	73.09	100	86.14
15	1	1000	120	10	30	27.41	79.8	54.94
16	5	1000	120	10	30	65.15	99.25	97.41
17	1	250	30	4	90	100	76.82	99.26
18	5	250	30	4	90	77.24	90.6	30.13
19	1	1000	30	4	90	92.11	73.4	93.08
20	5	1000	30	4	90	67.14	91.28	22.72
21	1	250	120	4	90	100	83.45	85.11
22	5	250	120	4	90	95.25	85.72	22.63
23	1	1000	120	4	90	93.33	81.12	73.16
24	5	1000	120	4	90	80.75	87.38	23.63
25	1	250	30	10	90	59.06	92.12	42.41
26	5	250	30	10	90	59.28	72.82	79.02
27	1	1000	30	10	90	63.18	87.97	40.05
28	5	1000	30	10	90	45.34	71.68	64.74
29	1	250	120	10	90	58.45	90.34	34.24
30	5	250	120	10	90	47.7	77.39	28.35
31	1	1000	120	10	90	71.56	83.63	26.25
32	5	1000	120	10	90	34.1	77.99	18.81
33	3	625	75	7	60	35.28	95.24	52.36
34	3	625	75	7	60	40.96	93.22	42.21
35	3	625	75	7	60	51.97	95.05	66.91
36	3	625	75	7	60	46.42	92.89	62.72

Table 4.3 Design Summary

Name	Units	Type	Low (-1)	High (+1)	Mean	Std. Dev.	Analysis	Model	Transform
A: ABFR	L/min	Factor (Numeric)	1	5	3	1.91	-	-	-
B: OC	ppm	Factor (Numeric)	250	1000	625	358.57	-	-	-
C: UVIT	min	Factor (Numeric)	30	120	75	43.03	-	-	-
D: pH	pH	Factor (Numeric)	4	10	7	2.87	-	-	-
E: MPD	n	Factor (Numeric)	30	90	60	28.69			
R_1: Flux	L/m^2 hr	Response (Numeric)	1.01	100	49.04	29.41	Factorial	2FI	None
R_2: ABS	%	Response (Numeric)	71.68	100	88.09	8.40	Polynomial	Cubic	None
R_3: TOC	%	Response (Numeric)	3.61	100	52.71	29.9	Factorial	2FI	None

4.5.1 First Response (Permeate Flux)

In this case, based on the ANOVA Table (Table 4.4), the model F-value of 54.78 implies that the model is significant. The main effects B and E and two-factor interaction (2FI) affects AD, AE and DE are significant model terms. The Table vividly shows that the model center points and consequently the model curvature is not significant and the ‘Lack of Fit’ F-value of 1.48 indicates that the lack of fit is not significant relative to the pure error. The ordinary R^2 with the value of 0.93 indicates that about 93 percent of the total variability is explained by the model. At last, the predictive model (Equation 4.8) for the first response based on the factor coefficients is presented in terms of coded factors as follows:

$$Flux = 49.04 + 0.96\ A - 3.15\ B + 0.22\ D + 21.82\ E + 4.88\ AD - 9.14\ AE - 16.91\ DE \tag{4.8}$$

Table 4.4 ANOVA table for permeate flux

Source	Sum of Squares	dof	Mean Square	F-Value	p-value (Prob. > F)	
Model	28164.65	7	4023.52	54.78	<0.0001	**significant**
A-ABFR	29.58	1	29.58	0.40	0.5310	not significant
B-OC	316.90	1	316.90	4.31	0.0474	significant
D-pH	1.48	1	1.48	0.02	0.8882	not significant
E-MPD	15228.88	1	15228.88	207.34	<0.0001	significant
AD	762.53	1	762.53	10.38	0.0033	significant
AE	2674.36	1	2674.36	36.41	<0.0001	significant
DE	9150.92	1	9150.92	124.59	<0.0001	significant
Curvature	130.48	1	130.48	1.78	0.1937	**not significant**
Residual	1983.15	27	73.45			
Lack of Fit	1828.88	24	76.20	1.48	0.4241	**not significant**
Pure Error	154.27	3	51.42			
Cor. Total	30278.28	35				

To check the normality of residuals, the normal probability plot of residuals (Figure 4.13) was selected. The plot clearly shows that the residuals pursue a straight line. Meanwhile, to find the optimum value of the permeate flux based on the contour trend, the maximum flux (100 L/m^2 hr) is achieved at the A$^-$, B$^-$, D$^-$, and E$^+$ corner of the contour plot (Figure 4.14).

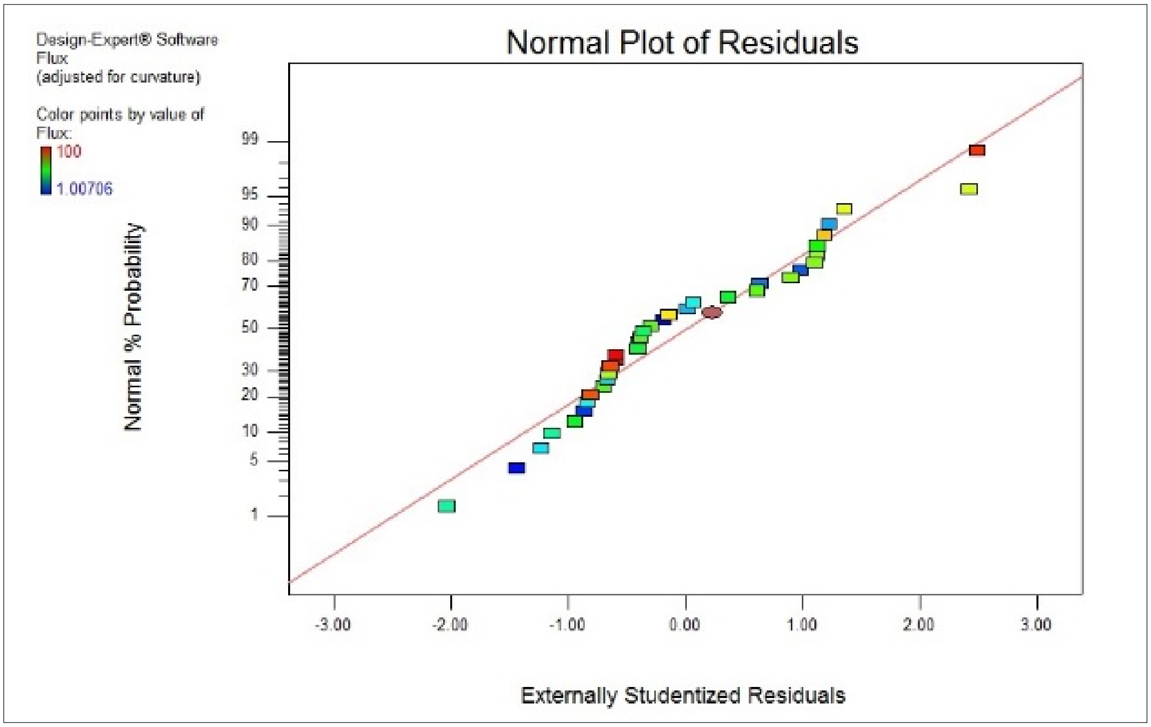


Figure 4.13 Normal probability plot of residuals for permeate flux

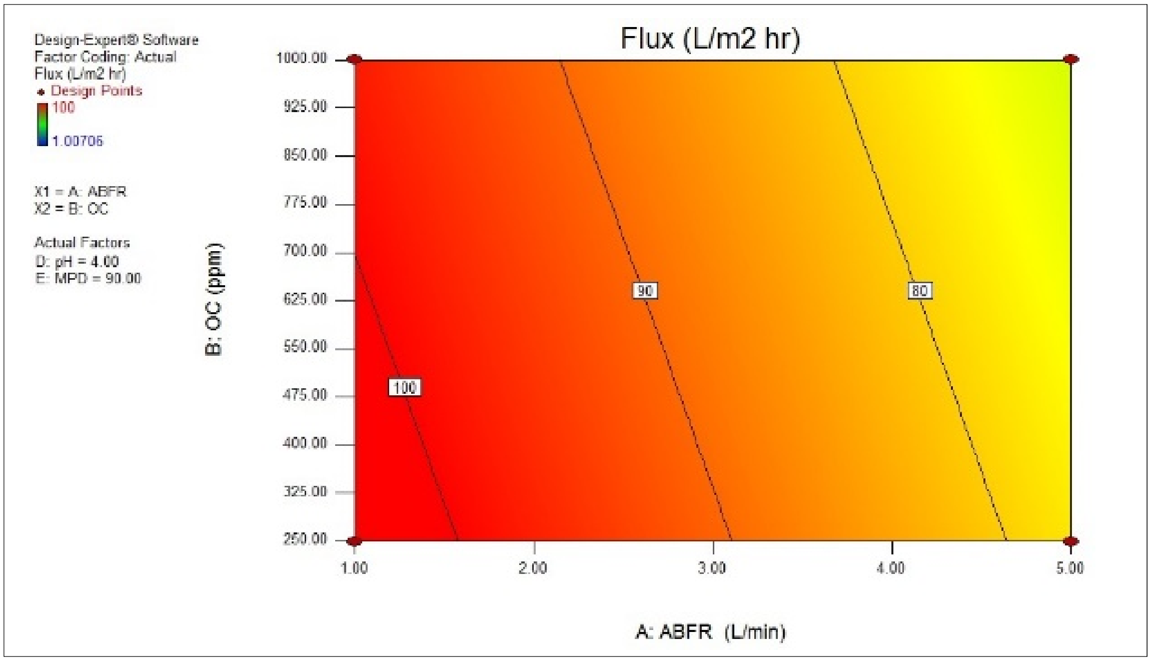


Figure 4.14 Contour plot for permeate flux

4.5.2 Second Response (Absorbance)

After finding the optimum result of the first response, the second response was considered. From the ANOVA table, it is evident that A, D, and E as the main effects, AD and AE as the 2FI effects, ACD and ADE as the 3FI effects, and A^2 offer a quadratic model for the second response. The ANOVA table is presented to confirm statistically the significant effects and summarized the test performed (Table 4.5).

Table 4.5 ANOVA table for absorbance

Source	Sum of Squares	dof	Mean Square	F-Value	p-value (Prob. > F)	
Model	2392.16	14	170.87	47.75	<0.0001	**significant**
A-ABFR	313.68	1	313.68	87.65	<0.0001	significant
C-UVIT	0.44	1	0.44	0.12	0.7288	not significant
D-pH	42.44	1	42.44	11.86	0.0024	significant
E-MPD	678.84	1	678.84	189.69	<0.0001	significant
AC	1.17	1	1.17	0.33	0.5739	not significant
AD	210.25	1	210.25	58.75	<0.0001	significant
AE	512.96	1	512.96	143.34	<0.0001	significant
CD	0.38	1	0.38	0.11	0.7489	not significant
CE	8.94	1	8.94	2.50	0.1290	not significant
DE	0.85	1	0.85	0.24	0.6312	not significant
A^2	162.56	1	162.56	45.42	<0.0001	significant
ACD	92.99	1	92.99	25.99	<0.0001	significant
ACE	10.60	1	10.60	2.96	0.0999	significant
ADE	356.06	1	356.06	99.49	<0.0001	significant
Residual	75.15	21	3.58			
Lack of Fit	70.72	18	3.93	2.66	0.2287	**not significant**
Pure Error	4.43	3	1.48			
Cor. Total	2467.32	35				

The Model F-value of 47.75 implies the model is significant. Meanwhile, the 'Lack of Fit' F-value of 2.66 indicates the lack of fit is not significant relative to the pure error and there is a 22.87% probability that the 'Lack of Fit' F-value could happen due to noise. In this specific case, the R^2 with the value of 0.9695 indicates that about 97 percent of the total variability is explained by the model. The predictive model (Equation 4.9) is presented in terms of coded factors as follows, for the second response:

$$ABS = +94.10 + 3.13\ A + 0.18\ C - 1.15\ D - 4.61\ E + 0.20\ AC - 2.56\ AD - 4.004\ AE - 0.11\ CD + 0.53\ CE + 0.16\ DE - 6.76\ A\hat{}2 + 1.70\ ACD - 0.58\ ACE - 3.34\ ADE \tag{4.9}$$

From the normal probability plot of the residuals (Figure 4.15), it can be understood that the residuals generally fall on a straight line indicating that the errors are distributed normally. The 3D surface plot was used to find the optimum value of the absorbance (Figure 4.16). Oil concentration (B) due to the insignificant effect does not have any effect on this plot and at the moderately high level of the ABFR (A^+) the maximum absorbance (ABS: 100%) is achieved at the lowest level of the UVIT (C^-), the lowest level of the pH (D^-) and the lowest level of the MPD (E^-).

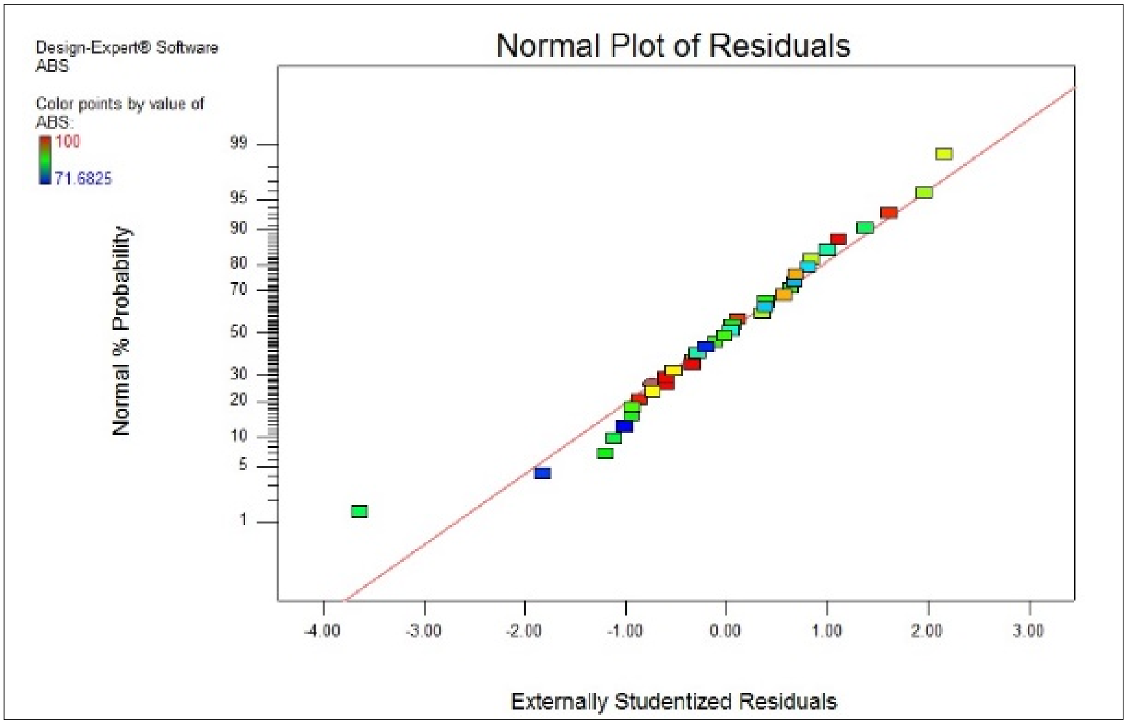


Figure 4.15 Normal probability plot of residuals for absorbance

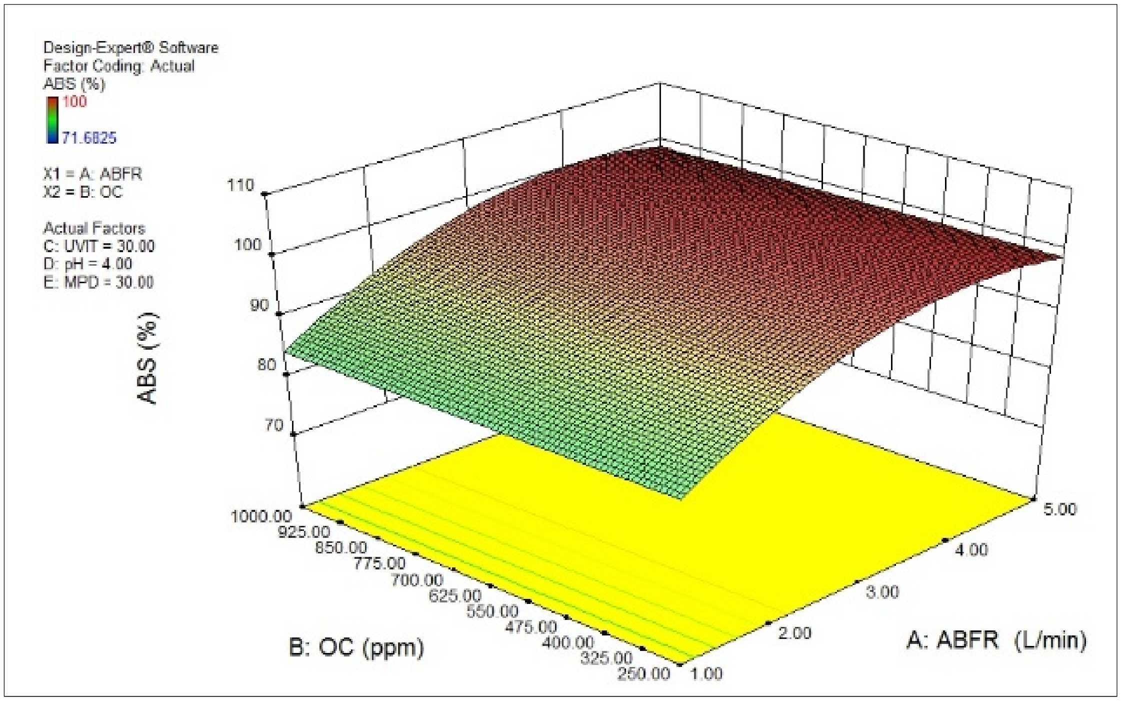


Figure 4.16 3D surface plot for absorbance

4.5.3 Third Response (Total Organic Carbon Degradation)

After finding the optimum result of the first and second responses, the third response was considered to finalize the analysis section of the responses. In this case, based on the ANOVA Table (Table 4.6), the model F-value of 19.64 implies that the model is significant. The main effects A, D, and E and two-factor interaction (2FI) effects AC, AD, AE, CD, CE, and DE are significant model terms. The Table vividly shows that the model center points and consequently the model curvature is not significant and the 'Lack of Fit' F-value of 1.11 indicates that the lack of fit is not significant relative to the pure error. The R^2 with the value of 0.8896 indicates that about 89 percent of the total variability is explained by the model. At last, the predictive model (Equation 4.10) for the third response based on the factor coefficients is presented in terms of coded factors as follows:

$$TOC = +52.71 - 6.97\ A - 0.82\ C + 3.83\ D - 3.32\ E - 5.44\ AC + 21.60\ AD - 5.76\ AE - 3.61\ CD - 9.13\ CE - 11.071\ DE \tag{4.10}$$

Table 4.6 ANOVA table for total organic carbon

Source	Sum of Squares	dof	Mean Square	F-Value	p-value (Prob. > F)	
Model	26345.71	10	2634.57	19.64	<0.0001	**significant**
A-ABFR	1552.57	1	1552.57	11.58	0.0023	significant
C-UVIT	21.62	1	21.62	0.16	0.6916	not significant
D-pH	469.60	1	469.60	3.50	0.0735	significant
E-MPD	352.81	1	352.81	2.63	0.1179	significant
AC	948.15	1	948.15	7.07	0.0137	significant
AD	14933.48	1	14933.48	111.35	<0.0001	significant
AE	1060.00	1	1060.00	7.90	0.0097	significant
CD	417.75	1	417.75	3.11	0.0903	significant
CE	2667.17	1	2667.17	19.89	0.0002	significant
DE	3922.57	1	3922.57	29.25	<0.0001	significant
Curvature	50.07	1	50.07	0.37	0.5469	**not significant**
Residual	3218.64	24	134.11			
Lack of Fit	2850.93	21	135.76	1.11	0.5436	**not significant**
Pure Error	367.70	3	122.57			
Cor. Total	29614.42	35				

To check the normality of residuals, the normal probability plot of residuals (Figure 4.17) was selected. The plot clearly shows that the residuals pursue a straight line. Meanwhile, to find the optimum value of the TOC based on the cube plot, the maximum TOC of the model with the value of 92.08% is gained at the A^+, C^-, D^+, and E^- corner of the cube plot (Figure 4.18).

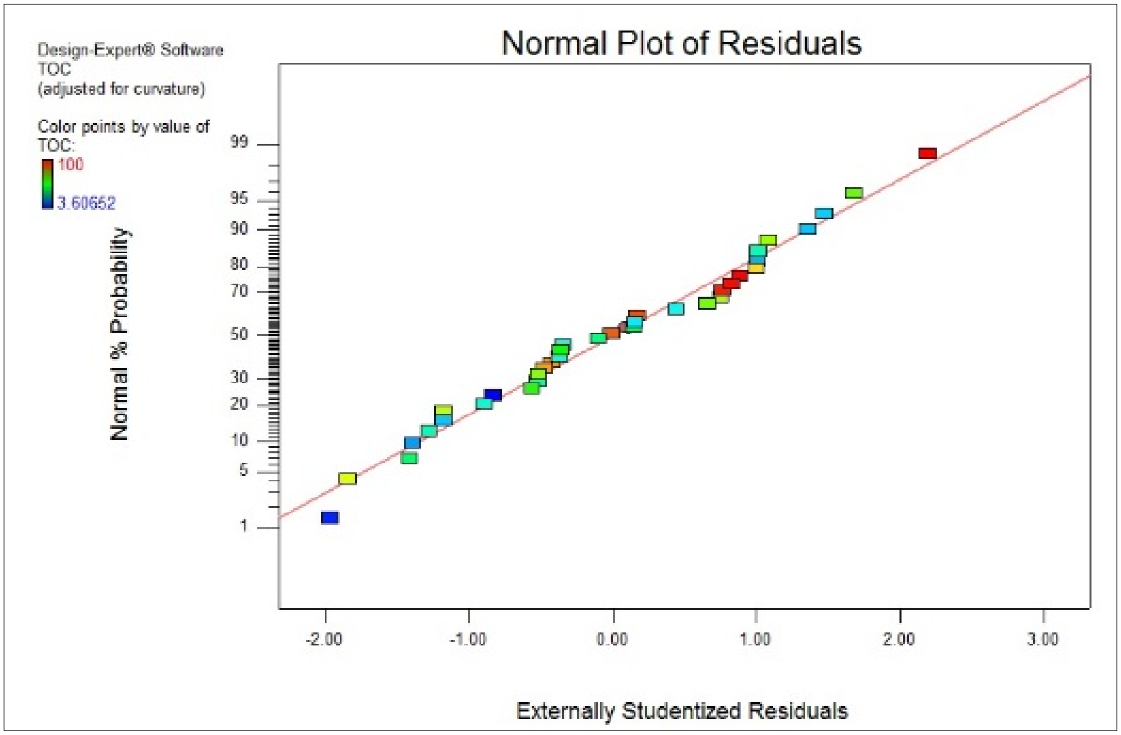


Figure 4.17 Normal probability plot of residuals for total organic carbon

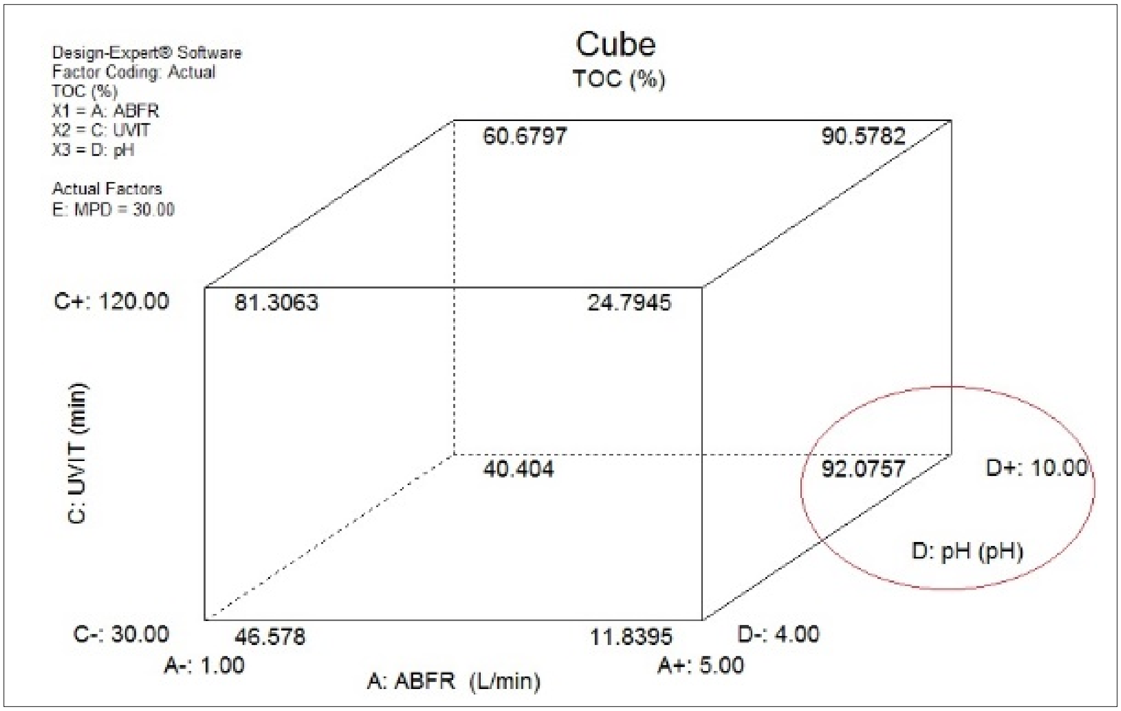


Figure 4.18 Cube plot for total organic carbon

4.5.4 Discussion

Ultimately, the optimum point within the range of the experimental model with the maximum desirability value of 82.5% is gained at the A^+ (5 L/min), B^- (250 ppm), C^+ (108 min), D^+ (10 pH), and E^- (30 n) with the approximate response values of permeate flux 62.5 L/m^2hr, ABS 100% and TOC 91% (Figure 4.19).

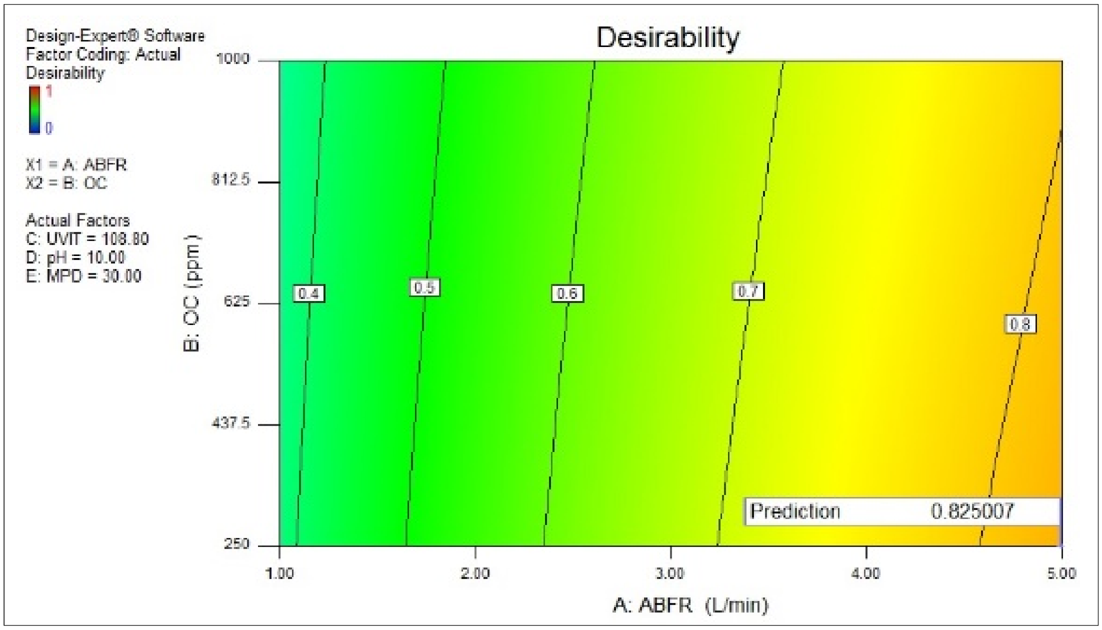


Figure 4.19 Desirability contour plot

4.5.5 Confirmation Test

To confirm the adequacy of the models obtained, eight confirmation runs were performed (Table 4.7). The first three runs are treatment combinations that were performed formerly, while the other runs are verification runs that have not been performed formerly, but they are within the range of the factors examined. The responses of these combinations are predicted using the 95% prediction interval and eight times the number of trials. The experimental and predicted response values are listed in Table 4.7. These values were compared to calculate and analyze absolute percentage errors. The range of the absolute percentage errors for permeate flux is ~ 1.77% to 10.45%, for ABS is ~ 0.16% to 9.98% and for TOC is ~ 0.68% to 13.31%.

It can be concluded from the confirmation tests that the experimental models developed are reasonably accurate. All the actual values for the confirmation runs are within the 95% prediction interval (PI) range.

Table 4.7 Confirmation experiments

Run	Input Factor					Response						Absolute Percentage Error		
						Experimental			Predicted					
	A: ABFR	B: OC	C: UVIT	D: pH	E: MPD	R_1: Flux	R_2: ABS	R_3: TOC	R_1: Flux	R_2: ABS	R_3: TOC	R_1	R_2	R_3
1	3	625	75	7	60	45.28	95.24	52.36	49.04	94.10	52.71	**8.31**	**1.20**	**0.68**
2	5	1000	120	4	30	12.40	98.54	28.60	12.18	98.38	24.79	**1.77**	**0.16**	**13.31**
3	1	250	30	10	90	59.06	92.12	39.80	60.61	89.30	41.39	**2.61**	**3.06**	**3.98**
4	5	625	30	7	90	69.99	90.64	50.00	62.68	81.60	52.07	**10.45**	**9.98**	**4.12**
5	3	1000	75	10	30	45.36	96.39	66.52	41.21	97.39	70.93	**9.16**	**1.04**	**6.64**
6	1	625	120	4	60	56.63	90.36	91.34	52.75	84.53	85.68	**6.85**	**6.45**	**6.19**
7	2	800	45	9	40	33.31	95.85	49.09	35.54	92.63	53.04	**6.68**	**3.36**	**8.07**
8	4	435	90	5	80	64.51	85.10	39.04	68.36	91.82	37.86	**5.97**	**7.90**	**3.04**

4.5.6 Conclusion

According to the DOE phase of the project, empirical models based on the design of experiment (DOE) methodology using Design-Expert software were obtained to develop and optimize SMPUF operating process conditions for the refinery-produced oily wastewater treatment. The design of experiments provided statistical analysis to determine which SMPUF process inputs have a significant effect on the membrane performances to optimize the performance of the system. Ultimately, the optimum point within the range of the experimental model with the maximum desirability value of 82.5% is gained at the A^+ (5 L/min), B^- (250 ppm), C^+ (108 min), D^+ (10 pH), and E^- (30 n) with the approximate response values of permeate flux 62.5 L/m^{2}hr, ABS 100% and TOC 91%. The presented work can be considered as a reference for future work since there is no known study considering three diverse factors as responses of experimental design methodology in the SMPUF system for the treatment of refinery-produced oily wastewater.

4.6 Artificial Intelligence Phase (AI/ML)

Based on the minimum and maximum range of input variables in the actual form, which is tabulated in Table 3.3 of the project, 153 non-negative experimental records were prepared. The experimental values of the five independent variables together with the three responses are tabulated in APPENDIX.

36 treatment combinations have already been collected to perform the DOE phase based on the minimum and maximum range of the input variables in the real form (Table 4.2 and Table 4.3). As a result, for the AI/ML phase, the total number of treatment combinations considered is $153 + 36 = \mathbf{189}$.

4.6.1 Data Visualization

A correlation matrix as a multivariate descriptive statistic was utilized. In other words, a correlation matrix represents the linear relationships between pairs of variables in each dataset. Each row and column represent a variable in the correlation matrix. Each value is Pearson's correlation coefficient between the variables defined by the corresponding row and column. In Figure 4.20, the correlation coefficient of the dataset was computed by Python and varies between -1 and +1.

The bivariate correlation coefficient of -1 and +1 means that all data points fall on a straight line and have strong linearity. The correlation coefficient around zero means weak linearity. And the correlation coefficient of zero means there is no linear correlation between variables and the curvilinearity of the relationship should be evaluated. Meanwhile, If the variables are independent, Pearson's correlation coefficient is zero, but the converse is not true because the correlation coefficient only detects linear dependencies between variables (Croxton, Cowden, & Klein, 1968).

$$\begin{aligned} X,Y \text{ independent} \quad &\Rightarrow \quad \rho_{X,Y}=0 \quad (X,Y \text{ uncorrelated}) \\ \rho_{X,Y}=0 \quad (X,Y \text{ uncorrelated}) \quad &\not\Rightarrow \quad X,Y \text{ independent} \end{aligned} \tag{4.11}$$

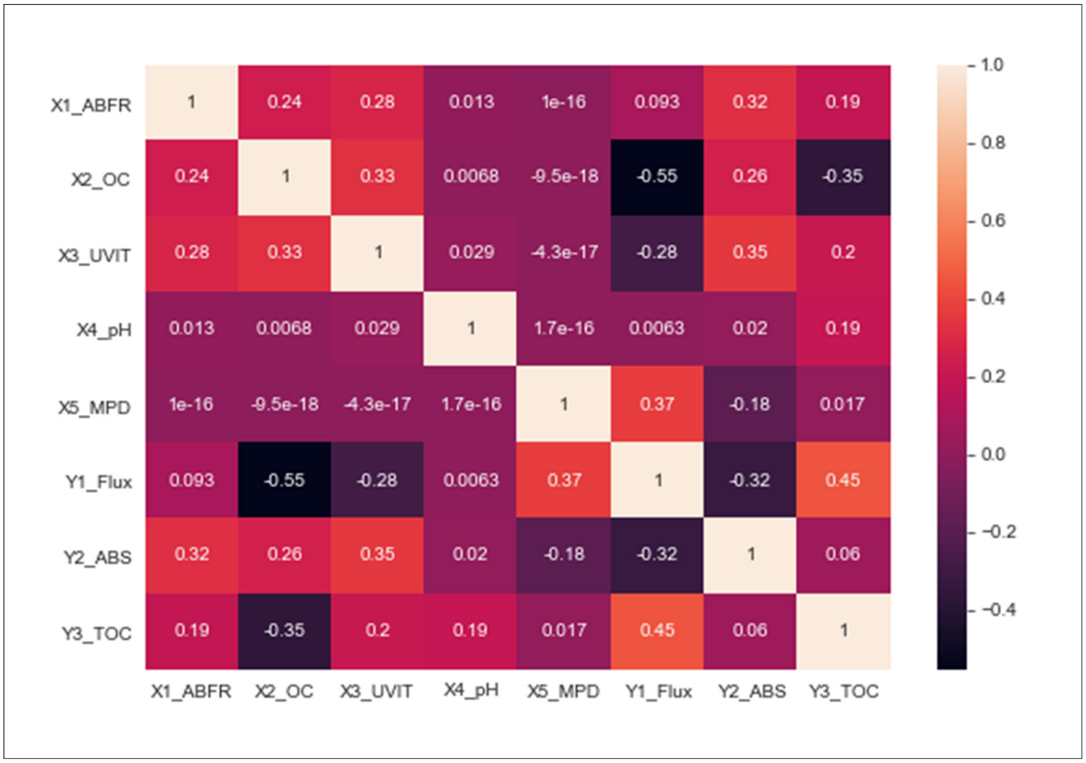


Figure 4.20 Correlation matrix of the dataset

Furthermore, the correlation matrix of the mean-centered dataset was computed by Python for better recognition of the correlations (Figure 4.21). In the regression model, mean centering subtracts the mean of the variable from all observations to reduce multicollinearity in the dataset. Multicollinearity consists of high intercorrelations between two or more input factors in the data set. This happens when two or more input factors are expected to affect the value of an output factor independently but are observed to be highly correlated. One of the drawbacks of multicollinearity is misleading results in determining how well each input factor is capable to predict an output factor.

The correlation matrix in Figure 4.21 uses mean-centered data. If the correlation coefficient is positive, the two variables in the correlation matrix increase or decrease together, i.e., they are correlated. Otherwise, if the correlation coefficient is negative, one increases when the other decreases, i.e., the variables are inversely correlated. (Iacobucci, Schneider, Popovich et al., 2016)

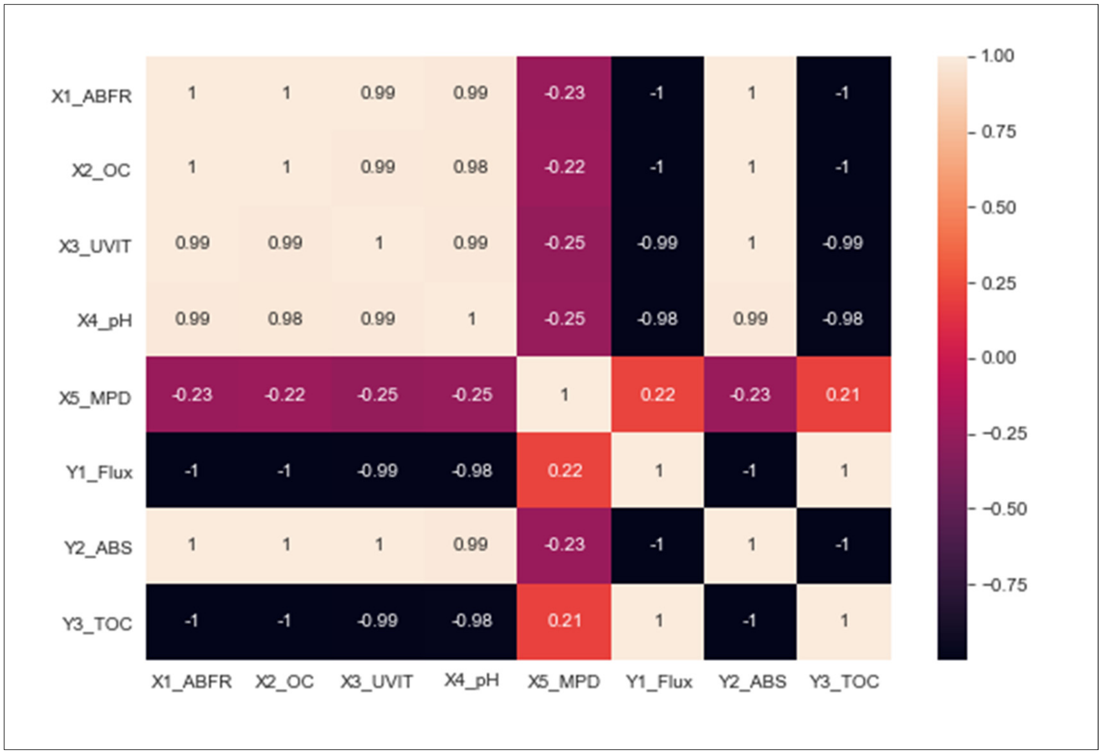


Figure 4.21 Correlation matrix using mean-centered data

The correlation matrix clearly shows the linear relationships between the factors in the positive and negative directions. However, the correlation matrix may not show desirable relationship results for nonlinear relationships. To overcome this problem, non-linear rank-based correlation coefficients such as Spearman's rank correlation coefficient and Kendall rank correlation coefficient, as well as Predictive Power Score (PPS) analysis, can be used to show non-linear relationships between variables.

4.6.2 Data Preprocessing

4.6.2.1 Data Normalization

Machine learning algorithms attempt to recognize patterns in the dataset by comparing features. However, when features are at drastically different scales, it is difficult to recognize patterns in the dataset. To overcome this problem, the normalization technique is utilized to make each data point have the same scale so that the features have the same importance.

In this work, maximum absolute scaling (MaxAbsScaler) is used to scale the data to its maximum value. Then, MaxAbsScaler divides every observation by the maximum value of the variable. MaxAbsScaler is similar to MinMaxScaler, except that values are mapped in the range between 0 and 1. MaxAbsScaler is given by:

$$xscaled = \frac{x}{max(x)} \tag{4.12}$$

Considering five input variables and three output variables, Figure 4.22 shows the correlation matrix between input variables after MaxAbsScaler normalization.

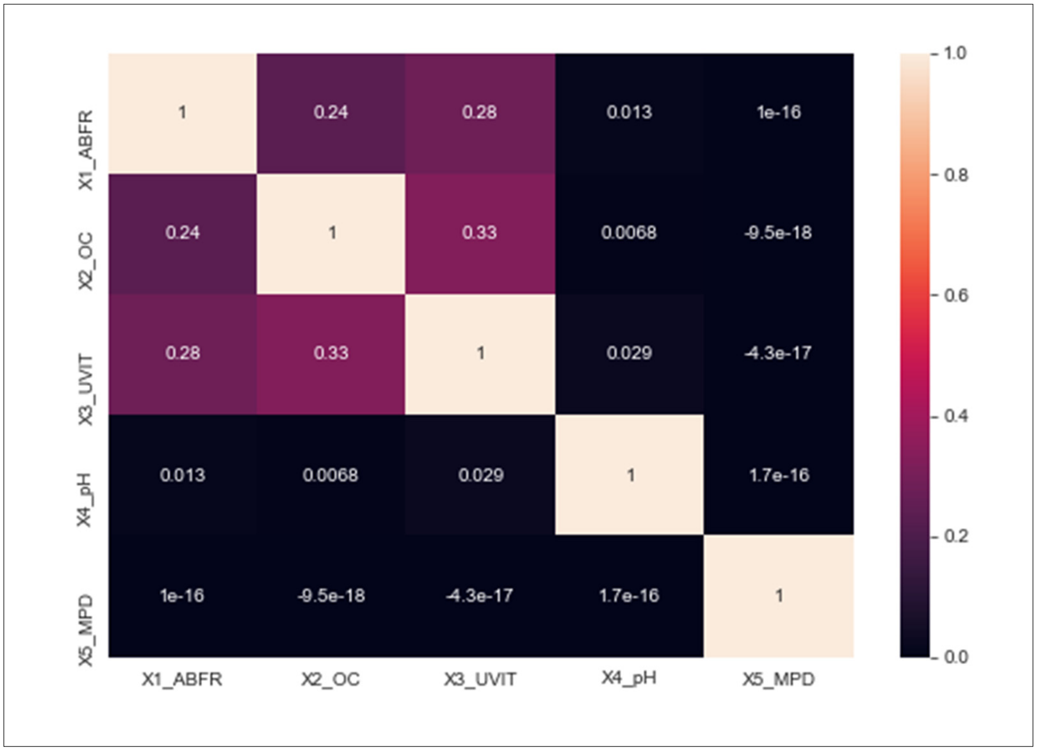


Figure 4.22 Correlation matrix of input variables after normalization

4.6.2.2 Training and Testing Dataset

The train-test split approach helps evaluation of the performance of a machine learning algorithm. This approach can be used for supervised and unsupervised machine learning. The approach requires dividing the dataset into **train** and **test** datasets. The training dataset is used to fit the model. The test dataset is not used to train the model. The model uses the test dataset for prediction. Then the predicted values are compared to the expected values. In other words, the training dataset is used to fit the machine learning model, and the test dataset is used to evaluate the performance of the machine learning model. The goal is to evaluate the performance of the machine learning model on a dataset that was not used to train the model (Jason Brownlee in 2020 in Python Machine Learning). In this work, 80% of the dataset was used for training and 20% of the dataset was used for testing. The total number of treatment combinations was $153 + 36 = \mathbf{189}$.

4.6.3 Multi-Output Regression in Machine Learning

In this work, the prediction model has been classified as a multi-output regression model based on the dataset of five inputs and three output variables. Three numeric output factors were predicted by giving five input variables. The multi-output regression model was utilized in **supervised learning**, considering inputs and corresponding outputs in the dataset. The supervised learning technique uses labeled data which consists of a set of inputs and corresponding outputs. In the multi-output regression model, the outputs are typically dependent upon the inputs and each other. This means that the outputs are not independent and may require a model that predicts output variables together. More details on the implementation of the multi-output regression model are provided by Hunt (2019) and Jason Brownlee in 2021 in Ensemble Learning.

4.6.3.1 Multi-Output Regression Algorithms under Comparison

In this work, several multi-output regression models as described below were compared, such as **K-Neighbors Regression**, **Decision Tree Regressor**, **Random Forest Regression**, **Multi-Output Regressor**, and **Neural Network**. Before starting to explain each model, the terms Machine learning (ML), Deep Learning (DL), and Artificial Intelligence (AI) are discussed, which are used interchangeably, however, they are different. AI is defined as learning and performing suitable techniques to solve problems and achieve desirable goals appropriate to the context. Instead of following explicit rules, ML uses models and algorithms to learn from data. DL is a form of ML that uses large and multi-layered artificial neural networks. Neural networks are computational algorithms inspired by brain or biological networks for information processing (Olczak et al., 2021). As shown in Figure 4.23, Spiking Neural Network (SNN) and Deep Learning (DL) are also brain-inspired models. On the other hand, K-Neighbors Regression, Decision Tree Regressor, Random Forest Regression, and Multi-Output Regressor are ML regression algorithms.

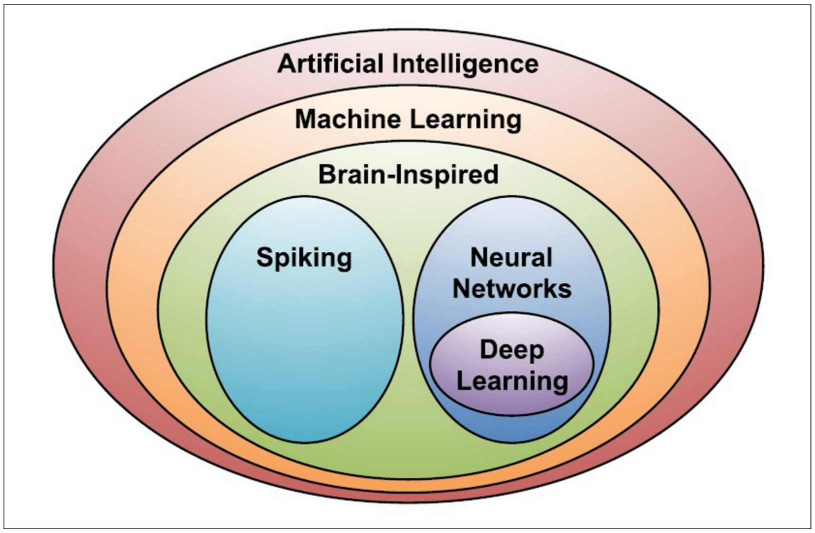


Figure 4.23 AI, ML, NN, DL, and SNN
taken from Sze, Chen, Yang, & Emer (2017, p. 2296)

4.6.3.2 K-Neighbors Regression

K-nearest neighbor uses all available elements and predicts the numerical target based on similarity measures known as distance functions (Aishwarya Singh in 2020 in A Practical Introduction to K-Nearest Neighbors Algorithm for Regression). The distance can be calculated using one of the distance metrics below:

The Euclidean distance uses the square root of the sum of the differences between the x and y data points:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(x_1 - y_2)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \qquad (4.13)$$

The Manhattan distance uses the sum of the absolute values of the differences between the x and y data points:

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \qquad (4.14)$$

The Minkowski distance can be interpreted as the generalization of Euclidean and Manhattan distances. The Minkowski distance introduced a parameter p which takes 1 or 2. In other words, when $p = 1$, the Minkowski distance becomes a Manhattan distance. Otherwise, for $p = 2$, the Minkowski distance becomes Euclidean distance:

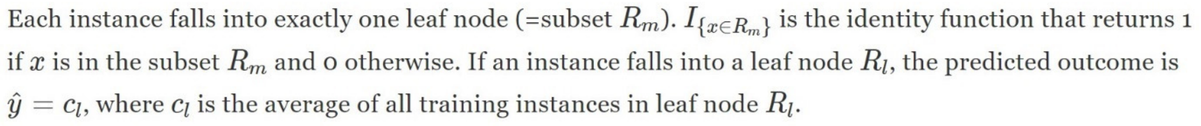
$$minkowski(x, y) = \left(\sum_{i=1}^{n} \left(x_i - y_i\right)^p\right)^{1/p} \qquad (4.15)$$

(Aishwarya Singh in 2020 in A Practical Introduction to K-Nearest Neighbors Algorithm for Regression)

4.6.3.3 Decision Tree Regressor

Decision trees apply binary rules to compute a target value in regression. Each tree is a relatively simple model. The model has branches, nodes, and leaves. Regression models such as linear regression and logistic do not perform well if the relationship between the features and the outcome is nonlinear, and the features interact with each other. The decision tree partitions the data points multiple times to address those challenges according to specific cut-off values in the features. Different subsets of the dataset are created through splitting, where each instance belongs to one subset. After splitting, terminal or leaf nodes represent the final subsets, and internal nodes or split nodes represent the intermediate subsets. The average outcome of the training data in a node is used to predict the outcome in each leaf node (Christoph Molnar (2022), Interpretable Machine Learning, A Guide for Making Black Box Models Explainable). The following formula describes the relationship between the outcome y and features x:

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\} \qquad (4.16)$$



Each instance falls into exactly one leaf node (=subset R_m). $I_{\{x \in R_m\}}$ is the identity function that returns 1 if x is in the subset R_m and 0 otherwise. If an instance falls into a leaf node R_l, the predicted outcome is $\hat{y} = c_l$, where c_l is the average of all training instances in leaf node R_l.

4.6.3.4 Random Forest Regression

A Random Forest technique can be applied to perform both regression and classification tasks. It uses multiple decision trees and Bootstrap Aggregation, i.e., bagging. In the Random Forest method, bagging trains each decision tree on a different data sample, where data sampling is performed with replacement (Krishni in 2018 in A Beginners Guide to Random Forest Regression).

4.6.3.5 Multi-Output Regressor

A multi-output regressor is a regression approach that requires predicting two or more numeric values or outputs for given input examples. In other words, the multi-output regressor model attempts to predict dependent variables based on the value of two or more independent variables (Watt, Borhani, & Katsaggelos, 2020). In our case, we have five independent variables and three dependent variables.

4.6.3.6 Neural Network and Deep Learning

Neural networks can be defined as algorithms inspired by the human brain to recognize patterns. Neural network algorithms interpret sensory data through machine perception, labeling, or raw clustering input data. Deep learning is one of the neural network approaches that use multiple layers, consisting of nodes. A node is just a place where computation happens, loosely patterned on a neuron in the human brain, which fires when it encounters enough stimuli. As shown in Figure 4.24, a node combines the dataset input with coefficients or weights that amplify or attenuate the input. Assigning significance to inputs depends on the task the algorithm is trying to learn. The task can be regression or classification. The input-weight products are summed, then the sum is passed through the activation function to determine if and how far that signal needs to progress further through the network to produce the result (Chris Nicholson in 2020 in A Beginner's Guide to Neural Networks and Deep Learning).

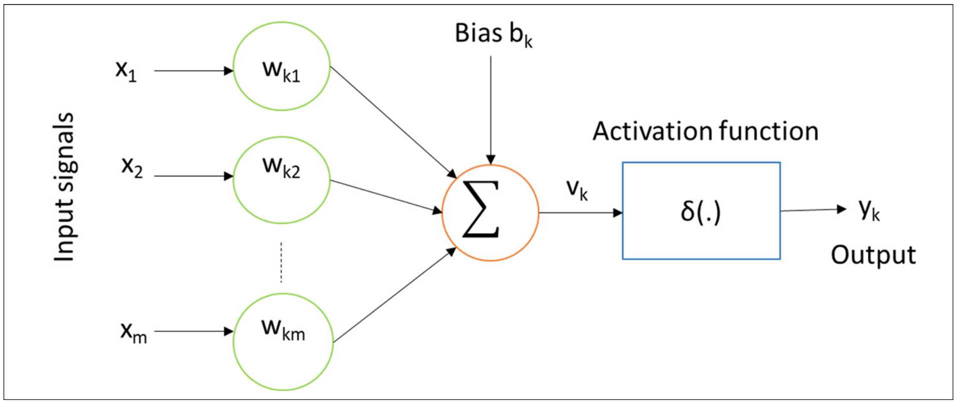


Figure 4.24 A basic model of neuron networks
taken from Alom et al. (2018, p. 6)

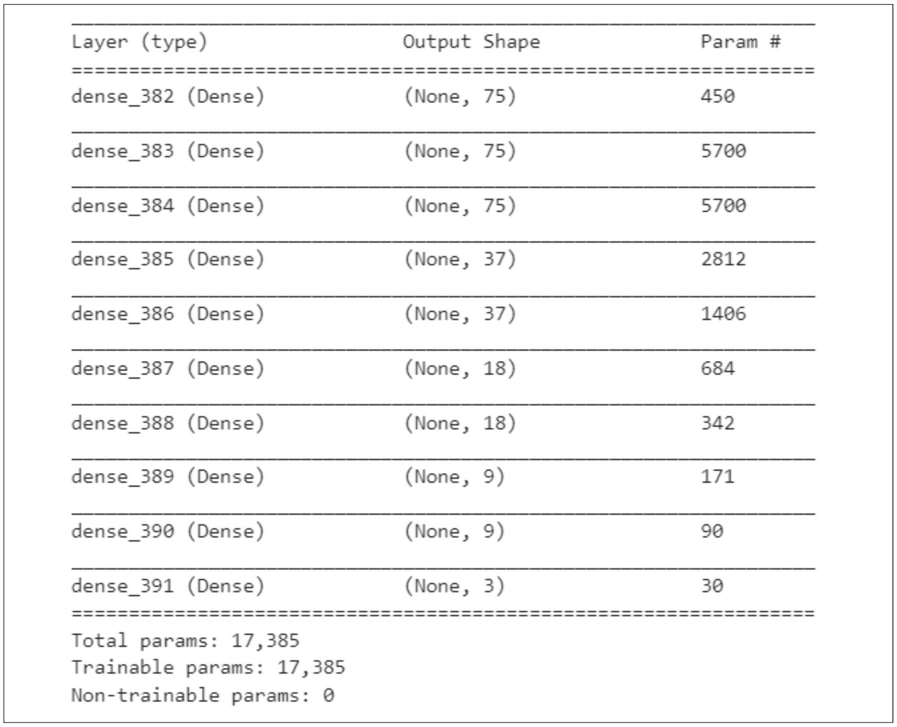
In prediction, Multi-Layer Perceptron (MLP) is utilized. In artificial neural networks (ANN), multi-layer neural networks can be called Multi-Layer Perceptron (MLP), the most useful type of neural network. A perceptron is a single-neuron model that was a precursor to larger neural networks (Jason Brownlee in 2022 in Crash Course on Multi-Layer Perceptron Neural Networks). In this work, to highlight the uses of multiple layers, the model was called Multi-Layer Perceptron (MLP).

4.6.4 Develop Multi-Output Model

This section utilized Multi-Layer Perceptron (MLP) and five input variables to predict three output variables. The prediction was classified into two categories: Multi-Output Model without PCA and Multi-Output Model with PCA. Principal Component Analysis (PCA) is mainly used in unsupervised machine learning algorithms for information compression and dimensionality reduction (Harsha Goonewardana in 2019 in PCA: Application in Machine Learning).

4.6.4.1 Multi-Output Model without PCA

The output of the Multi-Output Models without Principal Component Analysis (PCA) is described in this section. In Figure 4.25, the Multi-Layer Perceptron (MLP) used an input layer of 75 neurons, 8 hidden layers, and an output layer of 3 neurons. The three neurons of the output layer corresponding to 3 output variables.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_382 (Dense)            (None, 75)                450
_________________________________________________________________
dense_383 (Dense)            (None, 75)                5700
_________________________________________________________________
dense_384 (Dense)            (None, 75)                5700
_________________________________________________________________
dense_385 (Dense)            (None, 37)                2812
_________________________________________________________________
dense_386 (Dense)            (None, 37)                1406
_________________________________________________________________
dense_387 (Dense)            (None, 18)                684
_________________________________________________________________
dense_388 (Dense)            (None, 18)                342
_________________________________________________________________
dense_389 (Dense)            (None, 9)                 171
_________________________________________________________________
dense_390 (Dense)            (None, 9)                 90
_________________________________________________________________
dense_391 (Dense)            (None, 3)                 30
=================================================================
Total params: 17,385
Trainable params: 17,385
Non-trainable params: 0
```

Figure 4.25 MLP before applying PCA

Figures 4.26 and 4.27 show the minimization of the error function in MLP. The Mean Squared Error (MSE) is utilized as the error function.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{4.17}$$

n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

As metrics, the Mean Squared Error (MSE) was compared to the Mean Absolute Error (MAE). The results in Figures 4.26 and 4.27 show that the MLP model minimized the mean squared error better than the mean absolute error.

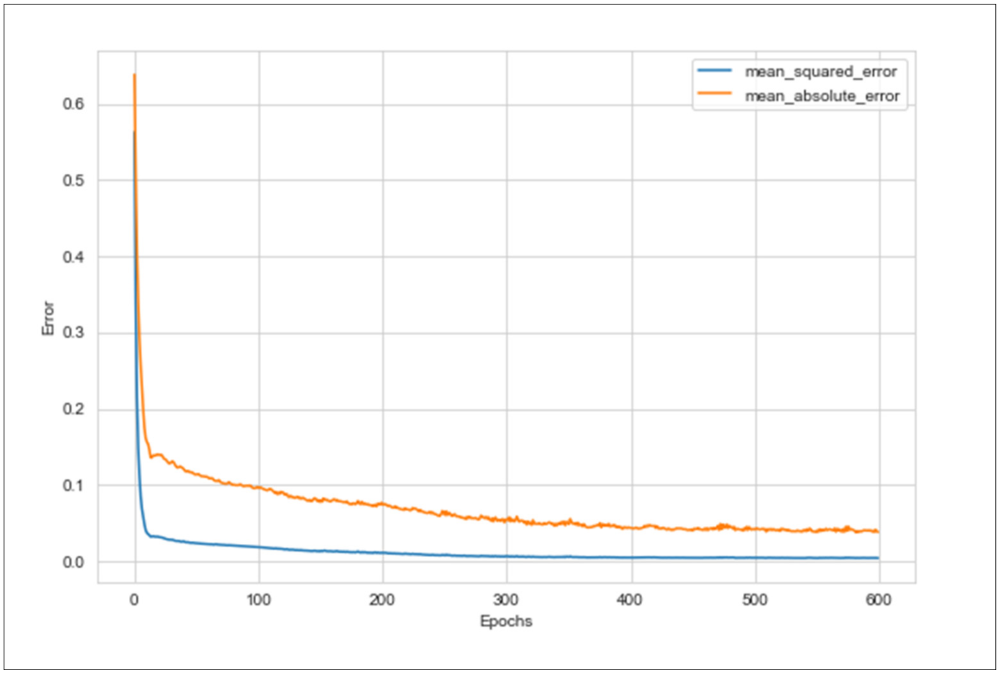


Figure 4.26 MLP error function minimization

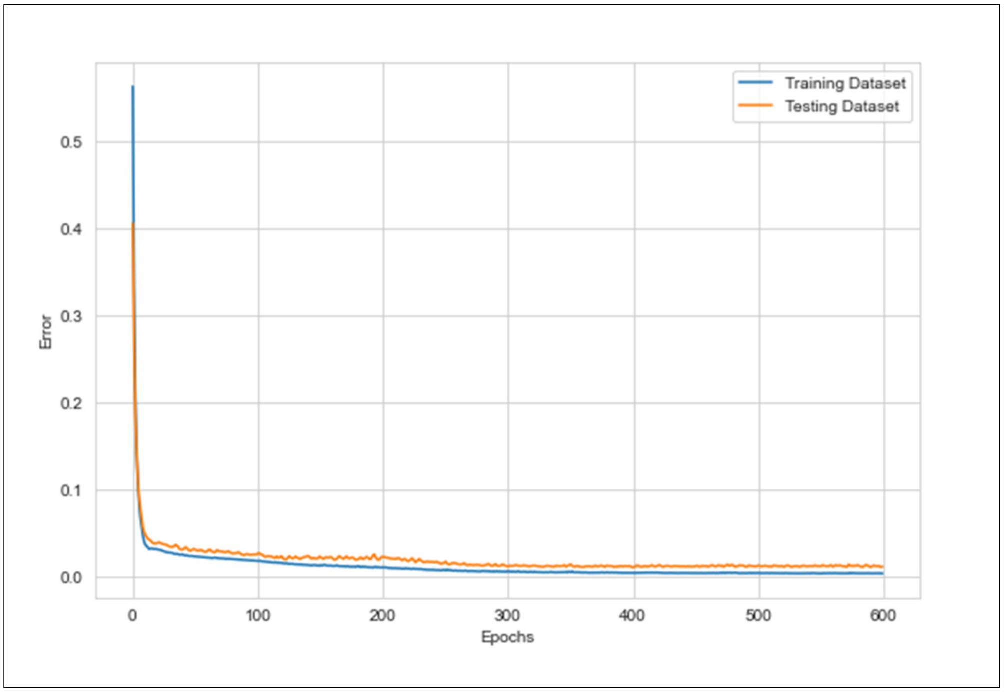


Figure 4.27 MLP Error function minimization based on dataset splitting

Figures 4.28 to 4.33 show the comparison between the actual data, i.e., the ground truth, and the predicted data using the multiple-output models compared in this study.

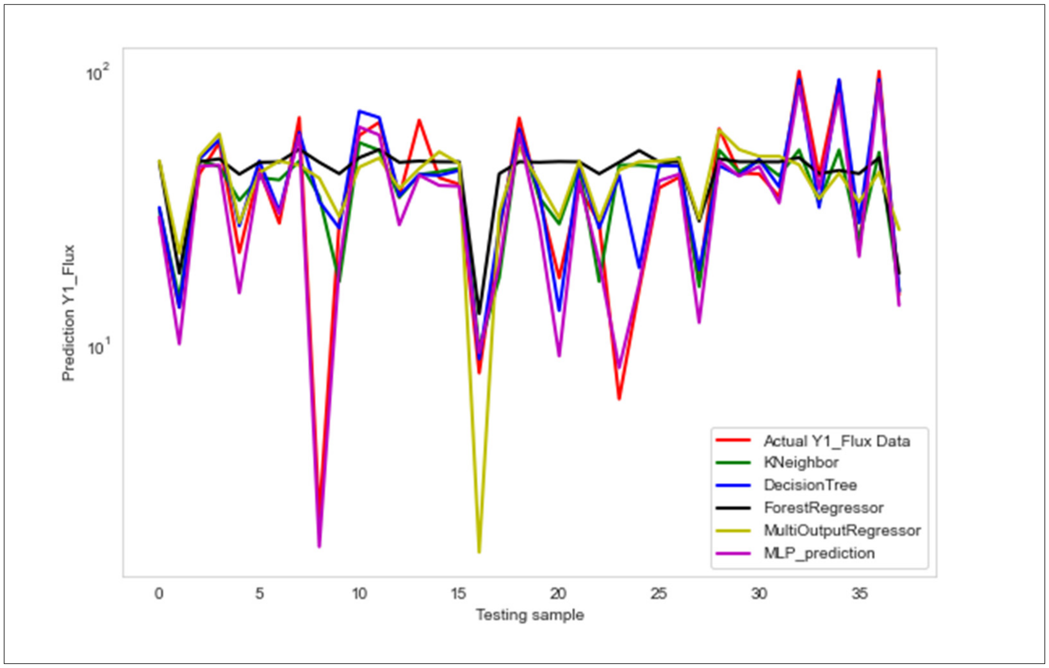


Figure 4.28 Y1: Permeate flux actual data and predicted data

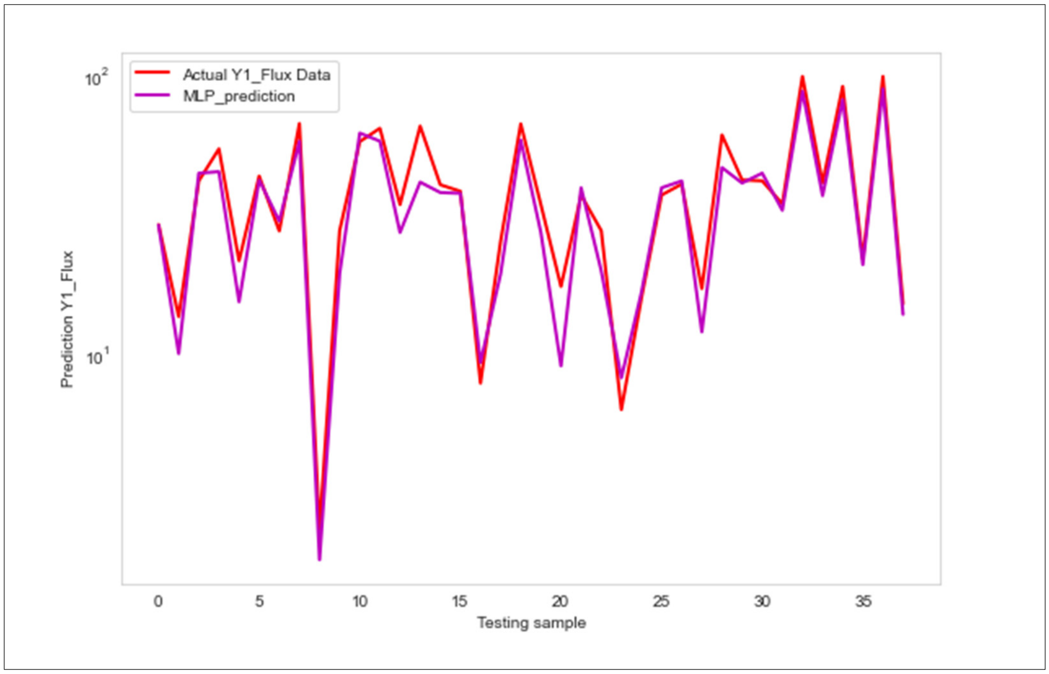


Figure 4.29 Y1: Permeate flux actual data and predicted data using the best-fit model

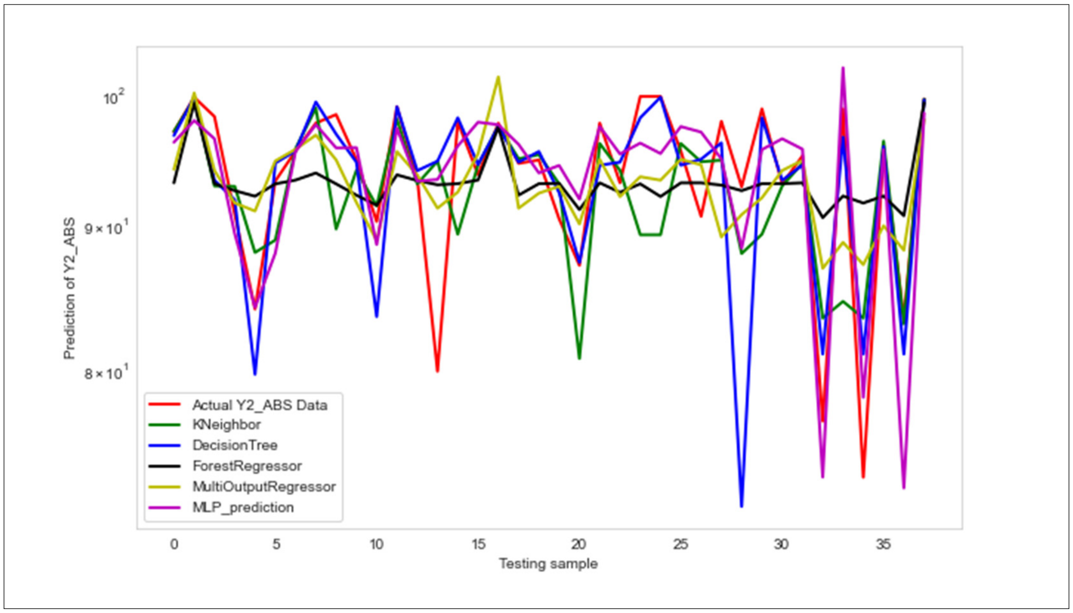


Figure 4.30 Y_2: ABS actual data and predicted data

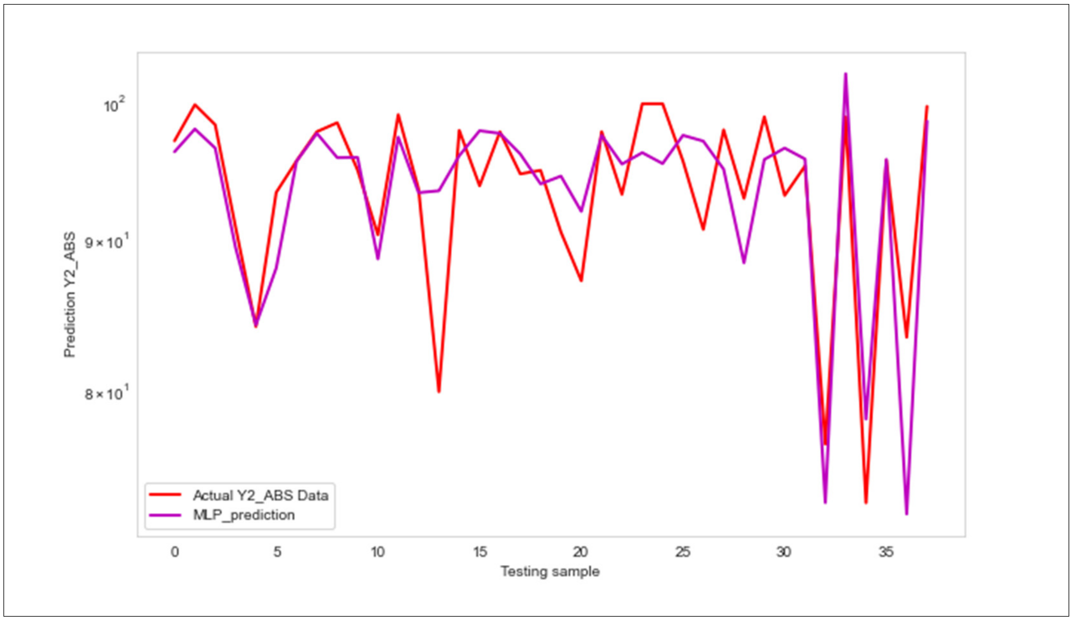


Figure 4.31 Y_2: ABS actual data and predicted data using the best-fit model

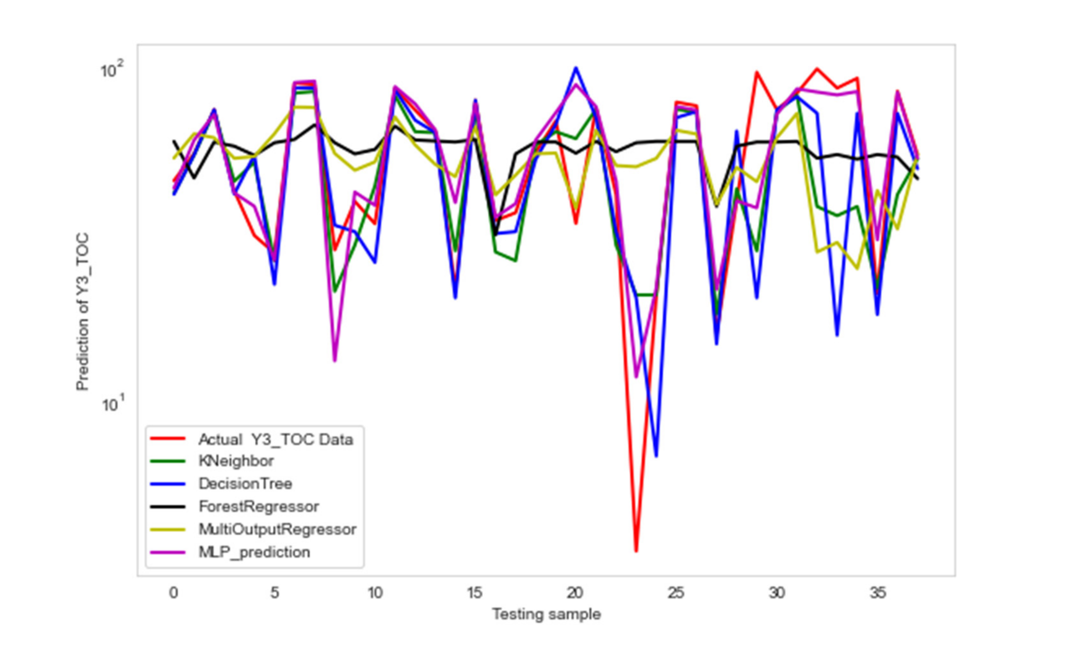


Figure 4.32 Y_3: TOC Degradation actual data and predicted data

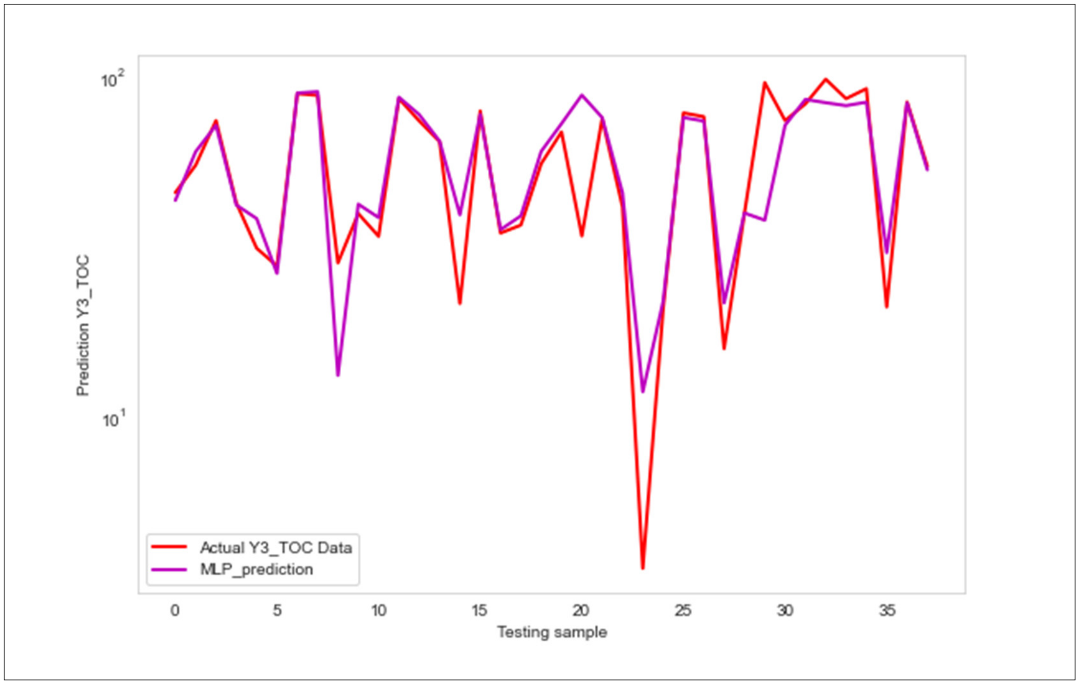


Figure 4.33 Y_3: TOC Degradation actual data and predicted data using the best-fit model

Figures 4.34 to 4.37 show the comparison of the Mean Squared Errors (MSE) of the prediction using different prediction models in comparison. The results show that the MLP performed well based on the output variables. As shown in Figure 4.37, MLP provides better performance than other models.

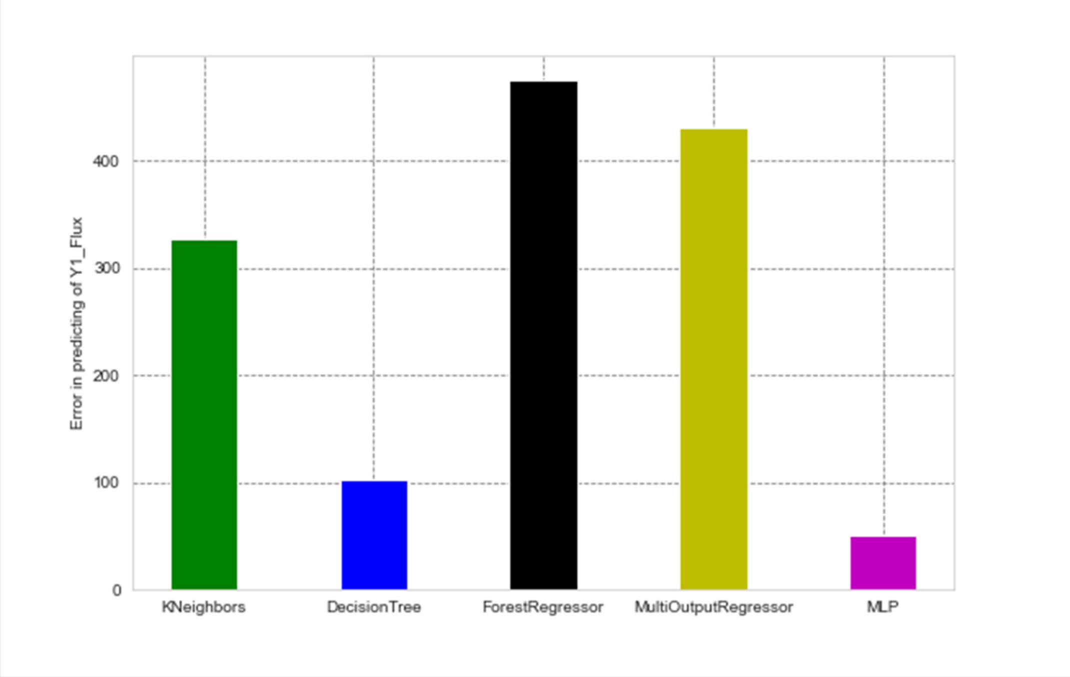


Figure 4.34 Y1: Permeate flux prediction MSE

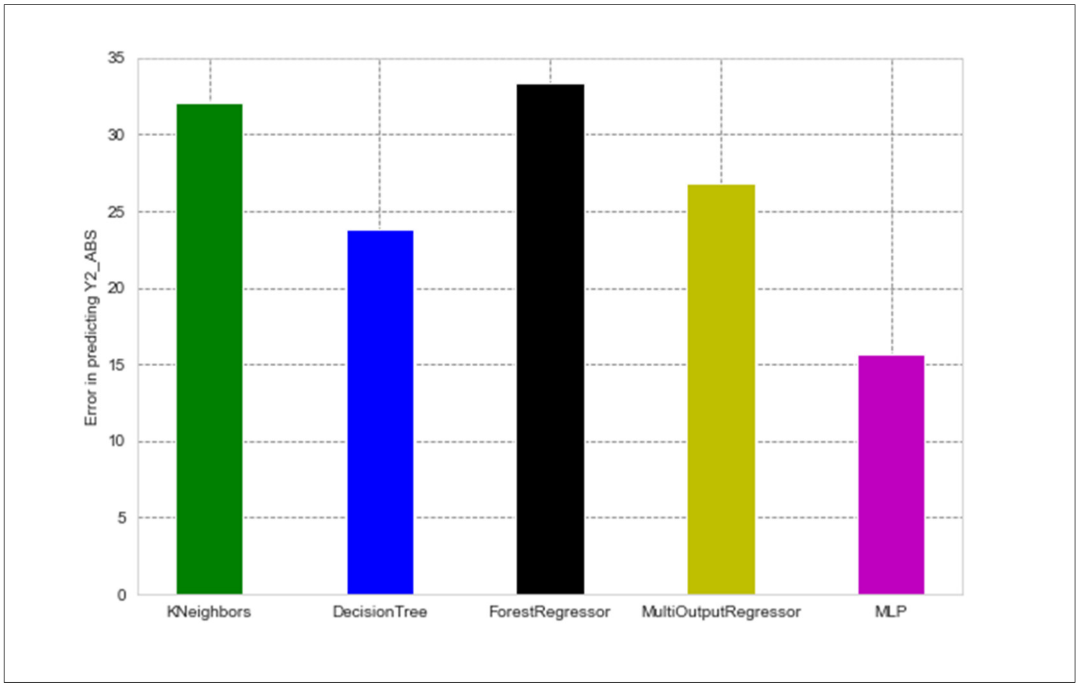


Figure 4.35 Y_2: ABS prediction MSE

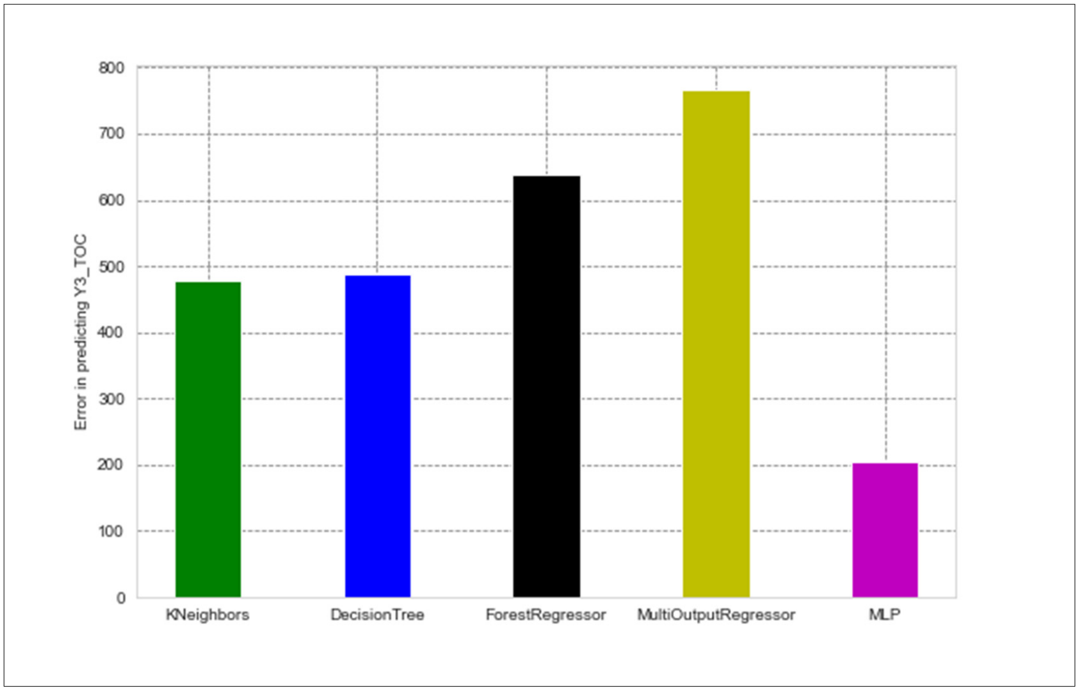


Figure 4.36 Y_3: TOC Degradation prediction MSE

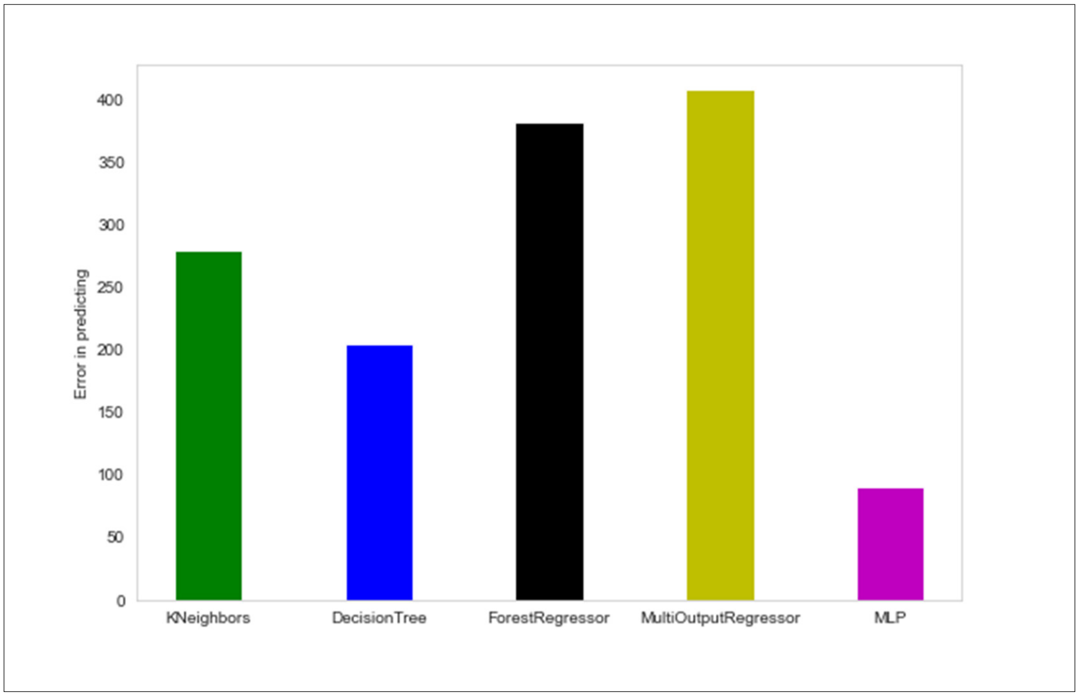


Figure 4.37 MSE comparison summary of all output variables

4.6.4.2 Multi-Output Model with PCA

Using the correction matrix, the eigenvalue was calculated to recognize where the most variance occurs. the eigenvector with the largest eigenvalue will be the direction in which the most variance occurs. The results in Figure 4.38 show that the greatest variance occurs at variables 6, 7, and 8 which corresponds to the ground truth or dependent variables, which is correct because the dependent variables depend on input variables. The calculated variance based on eigenvalues are:

$$0.0075, 0.0244, 0.0294, 0.0358, 0.0576, 0.1571, 0.3114, 0.3769$$

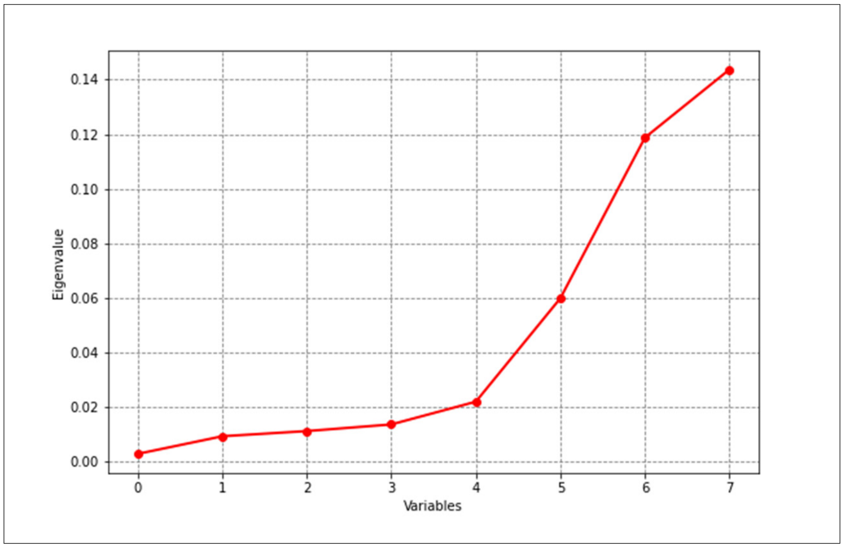


Figure 4.38 Eigenvalues based on the correlation matrix of all variables

In Figure 4.39, the eigenvalues are also calculated based on the input variables, and then the variance is computed:

$$0.1438, 0.1601, 0.1992, 0.2000, 0.2969$$

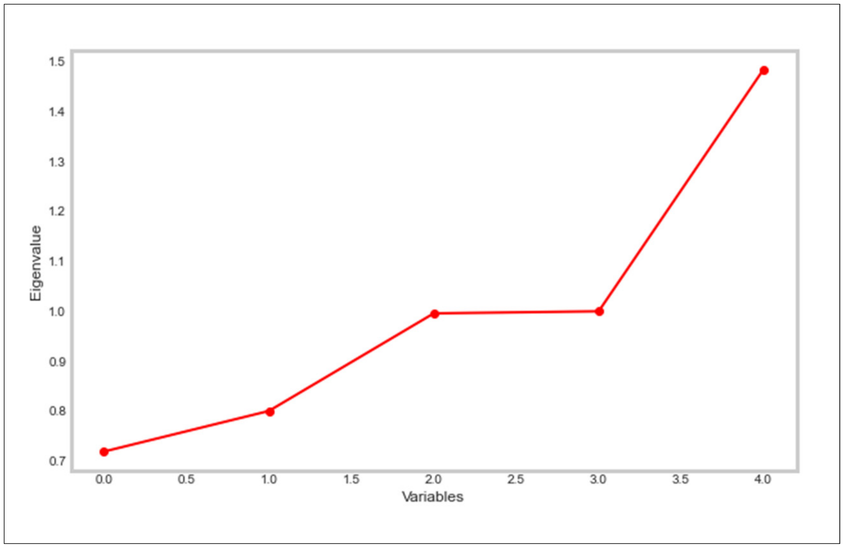


Figure 4.39 Eigenvalues based on the correlation matrix of input variables

In Figure 4.40, variance and PCA were considered. All output variables can be captured by considering all input variables. However, if all variables are utilized, there will be no use of PCA. Therefore, in Figure 4.40, the number of components was chosen to cover 95% of the variance. Consequently, four components were utilized (Figure 4.41).

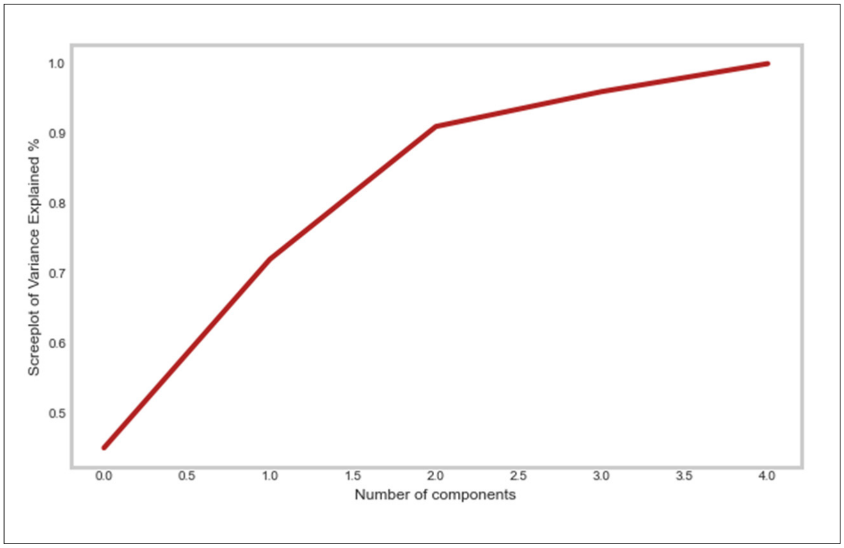


Figure 4.40 Variance and PCA

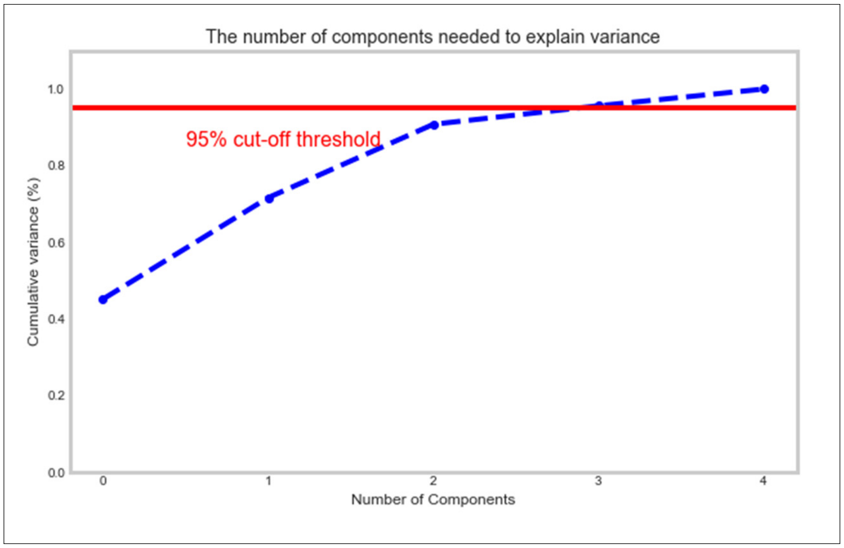


Figure 4.41 Number of components based on the variance

In Figure 4.42, to capture the contribution of each variable to the PCA components, a correlation matrix was used.

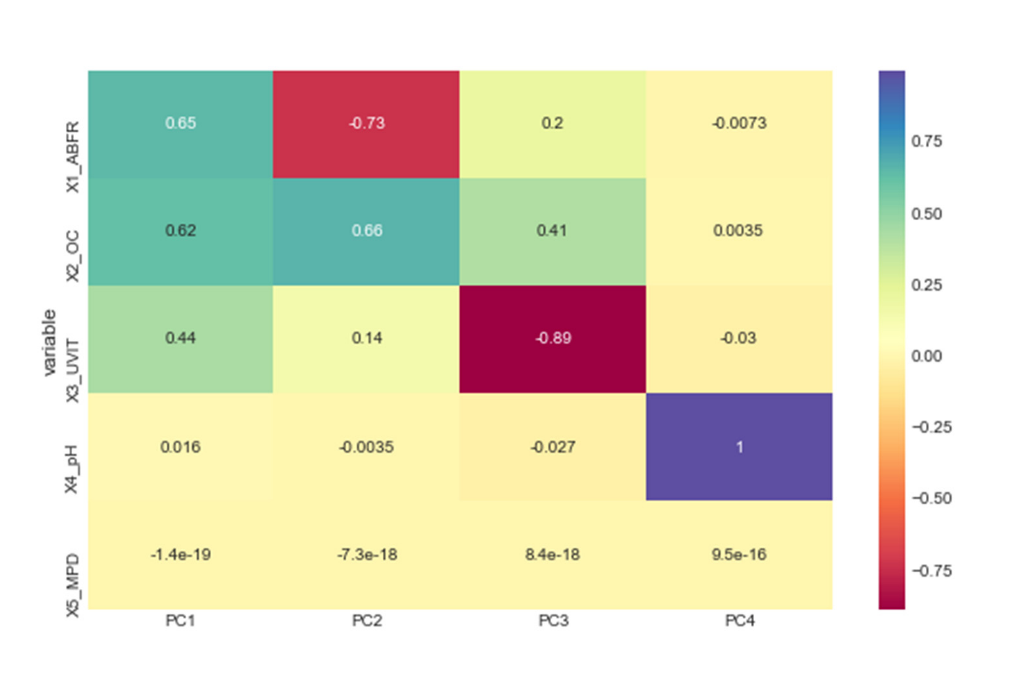
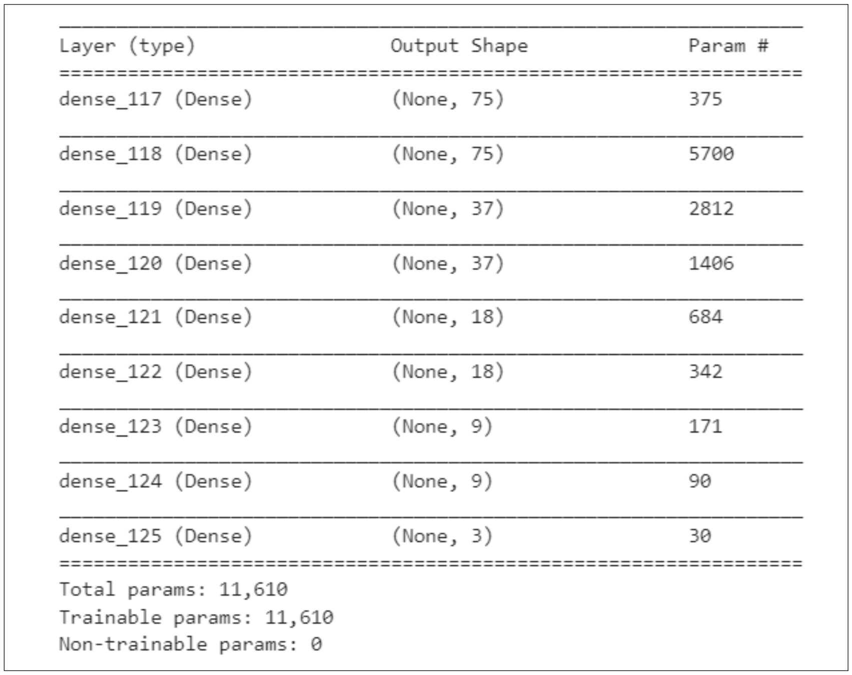


Figure 4.42 Correlation matrix plot for PCA loads

After the application of the PCA, four components were considered as inputs of the MLP. In Figure 4.43, the Multi-Layer Perceptron (MLP) used an input layer of 75 neurons, 7 hidden layers, and an output layer of 3 neurons. The three neurons of the output layer corresponding to 3 output variables.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_117 (Dense)            (None, 75)                375
_________________________________________________________________
dense_118 (Dense)            (None, 75)                5700
_________________________________________________________________
dense_119 (Dense)            (None, 37)                2812
_________________________________________________________________
dense_120 (Dense)            (None, 37)                1406
_________________________________________________________________
dense_121 (Dense)            (None, 18)                684
_________________________________________________________________
dense_122 (Dense)            (None, 18)                342
_________________________________________________________________
dense_123 (Dense)            (None, 9)                 171
_________________________________________________________________
dense_124 (Dense)            (None, 9)                 90
_________________________________________________________________
dense_125 (Dense)            (None, 3)                 30
=================================================================
Total params: 11,610
Trainable params: 11,610
Non-trainable params: 0
```

Figure 4.43 MLP after applying PCA

Figures 4.44 and 4.45 show the minimization of the error function in the MLP after applying PCA. MSE was used as an error function. As metrics, the MSE was compared to the MAE. The results show that the MLP model minimized the MSE better than the MAE.

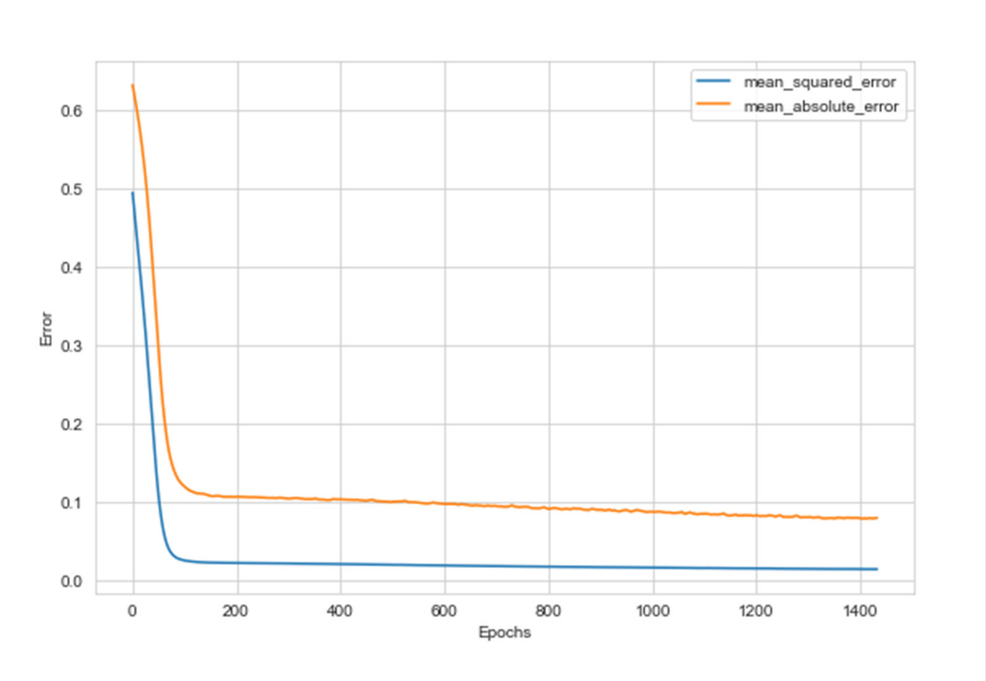


Figure 4.44 MLP error minimization after applying PCA

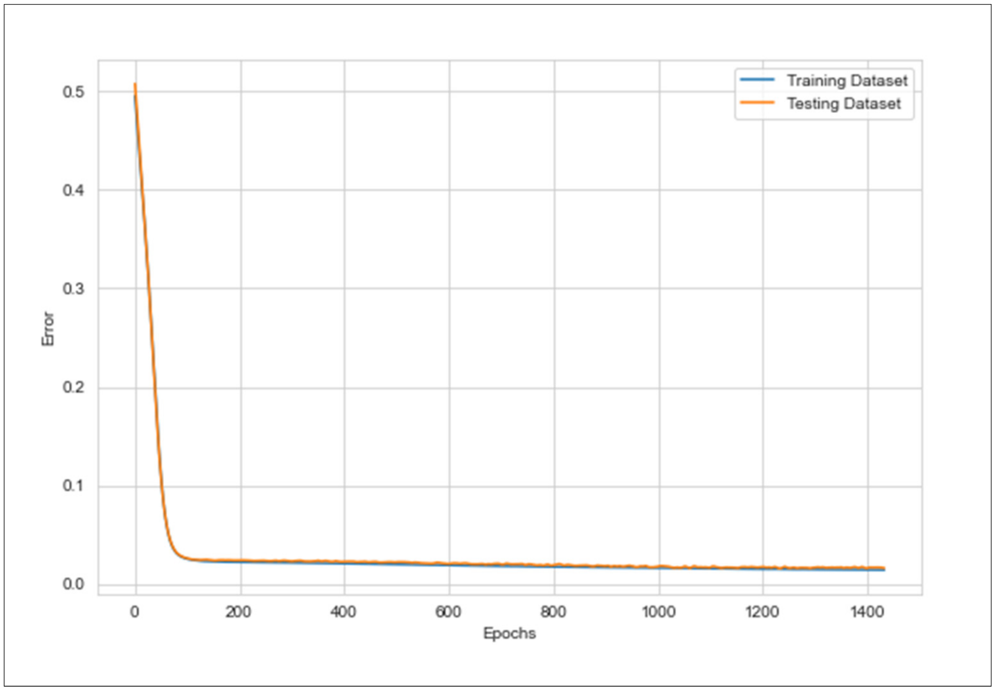


Figure 4.45 MLP error minimization based on dataset splitting after applying PCA

Figures 4.46 to 4.51 show actual and predicted data after the PCA application using different prediction models.

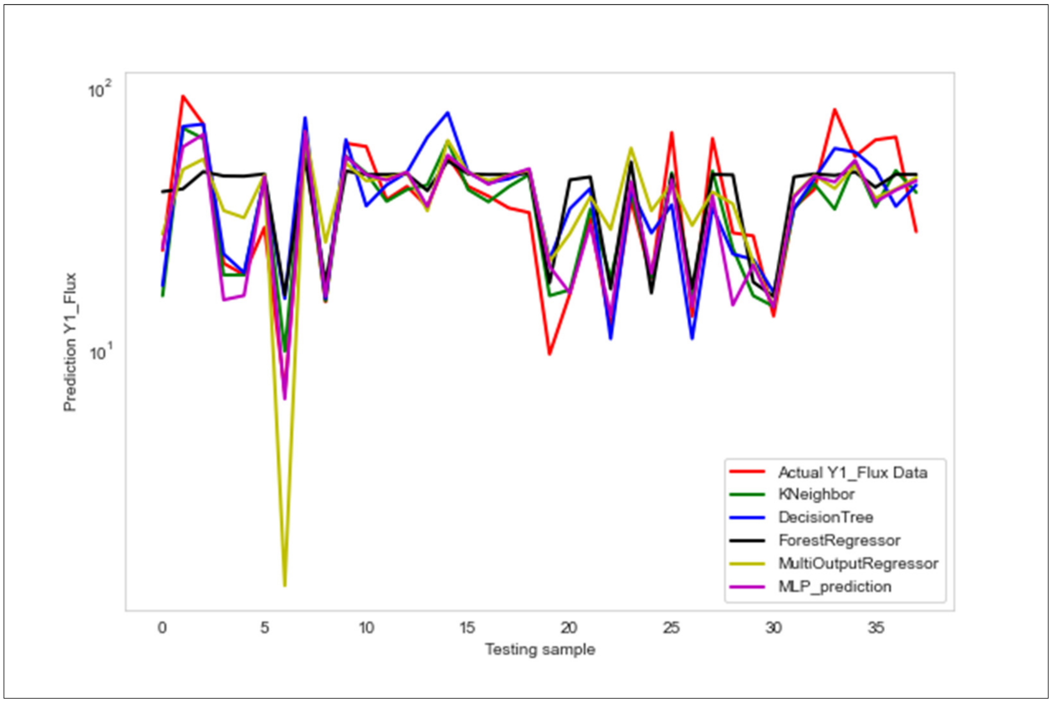


Figure 4.46 Y1: Permeate flux actual and predicted data after applying PCA

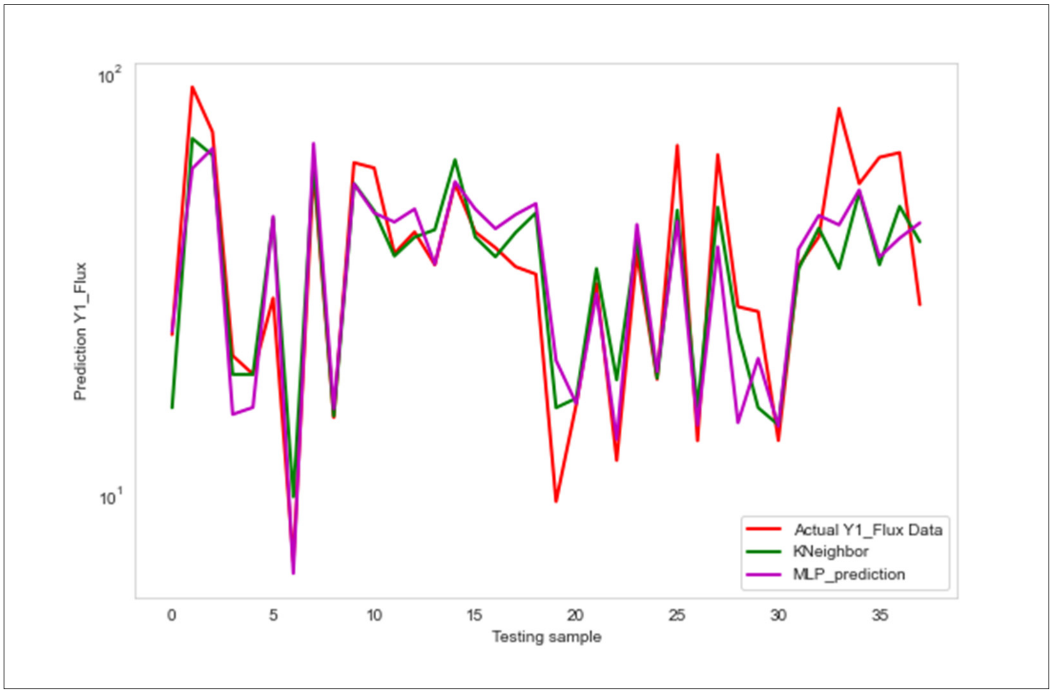


Figure 4.47 Y1: Permeate flux actual and predicted data with best-fit models after applying PCA

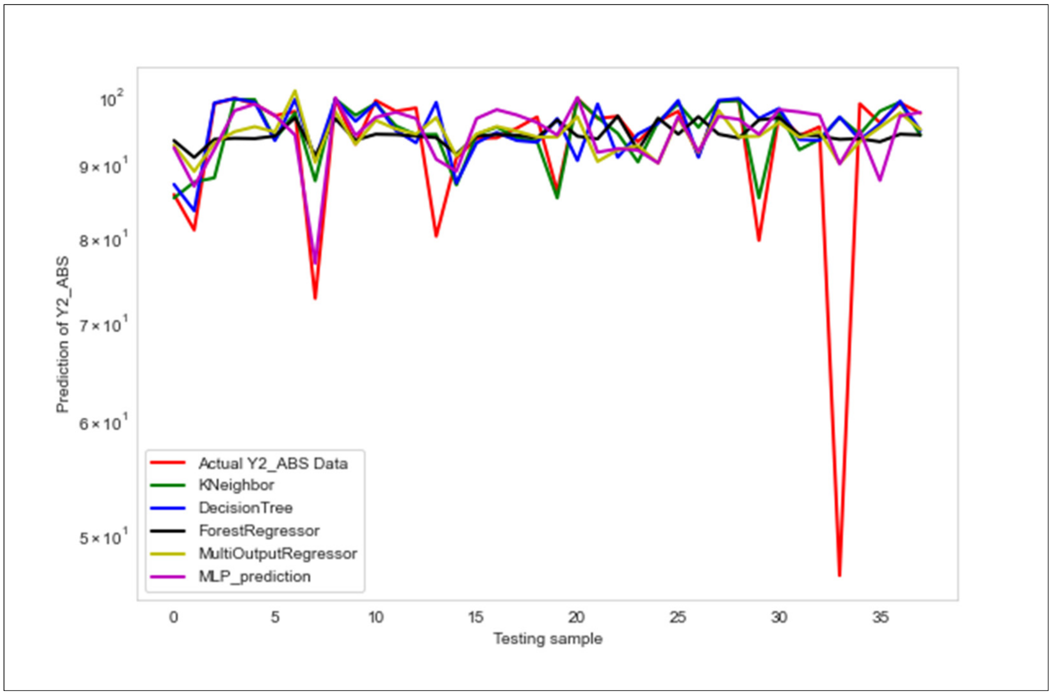


Figure 4.48 Y2: ABS actual and predicted data after applying PCA

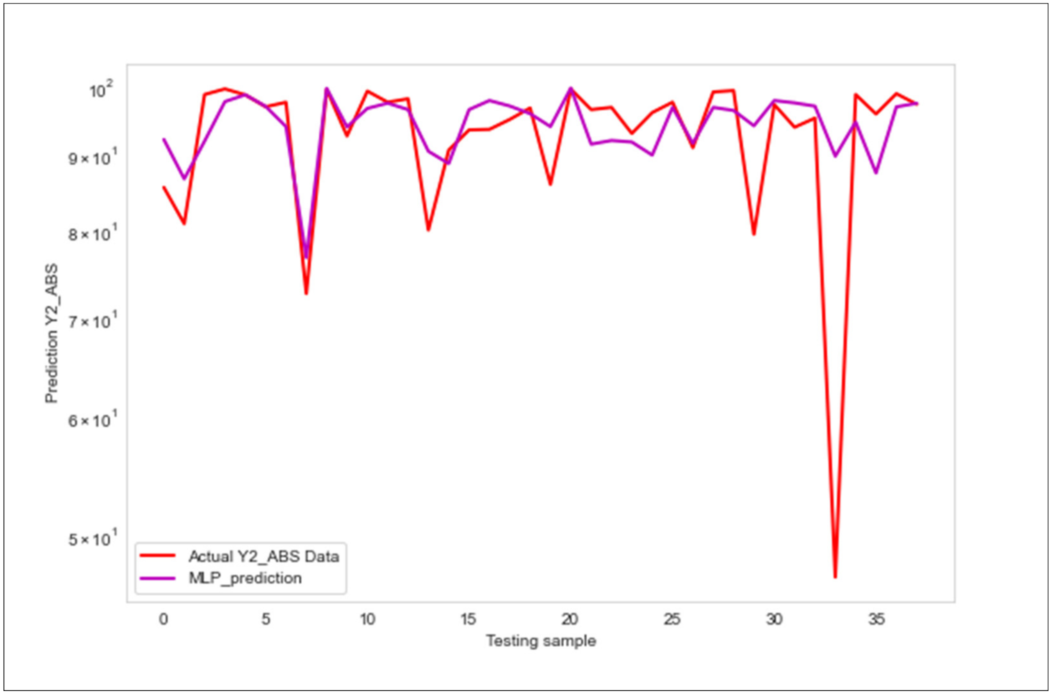


Figure 4.49 Y_2: ABS actual and predicted data with best-fit model after applying PCA

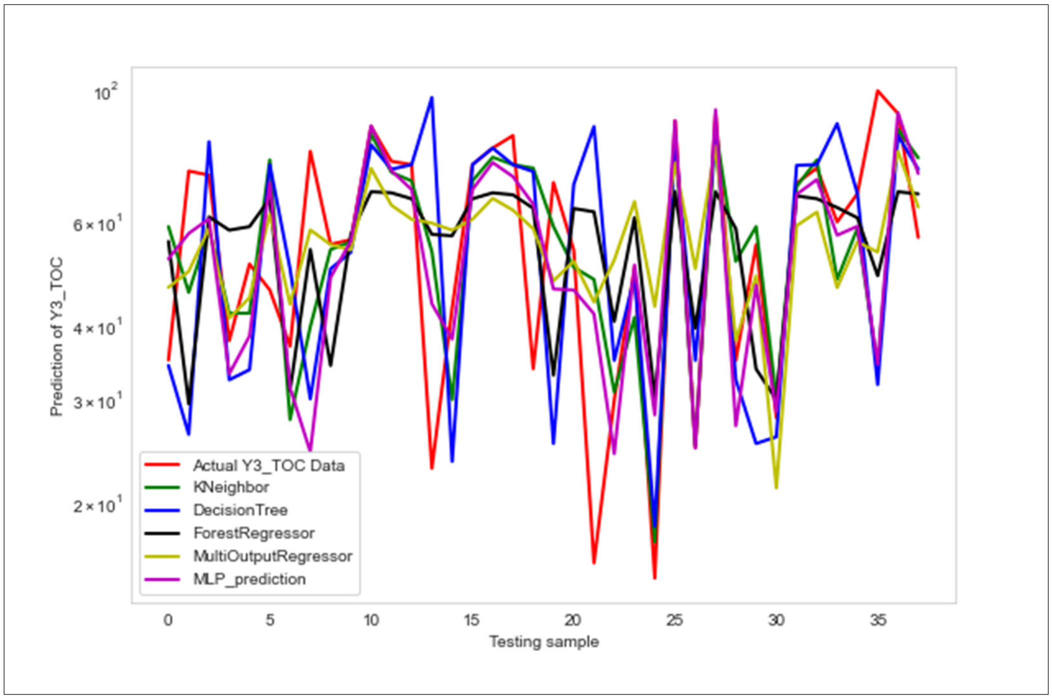


Figure 4.50 Y_3: TOC degradation actual and predicted data after applying PCA

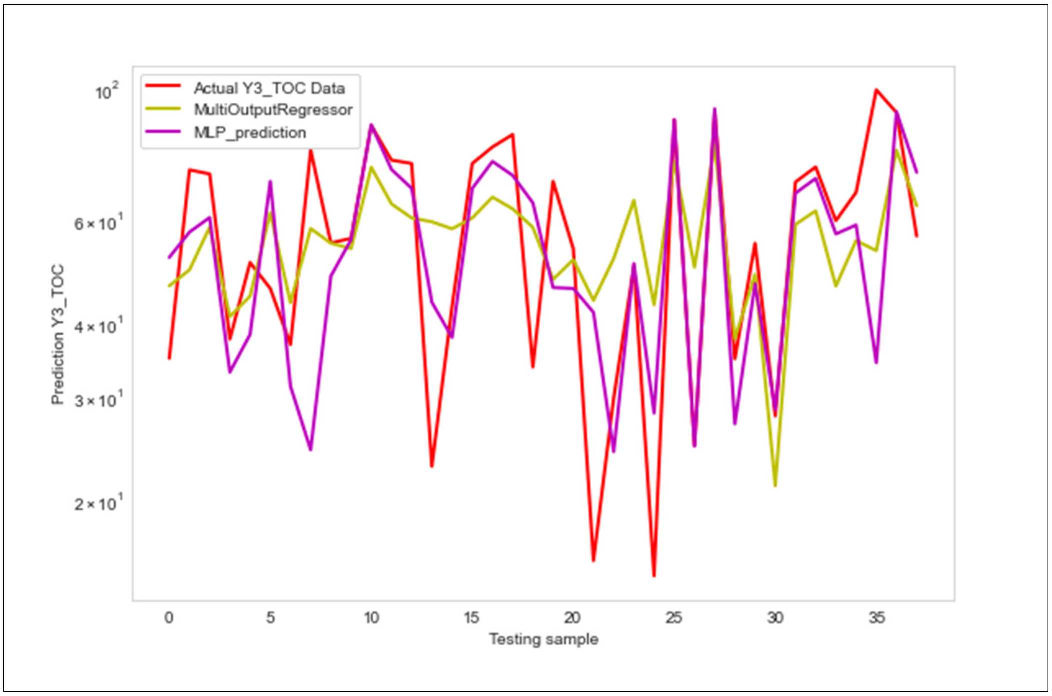


Figure 4.51 Y3: TOC degradation actual and predicted data with best-fit models after applying PCA

Figures 4.52 to 4.55 show the comparison of the MSE of the prediction using different prediction models. Figure 4.55 shows the MSE summary of the prediction models in comparison. The results show that the MLP performed better than other models. In conclusion, with or without PCA application, MLP has better performance.

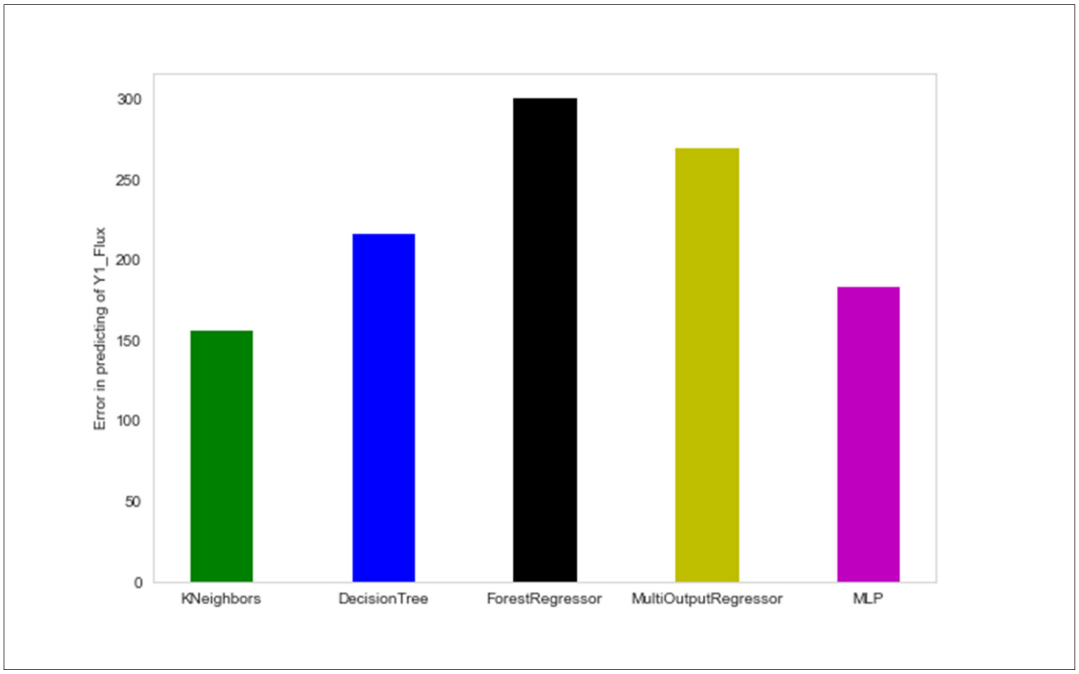


Figure 4.52 Y1: Permeate flux prediction MSE after applying PCA

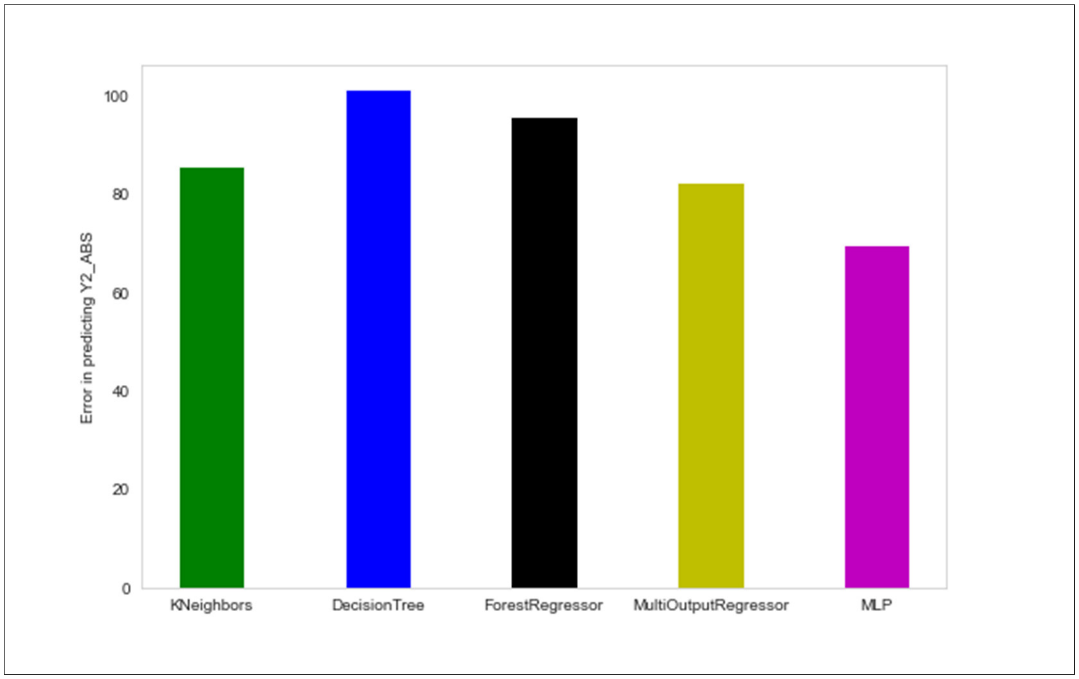


Figure 4.53 Y2: ABS prediction MSE after applying PCA

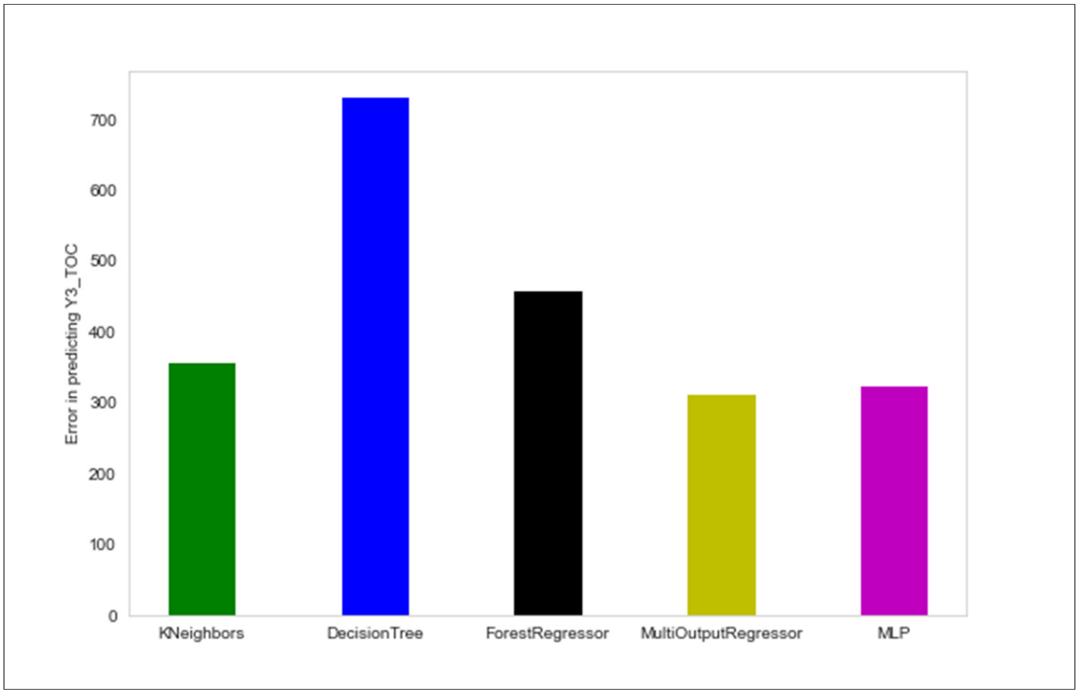


Figure 4.54 Y_3: TOC degradation prediction MSE after applying PCA

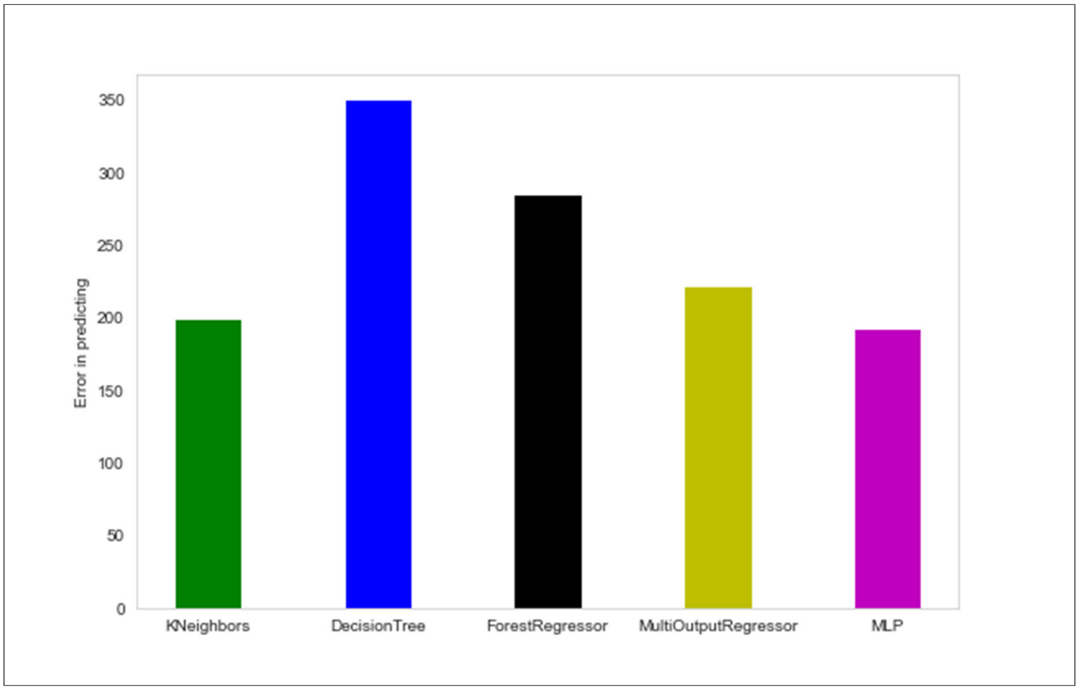


Figure 4.55 MSE comparison summary of all output variables after applying PCA

4.6.5 Multi-Layer Perceptron (MLP) and Design of Experiments (DOE) Comparison

In this section, the best-performed Multi-Output AI/ML model i.e., MLP with and without PCA application is compared to the Design of Experiments (DOE) model. The results of the MLP and DOE comparison in terms of MAE and MSE are as follows:

Table 4.8 DOE model, MAE and MSE comparison

Run	Factor					Response						Absolute Error			(Error)^2		
						Experimental			Predicted (DOE model)								
	A: ABFR	B: OC	C: UVIT	D: pH	E: MPD	R1: Flux	R2: ABS	R3: TOC	R1: Flux	R2: ABS	R3: TOC	Y1 (R1)	Y2 (R2)	Y3 (R3)	Y1 (R1)	Y2 (R2)	Y3 (R3)
1	3	625	75	7	60	45.28	95.24	52.36	49.04	94.10	52.71	3.76	1.14	0.35	14.16	1.30	0.13
2	5	1000	120	4	30	12.40	98.54	28.60	12.18	98.38	24.79	0.22	0.16	3.81	0.05	0.02	14.49
3	1	250	30	10	90	59.06	92.12	39.80	60.61	89.30	41.39	1.54	2.82	1.59	2.38	7.95	2.52
4	5	625	30	7	90	69.99	90.64	50.00	62.68	81.60	52.07	7.32	9.04	2.06	53.52	81.75	4.25
5	3	1000	75	10	30	45.36	96.39	66.52	41.21	97.39	70.93	4.15	1.00	4.41	17.26	1.00	19.49
6	1	625	120	4	60	56.63	90.36	91.34	52.75	84.53	85.68	3.88	5.83	5.66	15.04	33.95	32.01
7	2	800	45	9	40	33.31	95.85	49.09	35.54	92.63	53.04	2.23	3.22	3.96	4.95	10.37	15.67
8	4	435	90	5	80	64.51	85.10	39.04	68.36	91.82	37.86	3.85	6.72	1.19	14.85	45.17	1.41
												3.37	**3.74**	**2.88**	**15.28**	**22.69**	**11.24**
												MAE			**MSE**		

Table 4.9 MLP without PCA application, MAE and MSE comparison

Run	Factor					Response						Absolute Error			(Error)^2		
						Experimental			Predicted (MLP model without PCA)								
	A: ABFR	B: OC	C: UVIT	D: pH	E: MPD	R1: Flux	R2: ABS	R3: TOC	R1: Flux	R2: ABS	R3: TOC	Y1 (R1)	Y2 (R2)	Y3 (R3)	Y1 (R1)	Y2 (R2)	Y3 (R3)
1	3	625	75	7	60	45.28	95.24	52.36	43.96	93.48	65.26	1.32	1.76	12.90	1.75	3.09	166.51
2	5	1000	120	4	30	12.40	98.54	28.60	1.96	95.91	13.35	10.44	2.63	15.26	108.90	6.90	232.72
3	1	250	30	10	90	59.06	92.12	39.80	66.56	88.21	37.70	7.49	3.90	2.10	56.15	15.24	4.41
4	5	625	30	7	90	69.99	90.64	50.00	53.50	87.48	37.73	16.49	3.16	12.28	272.07	10.01	150.71
5	3	1000	75	10	30	45.36	96.39	66.52	35.31	91.42	80.21	10.05	4.97	13.69	100.95	24.70	187.37
6	1	625	120	4	60	56.63	90.36	91.34	57.73	93.83	88.06	1.10	3.47	3.28	1.22	12.04	10.78
7	2	800	45	9	40	33.31	95.85	49.09	28.78	89.55	46.40	4.53	6.30	2.69	20.53	39.66	7.22
8	4	435	90	5	80	64.51	85.10	39.04	67.86	91.77	39.21	3.36	6.67	0.17	11.26	44.55	0.03
												6.85	**4.11**	**7.80**	**71.60**	**19.53**	**94.97**
												MAE			**MSE**		

Table 4.10 MLP with PCA application, MAE and MSE comparison

Run	Factor					Response						Absolute Error			(Error)^2		
						Experimental			Predicted (MLP model with PCA)								
	A: ABFR	B: OC	C: UVIT	D: pH	E: MPD	R1: Flux	R2: ABS	R3: TOC	R1: Flux	R2: ABS	R3: TOC	Y1 (R1)	Y2 (R2)	Y3 (R3)	Y1 (R1)	Y2 (R2)	Y3 (R3)
1	3	625	75	7	60	45.28	95.24	52.36	52.39	93.95	68.96	7.11	1.29	16.60	50.57	1.66	275.69
2	5	1000	120	4	30	12.40	98.54	28.60	35.35	89.79	40.05	22.96	8.75	11.44	526.96	76.52	130.98
3	1	250	30	10	90	59.06	92.12	39.80	86.11	76.19	63.98	27.04	15.93	24.18	731.32	253.82	584.46
4	5	625	30	7	90	69.99	90.64	50.00	65.06	82.28	28.16	4.94	8.36	21.85	24.36	69.87	477.21
5	3	1000	75	10	30	45.36	96.39	66.52	31.07	93.13	50.80	14.29	3.26	15.72	204.24	10.66	247.16
6	1	625	120	4	60	56.63	90.36	91.34	30.59	92.65	45.75	26.04	2.28	45.59	678.03	5.22	2078.41
7	2	800	45	9	40	33.31	95.85	49.09	30.00	93.31	52.23	3.31	2.53	3.14	10.98	6.41	9.88
8	4	435	90	5	80	64.51	85.10	39.04	71.05	89.49	57.45	6.54	4.40	18.41	42.79	19.32	338.78
												14.03	**5.85**	**19.62**	**283.66**	**55.44**	**517.82**
												MAE			**MSE**		

According to Tables 4.8, 4.9, and 4.10, the MSE of the prediction models is tabulated in Table 4.11. The results show that DOE and MLP without PCA application performed better. In the MLP model, the output factors are dependent on the input factors and each other. This means that the output factors are not independent and the MSE of the prediction of all output variables together can be calculated, which is not possible based on the DOE model.

Table 4.11 MSE of the prediction models in comparison

Model	**MSE**			
	Y_1 (R_1)	**Y_2 (R_2)**	**Y_3 (R_3)**	**Y_t (R_t)**
DOE	15.28	22.69	11.24	-
MLP without PCA	71.60	19.53	94.97	66.43
MLP with PCA	283.66	55.44	517.82	285.64

4.6.6 An Optimization Approach to Find the Best Input Settings to Maximize Responses

Mathematical optimization was utilized to find the best input settings to maximize responses based on the dataset. Mathematical optimization is the mathematical approach to finding the best element regarding some criteria or constraints from a set of available alternatives. The following optimization problem is used to maximize each response for given inputs X, i.e., five input variables. Maximum and minimum values of the outputs are known to ensure that the optimization problem is bounded.

$$\max_{y} \textstyle\sum_{m=1}^{M} X\, y_m, \tag{4.18}$$

$$s.t{:}\ y_m \leq y_{max}, \tag{4.19}$$

$$y_m \geq y_{min,} \tag{4.20}$$

Where m is the size of output values.

The following optimization problem is used to find the best input settings that maximize all responses together for given responses Y, i.e., three output variables and the maximum and minimum values of the inputs.

$$\max_{x} \textstyle\sum_{n=1}^{N} Y\, x_n, \tag{4.21}$$

$$s.t{:}\ x_n \leq x_{max}, \tag{4.22}$$

$$x_n \geq x_{min,} \tag{4.23}$$

Where n is the size of input values.

To solve the above optimization problems, CVXPY and ECOS_BB were used as open-source tools to solve the optimization problems. CVXPY is a Python-based modeling language for convex optimization problems, and ECOS_BB is a mixed-integer nonlinear solver. The optimization results are tabulated as follows: (Tables 4.12, 4.13, and 4.14)

Table 4.12 DOE best input settings to maximize responses

DOE Model	**Input Factor**					**Max. Response**		
	A: ABFR	**B: OC**	**C: UVIT**	**D: pH**	**E: MPD**	**Y_1: Flux**	**Y_2: ABS**	**Y_3: TOC**
Y_1: Permeate Flux	1	250	-	4	90	100 L/m^{2}hr	-	-
Y_2: ABS	5	-	30	4	30	-	100%	-
Y_3: TOC	5	-	30	10	30	-	-	92.08%
Y_{total}: Local Optimum Point	5	250	108	10	30	62.5 L/m^{2}hr	100%	91%

Table 4.13 MLP without PCA best input settings to maximize responses

MLP without PCA	**Input Factor**					**Max. Response**		
	A: ABFR	**B: OC**	**C: UVIT**	**D: pH**	**E: MPD**	**Y_1: Flux**	**Y_2: ABS**	**Y_3: TOC**
Y_1: Permeate Flux	5	250	150	7	60	90.28 L/m^{2}hr	-	-
Y_2: ABS	5	250	60	7	60	-	100%	-
Y_3: TOC	0	1000	240	7	60	-	-	91.22%
Y_{total}: Local Optimum Point	5	10000	480	10	90	90.28 L/m^{2}hr	100%	91%

Table 4.14 MLP with PCA best input settings to maximize responses

MLP with PCA	**Input Factor**					**Max. Response**		
	A: ABFR	**B: OC**	**C: UVIT**	**D: pH**	**E: MPD**	**Y_1: Flux**	**Y_2: ABS**	**Y_3: TOC**
Y_1: Permeate Flux	3	1000	180	7	60	78.80 L/m^{2}hr	-	-
Y_2: ABS	5	1000	150	7	60	-	100%	-
Y_3: TOC	5	250	30	7	60	-	-	93.68%
Y_{total}: Local Optimum Point	5	10000	480	10	90	78.80 L/m^{2}hr	100%	94%

According to the results of the mathematical optimization, the best input settings to maximize responses based on the data set in MLP models with and without PCA were recognized and compared to the best input settings to maximize responses based on the DOE model.

However, the DOE is not capable to consider all three output factors simultaneously. Therefore, a desirability analysis was utilized to overcome this problem to find the most desirable local optimum point by considering all responses together. On the other hand, with the MLP, the output layer produces all three output factors simultaneously.

4.6.7 What are the Input Settings that Can Produce the Maximum Output Values?

Input search spaces were explored which produce each output to answer this question. In the following figures (Figures 4.56, 4.57, and 4.58), the x-axis represents the values of each variable, and the y-axis shows the ranges of the output variables. These figures clearly show that the output variables can reach the maximum value (100).

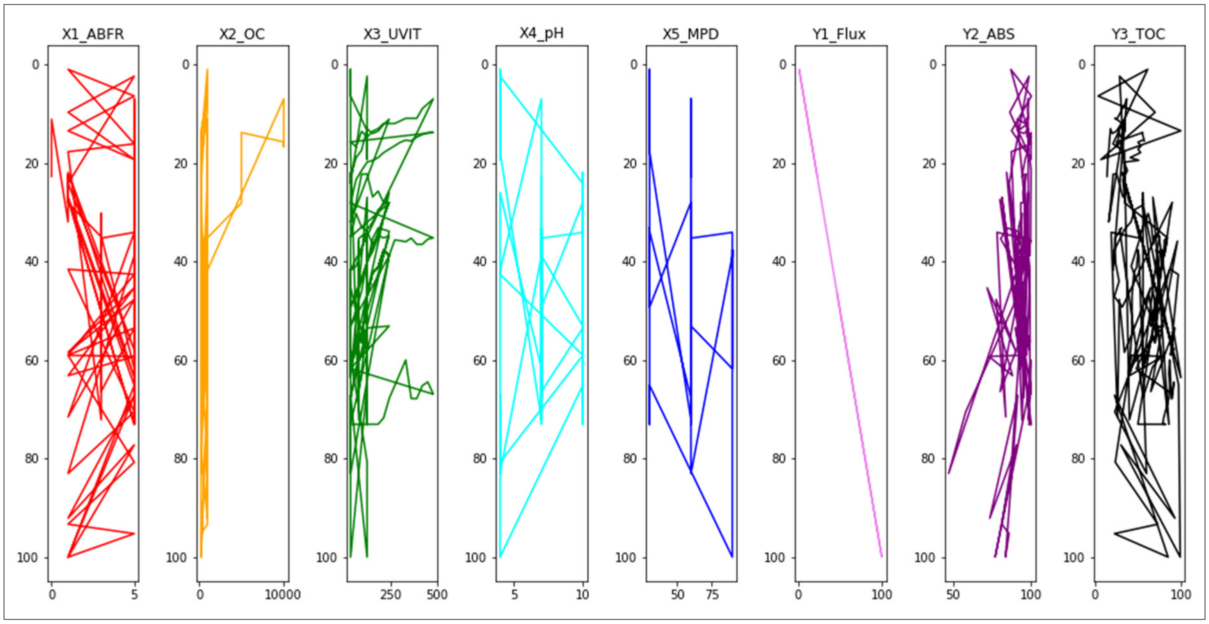


Figure 4.56 Input search spaces that produce Y_1: Permeate flux

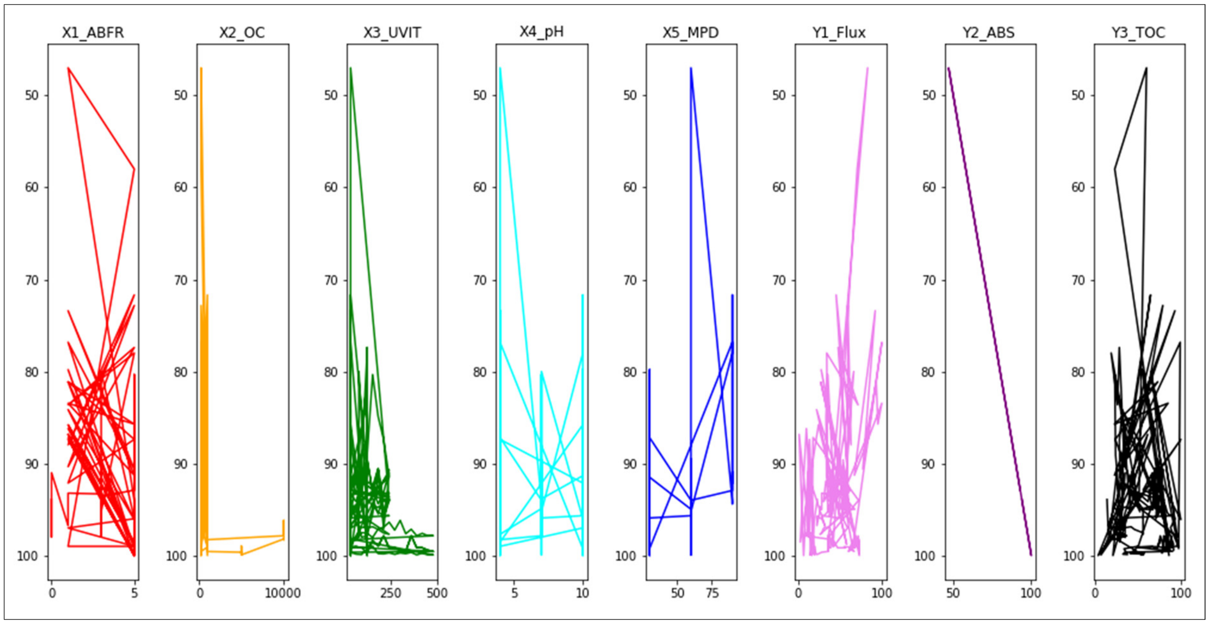


Figure 4.57 Input search spaces that produce Y_2: ABS

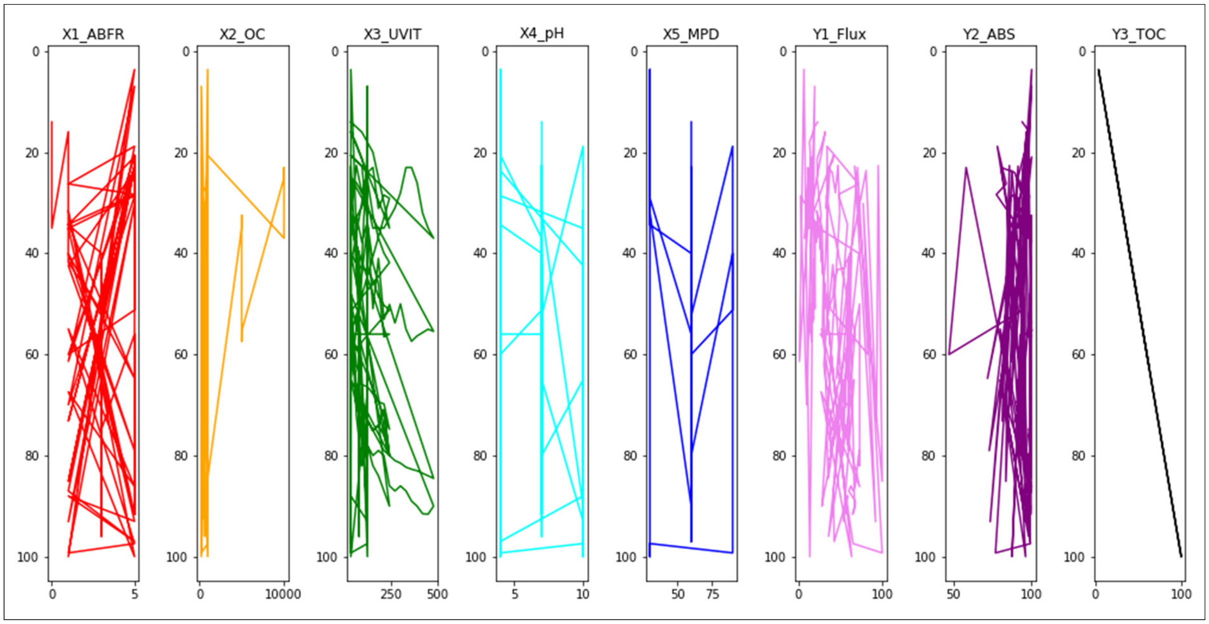


Figure 4.58 Input search spaces that produce Y_3: TOC degradation

After realizing that each output variable can reach 100 as its maximum value, the input settings that produce these maximum output values must be recognized. To achieve this goal, NumPy (NumPy, 2022, The fundamental package for scientific computing with Python) is a Python library that supports large multi-dimensional arrays and matrices for high-level mathematical functions such as divide-and-conquer approach (Aung, Phyo, Do, & Ogata, 2021) was used.

Given output variables, the argsort function of NumPy was used to return the indexes of the output variables in descending order of the variables' values. The argsort function uses the quicksort algorithm, which is based on the divide-and-conquer approach.

In the quicksort algorithm, an array is split into sub-arrays by selecting a pivot element, where the pivot is a selected element in the array. When splitting the array, the pivot element should be positioned so that elements less than the pivot are on the left side, while the elements greater than the pivot are on the right side of the pivot. Then, the left and right elements or sub-arrays are also divided by applying the same approach. The process continues until each sub-array has a single element. When the sub-array has a single element means that elements are sorted. Finally, the algorithm combined elements to form a sorted array (Programiz, Quicksort

Algorithm, 2022). The figure below is an illustrative example of the quicksort algorithm (Figure 4.59).

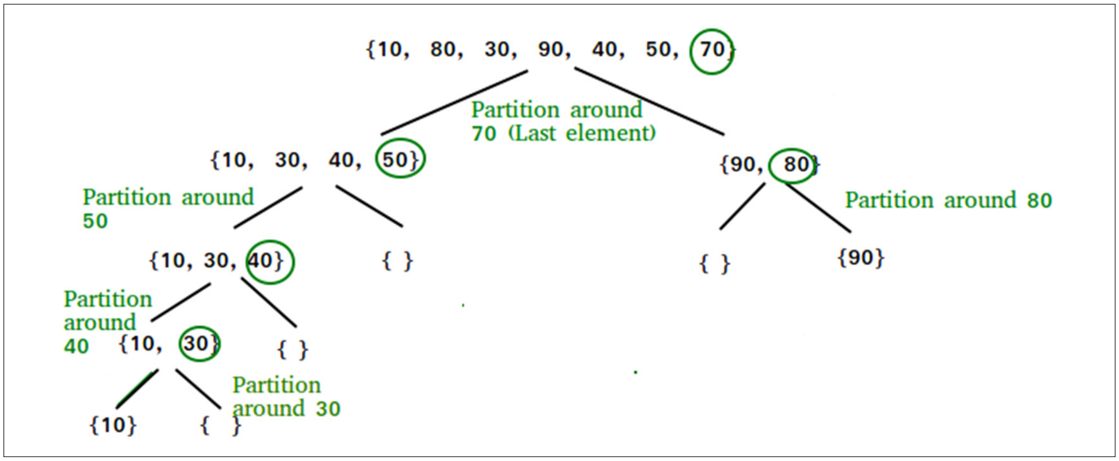
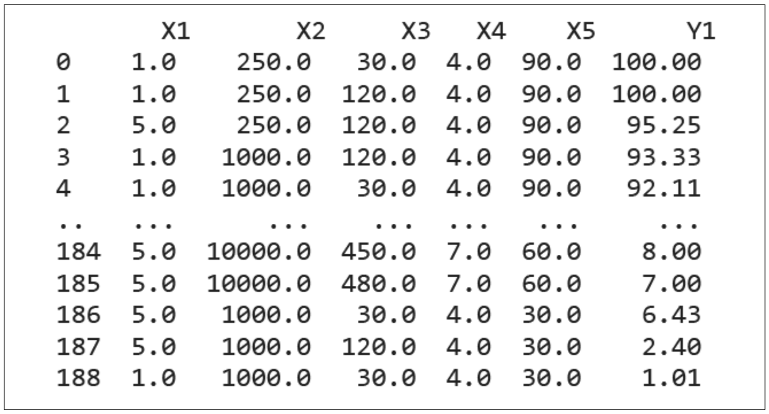


Figure 4.59 Illustrative example of Quicksort algorithm taken from GeeksforGeeks (2022, QuickSort)

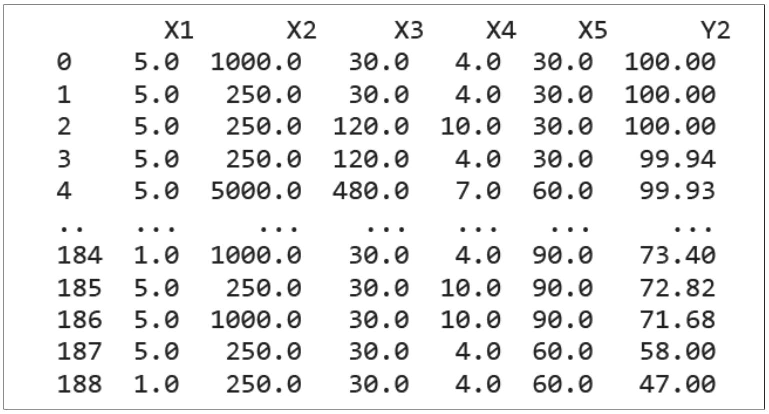
After performing the quicksort algorithm, the argsort function returns the indexes of the values of the output variables. Since input and output variables have the same length, we use the indexes of output variables returned by the argsort function and match the indexes with the input variables using the take function of NumPy. The take function returns the elements of the array along the mentioned axis and indices (GeeksforGeeks, Take function, 2022). Matching indexes of output variables with values of input variables help to return inputs that can produce maximum output variables.

Table 4.15 Inputs that can produce maximum Y_1



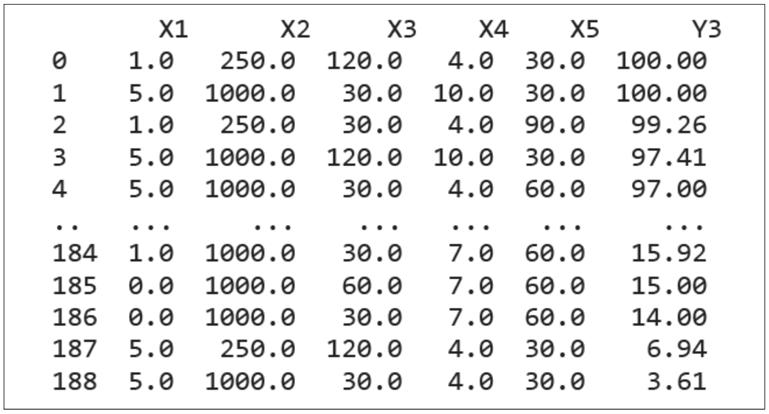
```
         X1       X2     X3   X4    X5      Y1
0     1.0    250.0   30.0  4.0  90.0  100.00
1     1.0    250.0  120.0  4.0  90.0  100.00
2     5.0    250.0  120.0  4.0  90.0   95.25
3     1.0   1000.0  120.0  4.0  90.0   93.33
4     1.0   1000.0   30.0  4.0  90.0   92.11
..    ...      ...    ...  ...   ...     ...
184   5.0  10000.0  450.0  7.0  60.0    8.00
185   5.0  10000.0  480.0  7.0  60.0    7.00
186   5.0   1000.0   30.0  4.0  30.0    6.43
187   5.0   1000.0  120.0  4.0  30.0    2.40
188   1.0   1000.0   30.0  4.0  30.0    1.01
```

Table 4.16 Inputs that can produce maximum Y_2



```
         X1      X2     X3    X4    X5      Y2
0     5.0  1000.0   30.0   4.0  30.0  100.00
1     5.0   250.0   30.0   4.0  30.0  100.00
2     5.0   250.0  120.0  10.0  30.0  100.00
3     5.0   250.0  120.0   4.0  30.0   99.94
4     5.0  5000.0  480.0   7.0  60.0   99.93
..    ...     ...    ...   ...   ...     ...
184   1.0  1000.0   30.0   4.0  90.0   73.40
185   5.0   250.0   30.0  10.0  90.0   72.82
186   5.0  1000.0   30.0  10.0  90.0   71.68
187   5.0   250.0   30.0   4.0  60.0   58.00
188   1.0   250.0   30.0   4.0  60.0   47.00
```

Table 4.17 Inputs that can produce maximum Y_3



```
         X1      X2     X3    X4    X5      Y3
0     1.0   250.0  120.0   4.0  30.0  100.00
1     5.0  1000.0   30.0  10.0  30.0  100.00
2     1.0   250.0   30.0   4.0  90.0   99.26
3     5.0  1000.0  120.0  10.0  30.0   97.41
4     5.0  1000.0   30.0   4.0  60.0   97.00
..    ...     ...    ...   ...   ...     ...
184   1.0  1000.0   30.0   7.0  60.0   15.92
185   0.0  1000.0   60.0   7.0  60.0   15.00
186   0.0  1000.0   30.0   7.0  60.0   14.00
187   5.0   250.0  120.0   4.0  30.0    6.94
188   5.0  1000.0   30.0   4.0  30.0    3.61
```

The input settings that can produce the maximum output values are tabulated in the following table (Table 4.18).

Table 4.18 Inputs that can produce maximum output values

MLP Model	Input Factor					Max. Response		
	A: ABFR	B: OC	C: UVIT	D: pH	E: MPD	Y_1: Flux	Y_2: ABS	Y_3: TOC
Y_1: Permeate Flux	1	250	30	4	90	100 L/m^2 hr	-	-
Y_1: Permeate Flux	1	250	120	4	90	100 L/m^2 hr	-	-
Y_2: ABS	5	1000	30	4	30	-	100%	-
Y_2: ABS	5	250	30	4	30	-	100%	-
Y_2: ABS	5	250	120	10	30	-	100%	-
Y_3: TOC	1	250	120	4	30	-	-	100%
Y_3: TOC	5	1000	30	10	30	-	-	100%
Y_{total}: Local Optimum Point	-	-	-	-	-	100 L/m^2 hr	100%	100%

4.6.8 Conclusion

Multi-Layer Perceptron (MLP) and five input variables were utilized to predict three output variables in this work. The results were grouped into two categories: multi-output model without PCA and multi-output model with PCA. MLP was compared to regression algorithms such as K-Neighbors Regression, Decision Tree Regressor, Random Forest Regression, and Multi-Output Regressor. The results show that MLP performed better than K-Neighbors Regression, Decision Tree Regressor, Random Forest Regression, and Multi-Output Regressor. However, based on the structure of the dataset, a strong correlation for some variables in the correlation matrix was not recognized. Some variables had positive correlations, where the variables increase or decrease together, i.e., the variables are positively correlated. On the other side, some pairs of variables had negative corrections, which means that one increases while the other decreases, i.e., they are inversely correlated. This correlation between variables makes it more difficult to find a model that performs well in all output variables in both the multi-output model without PCA and the multi-output model with PCA.

Furthermore, MLP and DOE were compared in terms of MSE. We considered human experts produced the DOE results. Therefore, MLP showed less performance than DOE results. However, the DOE is not capable to consider all three output factors simultaneously. On the other hand, with the MLP, the output layer produces all three output factors simultaneously. MLP with PCA produced 285.64 MSE. On the other hand, MLP without PCA produced 62.02 MSE. From the two MLP models (with PCA and without PCA), we conclude that MLP without PCA has a better performance compared to other machine learning algorithms in the comparisons of this study.

CONCLUSION

Conclusion

Some conventional oily wastewater treatment methods are incapable of treating nanoscale oil particles. These technologies are constrained by high operating costs, low efficiency, a large footprint, and high energy consumption. The use of ultrafiltration (UF) membrane technology for oily wastewater treatment has several advantages, including high efficiency in oil droplet removal, low energy consumption, minimal chemical use (only for cleaning), and no production of harmful by-products.

According to the scope of the project, this work focuses solely on the available membrane input and performance factors in the treatment of refinery-produced oily wastewater offered by the AMTEC. The performance characteristics of the membrane which are relevant for the measurement are **flux** or permeate volume per unit area and unit time, oil rejection or absorbance (**ABS**), and total organic carbon (**TOC**) degradation. AMTEC specialists proposed the three responses under consideration. The permeation flux (***F***) represents SMUF productivity. The rejection rates ($\boldsymbol{R_1}$ and $\boldsymbol{R_2}$) of PVDF hollow fibers are related to permeate water quality, which consists of absorbance (ABS) and total organic carbon (TOC) degradation. ***F*** represents SMUF productivity, and $\boldsymbol{R_1}$ and $\boldsymbol{R_2}$ represent, respectively, the filtration efficiency of hollow fiber UF membrane and photocatalytic degradation efficiency of TiO_2 catalysts. TOC degradation samples were collected from feed, whereas oil rejection samples were collected from permeate. The effect of the feed factors i.e., ABFR, OC, UVIT, pH, and MPD on membrane performance characteristics i.e., Flux, ABS, and TOC was studied to obtain optimum operating conditions for the SMPUF system.

In addition to significantly improving performance in degrading synthetic cutting oil wastewater, the laboratory-scale SMPR also produced high-quality permeate at a comparatively low operating cost. In addition, the PVDF-TiO_2 composite membrane under UV irradiation could effectively degrade the synthetic oily wastewater. As a function of permeate flux and quality, the effects of various important operating factors, including ABFR, OC, UVIT, pH, and MPD on membrane performances were examined. High feed concentrations had a tendency to cause a thicker oil layer to form, which was detrimental to photocatalytic degradation and permeate flux, but the use of a higher ABFR was able to mitigate the effects to some extent. When operated under optimum conditions, the fabricated membrane for TOC degradation and oil rejection was able to accomplish more than 90% TOC degradation. Overall, this study supports the development of SMPUF in the actual oily wastewater industry as well as provides helpful information for the simultaneous separation and degradation of oily wastewater.

A series of experiments are performed in the study to determine the effects of multiple factor levels, such as ABFR, OC, UVIT, pH, and MPD, on the output values of permeate flux, ABS, and TOC degradation. The submerged membrane photocatalytic ultrafiltration system's optimum process operating conditions were found and recorded. The influence of the simultaneous SMPUF oily wastewater treatment process conditions on membrane performances is demonstrated using a systematic experimental strategy based on DOE and AI/ML methodologies. The findings can be applied to the treatment of oily wastewater produced by refineries under various controllable conditions. And the findings will assist in determining the optimum submerged membrane photocatalytic ultrafiltration (SMPUF) system configuration. This could be a new paradigm for modeling and optimizing the operating conditions of submerged membrane photocatalytic ultrafiltration (SMPUF). The numerical models developed through this study facilitate early recognition of utilized variables on membrane performances.

PCA methodology as an unsupervised multivariate statistical projection learning method was utilized to quantify the amount of useful information contained in the data, as opposed to noise or meaningless variation, and visualize all the information contained in a datasheet. PCA was used for detecting patterns in high-dimensional data and expressing the data in a way that highlights similarities and differences and makes it easier to identify the variables that cause correlations and anticorrelations between samples. It helped to determine how one sample differs from another, which variables contributed the most to this difference, and whether those variables contributed in the same way (i.e., correlated) or independently. In this study, the original data matrix was transformed into a set of new variables called loadings which converted several correlated or potentially correlated variables into several uncorrelated variables known as principal components. It was demonstrated that the principal components contained 80% to 95% of the system variability and that the latent factors were sufficient to predict the system's future behavior.

DOE as an efficient and cost-effective statistical method to model and analyze process variations was utilized in this study. It was applied to process improvement in such a way that process can be defined using several controllable variables. Through utilizing DOE, the process input variables which had a significant effect on performances were determined, and the SMPUF system performance was optimized. ANOVA was utilized to analyze experimental outputs when several independent sources of variation were presented. Through utilizing DOE the influences of the simultaneous process conditions on the performances under controllable conditions were investigated, and the optimum point within the range of the experimental model with the maximum desirability value of 82.5% was gained.

Artificial Intelligence (AI) is defined as learning and performing suitable techniques to solve problems and achieve desirable goals appropriate to the context. ML models were used as algorithms to learn from data. Five multi-output regression models were considered in this study: K-Neighbors Regression, Decision Tree Regressor, Random Forest Regression, Multi-Output Regressor, and Artificial Neural Networks (ANNs). K-Neighbors Regression, Decision Tree Regressor, Random Forest Regression, and Multi-Output Regressor as not brain-inspired ML regression algorithms, and ANNs as brain-inspired ML regression algorithm. ANNs utilized in this study used multi-layer neural networks, hence called Multi-Layer Perceptron (MLP). Normalization technique was used to make each data point have the same scale so that the features have the same importance. Then train-test split approach helped to evaluate the performance of the ML algorithms. The training dataset was used to fit the ML models, and models utilized the test dataset for prediction. Then the predicted values were compared to the expected values. Comparison between the actual data, i.e., the ground truth and the predicted data, and comparison of the MSE of the prediction using different prediction models were performed. The results indicated that the MLP performed better than other ML models with and without PCA applications.

The best-performed multi-output model i.e., MLP with and without PCA application was compared to the DOE model. And MLP without PCA indicated better performance compared to other ML algorithms. In the MLP models, the MSE of the prediction of all output variables together can be calculated, which is not possible based on the DOE model. We also considered human experts produced the DOE results. Therefore, MLP showed less performance than DOE results.

A mathematical optimization problem was used to find the best input settings to maximize each response individually and all responses together. To solve the optimization problems, a Python-based modeling language for convex optimization problems and a mixed-integer nonlinear solver were used as open-source tools to solve the optimization problems.

Meanwhile, to find which input settings can produce the maximum output values, input search spaces were first explored to show that the output variables can reach the maximum values (100). Afterward, the input settings that produced these maximum output values were recognized using NumPy which supports large multi-dimensional arrays and matrices for high-level mathematical functions.

Highlights

- Since the experiments lasted for a long time, to ensure that the characteristics of the feed solution remained unchanged over time, feed solution was added from another reservoir to the photocatalytic reaction vessel to maintain the volume of the synthetic cutting oil solution to about 14 L.
- The permeate flow rate was recorded using a flow meter. To measure the absorbance (ABS) and total organic carbon (TOC), a Hach Spectrophotometer and a Total Organic Carbon Analyzer were used, respectively.
- TOC degradation samples were collected from feed, whereas oil rejection samples were collected from permeate.
- Since statistical approaches are not able to find the best input setting to optimize all responses together. And they are just able to find the best input setting to optimize each response individually. Therefore, a desirability analysis was used for the DOE phase to overcome this problem to find the most desirable local optimum point by considering all responses together.
- In the multi-output regression models, the output variables are typically dependent upon the input variables and each other. This means that the output variables are not independent and may require a model that predicts output variables together. That is why AI/ML models (multi-output regression models) were utilized.
- Meanwhile, determining an appropriate set of structural and learning parameter values for AI/ML models is a difficult task and is considered based on trial and error. The considered structures for the MLP models with and without PCA were obtained based on the trial and error between different structures and are highly integrated with the structure of the dataset.
- In addition, the structure of the dataset and the correlation between variables makes it more difficult to find a model that performs well in all output variables in both the multi-output model without PCA and the multi-output model with PCA.
- In the data visualization part, correlation analysis is performed completely with two different methods: Pearson’s correlation coefficient and mean-centered variables correlation coefficient, between all input and output variables. Correlation coefficients

around zero mean that there is no linear correlation between the variables and that there may be curvilinearity.

- The correlation matrix of the mean-centered dataset was computed for better recognition of the correlations. In the regression model, mean centering subtracts the mean of the variable from all observations to reduce multicollinearity in the dataset. Multicollinearity consists of high intercorrelations between two or more input factors in the data set. This happens when two or more input factors are expected to affect the value of an output factor independently but are observed to be highly correlated. One of the drawbacks of multicollinearity is misleading results in determining how well each input factor is capable to predict an output factor.
- The correlation matrix clearly shows the linear relationships between the factors in the positive and negative directions. However, the correlation matrix may not show desirable relationship results for nonlinear relationships. To overcome this problem, non-linear rank-based correlation coefficients such as Spearman's rank correlation coefficient and Kendall rank correlation coefficient, as well as Predictive Power Score (PPS) analysis, can be used to show non-linear relationships between variables.
- 36 treatment combinations were collected for the DOE phase. 44 observations were collected randomly for PCA. 153 treatment combinations plus 36 treatment combinations from the DOE phase were collected for the AI/ML phase. i.e., $153 + 36 = 189$ treatment combinations were considered for AI/ML phase. 80% of 189 treatment combinations were considered for training the AI/ML models. 20% of 189 treatment combinations were considered for testing the AI/ML models.
- In modeling, after creating a statistical or AI/ML model, the next phase is to monitor, control, or validate the created models based on the tracking errors, i.e., the residuals between the predicted and actual values. Therefore, to confirm the adequacy of the models obtained, 8 confirmation runs were performed for the validation of the DOE model, and MLP models with and without PCA. Please refer to section 4.6.5. In this section, the best-performed multi-output AI/ML model i.e., MLP with and without PCA application is compared to the DOE model in terms of MAE and MSE.
- Under optimum conditions, the SMPR can achieve high TOC degradation and oil rejection.

- Excessive module packing density degrades SMPR performances.
- A higher ABFR may mitigate the severe consequences of fouling during high feed concentration.
- Photocatalytic membranes provide superior permeate quality with less fouling for the treatment of oily wastewater.

RECOMMENDATIONS

Membrane technology is gradually transforming the treatment of water and wastewater. Much research has been conducted in this area over the years. However, there is still room for improvement in many areas. As fouling and high energy demand continue to be major issues in the processes, ongoing research is required to find a long-term solution, either through the implementation of stringent but low-cost pre-treatment processes or the development of fouling-resistant membranes. Continued research is required to fully comprehend the concepts, and thus the development of sustainable membranes will assist in making the process more viable for large-scale applications. Improvements are required to reduce the cost of the applications. More research should be done on membrane development and energy utilization in the future.

The nonlinearity and complicated process dynamics make it challenging to model and predict water quality. As a result, several research fields perceive the application of improved AI/ML models as a practical approach for analyzing historical data. The effectiveness of treatment and the sustainability of water use for end users can both be enhanced by models and forecasts of water quality variables. The protection of organisms and environmental pollution depend on wastewater treatment.

AI/ML approaches perform better when simulating complicated and nonlinear engineering problems. Different methods of water treatment demonstrated that AI/ML is a cutting-edge tool in the field of water and wastewater treatment that can overcome significant drawbacks of traditional modeling techniques and lower the likelihood of human mistakes. It also emphasizes how hybrid AI/ML models that incorporate several AI/ML techniques perform better and are more predictable.

Smart technology and AI/ML can be utilized to clarify and comprehend some of the most complicated problems affecting the water-based sectors. Some of the most important applications in water-based organizations and operations, including as those of water-treatment and wastewater-treatment facilities, natural systems, and water-based agriculture, have been successfully optimized, predicted, modeled, and controlled using AI/ML approaches.

Even if many of the research have been successfully published and reviewed, there are still a number of difficulties and restrictions with them. To further these intelligent applications, significant hurdles such as data management, public or governmental perspectives, predictability, and research transparency must be overcome. Although these difficulties and restrictions are undoubtedly evident, they do not negate the present research and advancements that indicate that AI/ML approaches and smart technologies have significant implications and possibilities for one of the most valuable resources on our planet.

APPENDIX

The experimental values of the five independent variables together with the responses

Standard	Variable					Response		
	X_1: ABFR (L/min)	X_2: OC (ppm)	X_3: UVIT (min)	X_4: pH (pH)	X_5: MPD (n)	Y_1: Flux (L/m^{2}hr)	Y_2: ABS (%)	Y_3: TOC (%)
						Experimental	Experimental	Experimental
1	0	1000	30	7	60	22.74	93.91	14.00
2	0	1000	60	7	60	18.95	96.30	15.00
3	0	1000	90	7	60	17.42	98.00	16.00
4	0	1000	120	7	60	13.06	97.86	18.00
5	0	1000	150	7	60	14.03	95.47	20.00
6	0	1000	180	7	60	13.55	91.23	25.00
7	0	1000	210	7	60	12.17	97.08	30.00
8	0	1000	240	7	60	11.13	91.00	35.00
9	1	1000	30	7	60	31.98	96.75	15.92
10	1	1000	60	7	60	28.11	96.01	18.35
11	1	1000	90	7	60	22.28	95.73	21.22
12	1	1000	120	7	60	22.28	96.75	25.33
13	1	1000	150	7	60	26.90	94.80	32.43
14	1	1000	180	7	60	25.72	94.71	36.94
15	1	1000	210	7	60	28.09	94.99	39.94
16	1	1000	240	7	60	28.09	93.22	41.82
17	3	1000	30	7	60	35.00	93.31	51.43
18	3	1000	60	7	60	30.26	94.15	61.97
19	3	1000	90	7	60	33.04	91.92	66.00
20	3	1000	120	7	60	33.91	91.92	66.35
21	3	1000	150	7	60	36.52	92.48	67.78
22	3	1000	180	7	60	34.78	90.53	69.42
23	3	1000	210	7	60	35.22	94.15	69.77
24	3	1000	240	7	60	34.78	93.59	74.81

Standard	Variable					Response		
	X$_1$: ABFR (L/min)	X$_2$: OC (ppm)	X$_3$: UVIT (min)	X$_4$: pH (pH)	X$_5$: MPD (n)	Y$_1$: Flux (L/m^{2}hr)	Y$_2$: ABS (%)	Y$_3$: TOC (%)
						Experimental	Experimental	Experimental
25	5	1000	30	7	60	61.76	92.89	56.00
26	5	1000	60	7	60	55.06	91.00	67.00
27	5	1000	90	7	60	56.47	94.92	70.00
28	5	1000	120	7	60	47.06	94.45	73.00
29	5	1000	150	7	60	42.31	93.76	75.00
30	5	1000	180	7	60	41.18	95.49	74.00
31	5	1000	210	7	60	37.65	93.13	71.00
32	5	1000	240	7	60	38.82	90.70	80.00
33	5	250	30	7	60	73.04	99.09	50.00
34	5	250	60	7	60	73.04	99.03	72.00
35	5	250	90	7	60	73.04	99.16	82.00
36	5	250	120	7	60	73.04	99.74	78.00
37	5	250	150	7	60	73.04	98.38	80.00
38	5	250	180	7	60	73.04	98.96	79.00
39	5	250	210	7	60	71.74	97.86	80.00
40	5	250	240	7	60	67.83	99.16	86.00
41	5	250	270	7	60	65.22	99.16	87.00
42	5	250	300	7	60	62.61	99.42	86.00
43	5	250	330	7	60	60.00	99.55	87.00
44	5	250	360	7	60	67.83	97.86	89.00
45	5	250	390	7	60	67.83	99.16	90.00
46	5	250	420	7	60	65.22	99.16	91.50
47	5	250	450	7	60	64.46	99.42	91.60
48	5	250	480	7	60	66.94	99.55	90.00
49	5	1000	30	7	60	61.76	99.09	48.00
50	5	1000	60	7	60	55.06	99.03	67.00
51	5	1000	90	7	60	56.47	99.16	70.00
52	5	1000	120	7	60	47.06	99.16	73.00

Standard	Variable					Response		
	X_1: ABFR (L/min)	X_2: OC (ppm)	X_3: UVIT (min)	X_4: pH (pH)	X_5: MPD (n)	Y_1: Flux (L/m^{2}hr)	Y_2: ABS (%)	Y_3: TOC (%)
						Experimental	Experimental	Experimental
53	5	1000	150	7	60	42.31	98.38	75.00
54	5	1000	180	7	60	41.18	98.96	74.00
55	5	1000	210	7	60	37.65	97.86	76.00
56	5	1000	240	7	60	38.82	99.16	80.00
57	5	1000	270	7	60	35.65	99.16	81.00
58	5	1000	300	7	60	35.78	99.42	81.50
59	5	1000	330	7	60	36.00	99.55	82.30
60	5	1000	360	7	60	35.27	97.86	82.70
61	5	1000	390	7	60	36.52	99.16	83.00
62	5	1000	420	7	60	36.52	99.16	83.50
63	5	1000	450	7	60	35.48	99.42	84.00
64	5	1000	480	7	60	35.22	99.56	84.50
65	5	5000	30	7	60	28.15	99.64	35.00
66	5	5000	60	7	60	23.38	99.85	32.46
67	5	5000	90	7	60	21.54	99.89	37.80
68	5	5000	120	7	60	19.85	99.27	33.76
69	5	5000	150	7	60	19.48	98.98	50.95
70	5	5000	180	7	60	18.92	99.79	41.00
71	5	5000	210	7	60	17.91	99.82	53.04
72	5	5000	240	7	60	17.20	99.71	50.10
73	5	5000	270	7	60	16.52	99.87	53.77
74	5	5000	300	7	60	15.82	99.72	50.00
75	5	5000	330	7	60	15.38	99.80	55.00
76	5	5000	360	7	60	15.54	99.66	57.39
77	5	5000	390	7	60	14.77	99.93	56.34
78	5	5000	420	7	60	14.25	99.74	55.66
79	5	5000	450	7	60	13.85	99.93	55.00
80	5	5000	480	7	60	13.84	99.93	55.43

Standard	Variable					Response		
	X_1: ABFR (L/min)	X_2: OC (ppm)	X_3: UVIT (min)	X_4: pH (pH)	X_5: MPD (n)	Y_1: Flux (L/m^{2}hr)	Y_2: ABS (%)	Y_3: TOC (%)
						Experimental	Experimental	Experimental
81	5	10000	30	7	60	15.74	98.20	25.00
82	5	10000	60	7	60	16.70	98.30	26.00
83	5	10000	90	7	60	13.57	97.50	28.00
84	5	10000	120	7	60	14.78	97.59	30.00
85	5	10000	150	7	60	13.57	98.04	35.00
86	5	10000	180	7	60	13.04	97.11	35.00
87	5	10000	210	7	60	12.00	96.92	35.00
88	5	10000	240	7	60	12.00	96.38	33.00
89	5	10000	270	7	60	12.00	97.24	31.00
90	5	10000	300	7	60	11.48	96.18	28.00
91	5	10000	330	7	60	10.96	97.63	23.00
92	5	10000	360	7	60	10.96	97.15	23.00
93	5	10000	390	7	60	10.96	97.85	26.00
94	5	10000	420	7	60	9.00	97.59	32.00
95	5	10000	450	7	60	8.00	97.85	35.00
96	5	10000	480	7	60	7.00	97.85	37.00
97	5	1000	30	4	60	41.80	98.28	20.58
98	5	1000	60	4	60	40.98	97.96	21.72
99	5	1000	90	4	60	37.70	96.92	22.69
100	5	1000	120	4	60	33.61	96.99	33.89
101	5	1000	150	4	60	31.97	96.87	41.95
102	5	1000	180	4	60	29.51	97.17	46.05
103	5	1000	210	4	60	28.46	97.55	56.51
104	5	1000	240	4	60	26.02	97.69	55.97
105	5	1000	30	7	60	61.76	94.90	56.00
106	5	1000	60	7	60	55.06	94.64	67.00
107	5	1000	90	7	60	56.47	92.93	70.00
108	5	1000	120	7	60	47.06	89.54	73.00

Standard	Variable					Response		
	X_1: ABFR (L/min)	X_2: OC (ppm)	X_3: UVIT (min)	X_4: pH (pH)	X_5: MPD (n)	Y_1: Flux (L/m^{2}hr)	Y_2: ABS (%)	Y_3: TOC (%)
						Experimental	Experimental	Experimental
109	5	1000	150	7	60	42.31	93.14	75.00
110	5	1000	180	7	60	41.18	90.72	77.00
111	5	1000	210	7	60	37.65	95.66	79.00
112	5	1000	240	7	60	38.82	93.84	80.00
113	5	1000	30	10	60	53.11	91.33	65.00
114	5	1000	60	10	60	50.00	93.23	70.00
115	5	1000	90	10	60	46.00	93.50	74.00
116	5	1000	120	10	60	40.00	93.88	78.00
117	5	1000	150	10	60	38.00	94.58	82.00
118	5	1000	180	10	60	35.00	95.29	84.00
119	5	1000	210	10	60	31.00	95.67	87.00
120	5	1000	240	10	60	28.00	95.67	90.00
121	5	1000	30	7	30	49.25	95.92	31.86
122	5	1000	60	7	30	47.15	94.73	22.58
123	5	1000	90	7	30	44.01	93.36	28.00
124	5	1000	120	7	30	40.33	89.09	23.99
125	5	1000	150	7	30	35.37	80.33	23.00
126	5	1000	180	7	30	33.20	84.88	31.11
127	5	1000	210	7	30	33.53	87.41	28.46
128	5	1000	240	7	30	33.87	91.49	29.09
129	5	1000	30	7	60	67.65	94.97	56.00
130	5	1000	60	7	60	61.47	95.00	67.00
131	5	1000	90	7	60	59.00	93.00	78.00
132	5	1000	120	7	60	53.53	89.60	73.00
133	5	1000	150	7	60	53.40	93.07	75.00
134	5	1000	180	7	60	53.30	90.64	74.00
135	5	1000	210	7	60	53.20	96.00	71.00
136	5	1000	240	7	60	53.10	94.00	80.00

Standard	Variable					Response		
	X_1: ABFR (L/min)	X_2: OC (ppm)	X_3: UVIT (min)	X_4: pH (pH)	X_5: MPD (n)	Y_1: Flux (L/m^{2}hr)	Y_2: ABS (%)	Y_3: TOC (%)
						Experimental	Experimental	Experimental
137	5	1000	30	7	90	61.76	92.93	40.00
138	5	1000	60	7	90	56.47	91.95	42.00
139	5	1000	90	7	90	55.06	90.89	42.93
140	5	1000	120	7	90	47.06	92.52	43.16
141	5	1000	150	7	90	42.31	92.68	45.32
142	5	1000	180	7	90	41.18	94.39	48.55
143	5	1000	210	7	90	37.65	93.25	50.24
144	5	1000	240	7	90	38.82	94.07	51.21
145	1	250	30	4	60	83.04	47	60
146	5	250	30	4	60	70.5	58	23
147	1	1000	30	4	60	41.58	99	87
148	5	1000	30	4	60	42.63	99	97
149	1	250	30	10	60	59.02	97	88
150	5	250	120	10	60	53.5	96	93
151	3	625	75	7	60	66.41	80	65
152	3	625	75	7	60	35.88	98	96
153	3	250	75	7	60	72.15	94.00	40.00

LIST OF BIBLIOGRAPHICAL REFERENCES

Al Aani, S., Bonny, T., Hasan, S. W., & Hilal, N. (2019). Can machine language and artificial intelligence revolutionize process automation for water treatment and desalination, Desalination, 458, 84-96. doi:https://doi.org/10.1016/j.desal.2019.02.005

Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Essen, B.C., Awwal, A.A., & Asari, V.K. (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. ArXiv, abs/1803.01164.

Altowayti, W. A. H., Shahir, S., Othman, N., Eisa, T. A. E., Yafooz, W. M. S., Al-Dhaqm, A., Soon, C. Y., et al. (2022). The Role of Conventional Methods and Artificial Intelligence in the Wastewater Treatment: A Comprehensive Review. Processes, 10(9), 1832. MDPI AG. Retrieved from http://dx.doi.org/10.3390/pr10091832

Antony, J., Capon, N. (1998). Teaching Experimental Design Techniques to Industrial Engineers. International Journal of Engineering Education 14 (5):335-343.

Aung MN, Phyo Y, Do CM, Ogata K. A Divide and Conquer Approach to Eventual Model Checking. Mathematics. (2021); 9(4):368. https://doi.org/10.3390/math9040368

Babanova, S., Artyushkova, K., Ulyanova, Y., Singhal, S., Atanassov, P. (2014). Design of experiments and principal component analysis as approaches for enhancing performance of gas-diffusional air breathing bilirubin oxidase cathode. Journal of Power Sources, 245, 389-397.

Busch, J., Cruse, A., & Marquardt, W. (2007). Modeling Submerged Hollow-Fiber Membrane Filtration for Wastewater Treatment. Journal of Membrane Science, 288(1-2), 94-111.

C.S. Ong, W.J. Lau, P.S. Goh, B.C. Ng, A.F. Ismail (2014). Investigation of submerged membrane photocatalytic reactor (sMPR) operating parameters during oily wastewater treatment process. Desalination, Volume 353, 2014, Pages 48-56, ISSN 0011-9164,

Cao, X., Ma, J., Shi, X., & Ren, Z. (2006). Effect of TiO2 Nanoparticle Size on the Performance of PVDF Membrane. Applied Surface Science, 253(4), 2003-2010.

Chen, H. Q., & Kim, A. S. (2006). Prediction of Permeate Flux Decline in Crossflow Membrane Filtration of Colloidal Suspension: ARadial Basis Function Neural Network Approach. Desalination, 192(1-3), 415-428.

Chen, H. Q., & Kim, A. S. (2006). Prediction of Permeate Flux Decline in Crossflow Membrane Filtration of Colloidal Suspension: ARadial Basis Function Neural Network Approach. Desalination, 192(1-3), 415-428.

Chenjing Cai, Shiwei Wang, Youjun Xu, Weilin Zhang, Ke Tang, Qi Ouyang, Luhua Lai, and Jianfeng Pei (2020). Transfer Learning for Drug Discovery. Journal of Medicinal Chemistry, 63 (16), 8683-8694

Croxton, Frederick E. & Klein, Sidney. & Cowden, Dudley J. (1968). Applied general statistics. London : Pitman

De Souza, D. I., Giacobbo, A., da Silva Fernandes, E., Rodrigues, M. A. S., de Pinho, M. N., & Bernardes, A. M. (2020). Experimental Design as a Tool for Optimizing and Predicting the Nanofiltration Performance by Treating Antibiotic-Containing Wastewater. Membranes (Basel), 10(7). doi:10.3390/membranes10070156

Dopar, M., Kusic, H., & Koprivanac, N. (2011). Treatment of simulated industrial wastewater by photo-Fenton process. Part I: The optimization of process parameters using design of experiments (DOE). Chemical Engineering Journal, 173(2), 267-279. doi:https://doi.org/10.1016/j.cej.2010.09.070

Hambli, R., Richir, S., Crubleau, P., Taravel, B. (2003). Prediction of Optimum Clearance in Sheet Metal Blanking Processes. The International Journal of Advanced Manufacturing Technology 22 (1):20-25. doi: 10.1007/s00170-002-1437-5.

He, X. (2017). Fabrication of Defect-Free Cellulose Acetate Hollow Fibers by Optimization of Spinning Parameters. Membranes (Basel), 7(2). doi:10.3390/membranes7020027

Hu, J., Kim, C., Halasz, P., Kim, J. F., Kim, J., & Szekely, G. (2021). Artificial intelligence for performance prediction of organic solvent nanofiltration membranes. Journal of Membrane Science, 619, 118513. doi:https://doi.org/10.1016/j.memsci.2020.118513

Hunt, J. (2019). A Beginners Guide to Python 3 Programming. Cham, Switzerland: Springer. Retrieved November 22, 2022. https://doi.org/10.1007/978-3-030-20290-3

Iacobucci, D., Schneider, M.J., Popovich, D.L. et al. Mean centering helps alleviate “micro” but not “macro” multicollinearity. Behav Res 48, 1308–1317 (2016). https://doi.org/10.3758/s13428-015-0624-x

Jawad, J., Hawari, A. H., & Zaidi, S. (2020). Modeling of forward osmosis process using artificial neural networks (ANN) to predict the permeate flux. Desalination, 484, 114427. doi:https://doi.org/10.1016/j.desal.2020.114427

Kamali, M., Appels, L., Yu, X., Aminabhavi, T. M., & Dewil, R. (2021). Artificial intelligence as a sustainable tool in wastewater treatment using membrane bioreactors. Chemical Engineering Journal, 128070. doi:https://doi.org/10.1016/j.cej.2020.128070

Kantardzic, M. (2011). Artificial Neural Networks Data Mining (pp. 199-234): John Wiley & Sons, Inc.

Li, L., Rong, S., Wang, R., & Yu, S. (2021). Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. Chemical Engineering Journal, 405, 126673. doi:https://doi.org/10.1016/j.cej.2020.126673

Lowe, M., Qin, R., & Mao, X. (2022). A Review on Machine Learning, Artificial Intelligence, and Smart Technology in Water Treatment and Monitoring. Water, 14(9), 1384. MDPI AG. Retrieved from http://dx.doi.org/10.3390/w14091384

Montgomery, D. C. (2008). Design and Analysis of Experiments. 7th ed. Hoboken, NJ: Wiley.

Montgomery, D. C., Runger, G. C. (2011). Applied Statistics and Probability for Engineers. Fifth ed: John Wiley & Sons.

Naessens, W., Maere, T., Gilabert-Oriol, G., Garcia-Molina, V., & Nopens, I. (2017). PCA as tool for intelligent ultrafiltration for reverse osmosis seawater desalination pretreatment. Desalination, 419, 188-196. doi:https://doi.org/10.1016/j.desal.2017.06.018

Obotey Ezugbe, E., & Rathilal, S. (2020). Membrane Technologies in Wastewater Treatment: A Review. Membranes, 10(5), 89. Retrieved from https://www.mdpi.com/2077-0375/10/5/89

Olczak J, Pavlopoulos J, Prijs J, Ijpma FFA, Doornberg JN, Lundström C, Hedlund J, Gordon M. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. Acta Orthop. 2021 Oct;92(5):513-525. doi: 10.1080/17453674.2021.1918389. Epub 2021 May 14. PMID: 33988081; PMCID: PMC8519529.

Peinemann, K.-V., & Pereira Nunes, S. (2010). Membrane Technology: Membranes for Water Treatment (Vol. 4): Wiley-VCH Verlag GmbH & Co. KGaA.

Peiris, R. H., Budman, H., Moresoli, C., & Legge, R. L. (2010). Understanding fouling behaviour of ultrafiltration membrane processes and natural water using principal component analysis of fluorescence excitation-emission matrices. Journal of Membrane Science, 357(1), 62-72. doi:https://doi.org/10.1016/j.memsci.2010.03.047

Peterson, K. L. (2007). Artificial Neural Networks and Their use in Chemistry Reviews in Computational Chemistry (pp. 53-140): John Wiley & Sons, Inc.

Rahmanian, B., Pakizeh, M., Mansoori, S. A. A., & Abedini, R. (2011). Application of Experimental Eesign Approach and Artificial Neural Network (ANN) for the Determination of Potential Micellar-Enhanced Ultrafiltration Process. Journal of Hazardous Materials, 187(1-3), 67-74. doi: 10.1016/j.jhazmat.2010.11.135

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.

Sze, V., Chen, Y., Yang, T., & Emer, J.S. (2017). Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Proceedings of the IEEE, 105, 2295-2329.

Uy, M., Telford, J. K. (2009). Optimization by Design of Experiment Techniques. Paper Read at Aerospace Conference, 2009 IEEE, 7-14 March 2009, at Big Sky, MT.

Wang, L., & Fu, K. (2007). Artificial Neural Networks Wiley Encyclopedia of Computer Science and Engineering: John Wiley & Sons, Inc.

Watt, J., Borhani, R., & Katsaggelos, A. (2020). Machine learning refined: foundations, algorithms, and applications (Second edition.). Cambridge University Press.

Yu, H., Li, X., Chang, H., Zhou, Z., Zhang, T., Yang, Y., Liang, H. (2020). Performance of hollow fiber ultrafiltration membrane in a full-scale drinking water treatment plant in China: A systematic evaluation during 7-year operation. Journal of Membrane Science, 613, 118469. doi:https://doi.org/10.1016/j.memsci.2020.118469

Yuliwati, E. (2012). Treatment of Refinery Prodused Wastewater Using Hydrophilic Polyvinylidene Fluoride Hollow Fiber Ultrafiltration Membrane. Ph.D Thesis, Universiti Teknologi Malaysia (UTM)

Yuliwati, E., & Ismail, A. F. (2011). Effect of Additives Concentration on the Surface Properties and Performance of PVDF Ultrafiltration Membranes for Refinery Produced Wastewater Treatment. Desalination, 273(1), 226-234.

Yuliwati, E., Ismail, A. F., Lau, W. J., Ng, B. C., Mataram, A., & Kassim, M. A. (2012). Effects of Process Conditions in Submerged Ultrafiltration for Refinery Wastewater Treatment: Optimization of Operating Process by Response Surface Methodology. Desalination, 287, 350-361.

Yuliwati, E., Ismail, A. F., Matsuura, T., Kassim, M. A., & Abdullah, M. S. (2011). Effect of Modified PVDF Hollow Fiber Submerged Ultrafiltration Membrane for Refinery Wastewater Treatment. Desalination, 283, 214-220.