Development of An Integrated AI-Based Online System for Lake Chlorophyll-a Concentration Modeling and Monitoring (CMMOS)

Yanbin Zhuang

A Thesis

in

The Department

of

Building, Civil and Environmental Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Applied Science (Civil Engineering) at

Concordia University

Montreal, Quebec, Canada

June 2024

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Yanbin Zhuang

Entitled: Development of An Integrated AI-Based Online System for Lake Chlorophyll-a Concentration Modeling and Monitoring

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Civil Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Catherine Mulligan Chair

Dr. Catherine Mulligan Examiner

Dr. Ali Nazemi Examiner

Dr. Zhi Chen Thesis Supervisor

Approved by Dr.S.Samuel Li, Chair of Department

12 June 2024

Dr. Mourad Debbabi, Dean, Faculty of Engineering and Computer Science

ABSTRACT

Development of An Integrated AI-Based Online System for Lake Chlorophyll-a Concentration Modeling and Monitoring (CMMOS)

Yanbin Zhuang

Concordia University, 2024

This study presents the development of the Chlorophyll-a (Chl-a) Modeling and Monitoring Online System (CMMOS), an innovative artificial intelligence (AI)-based tool designed to enhance the monitoring and prediction of Chl-a concentrations in lake ecosystems. Traditional methods of monitoring these concentrations face limitations in real-time data processing and handling complex environmental interactions. CMMOS addresses these challenges by integrating a sophisticated array of machine learning models, including Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Gradient Boosting Tree (GBT), Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), K Nearest Neighbors (KNN), Multiple Linear Regression (MLR), and Extreme Gradient Boosting (XGBoost). The system's efficacy was rigorously evaluated using comprehensive datasets from Lake Champlain and Lake Simcoe. In Lake Champlain, the RF model demonstrated high predictive accuracy with a Root Mean Squared Error (RMSE) of 1.4667 μg/L and a Mean Absolute Percentage Error (MAPE) of 27.89%. For Lake Simcoe, the R F model also showed superior performance with an RMSE of 0.2671 μg/L and a MAPE of 6.01%. These results highlight the robustness and reliability of the system across different environmental contexts.

Data preprocessing techniques such as Missing Value Imputation, Outlier Detection, and Feature Selection proved critical in enhancing the accuracy and reliability of these models. CMMOS contributes to the field of environmental science by offering a real-time, data-driven approach to lake water quality management. The system facilitates dynamic monitoring and predictive analysis, enabling stakeholders to make informed decisions promptly. It illustrates the substantial advantages of utilizing AI in ecological monitoring and management. Recommendations for future work include further optimization of machine learning models, exploration of ensemble techniques to refine predictive accuracy, expansion of the system to include more diverse environmental variables, and enhancements to the user interface to better

serve various stakeholders. This thesis lays a robust foundation for future advancements in AI applications for environmental monitoring, aiming to improve the sustainability and effectiveness of lake management practices.

ACKNOWLEDGEMENTS

I am writing to express my sincere gratitude to my supervisor, Professor Chen, for his professional guidance and selfless support throughout my postgraduate study. Professor Chen patiently listened to my thoughts and gave me valuable suggestions. He also taught me how to relentlessly pursue the true meaning of scientific research, which is of immeasurable significance to my research career and future path.

At the same time, I also want to thank every colleague in the laboratory. Their wisdom and efforts make our team full of vitality, and every moment of cooperation is a driving force for my scientific research ability. Our friendship and cooperation will be the treasure of my life.

I would also like to express my deep gratitude to my family for their constant encouragement and support. Their unwavering support has enabled me to overcome every challenge I have encountered along the road of scientific research. I extend my heartfelt thanks to Afzal Dar for his invaluable assistance with my thesis, especially in providing insightful feedback.

Finally, I would like to thank all those who participated and helped me with my research project. Without your help, the successful completion of this paper would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\hat{X}_i$ represents the imputed value for the missing value at position i.

X_j represents the observed values of the feature.

n represents the number of observed values.

IQR represents the interquartile range.

$Q3$ represents the 75th percentile (third quartile).

$Q1$represents the 25th percentile (first quartile).

χ^2 represents the chi-square statistic.

O_{ij} represents the observed frequency of the i_{th} category of the feature and the j_{th} category of the target variable.

E_{ij} represents the expected frequency of the i_{th} category of the feature and the j_{th} category of the target variable, assuming independence.

$X_{\text{normalized}}$ represents the normalized value of the data point.

X represents the original value of the data point.

$X_{\min}$ represents the minimum value of the feature.

$X_{\max}$ represents the maximum value of the feature.

y represents the dependent variable (the variable to be predicted).

$x_1, x_2, \ldots, x_n$ represent the independent variables (features).

$w_0, w_1, w_2, \ldots, w_n$ are the regression coefficients that represent the weights assigned to each independent variable.

ε represents the error term, representing the deviation between the predicted and actual values.

y_i is the target value for the i-th training sample.
$f(x_i)$ is the predicted value for the i-th training sample.

C is the regularization parameter that balances the trade-off between the margin and the error.

$f(x)$ represents the predicted value for the input data point x.

w_i is the weight assigned to the i-th leaf node.

y_i is the predicted value at the i-th leaf node.

$f(x)$ represents the predicted class label for the input data point x.

$I(y_i = y)$ is an indicator function that returns 1 if the predicted class label of the i-th decision tree is equal to y and 0 otherwise.

N is the number of trees in the ensemble.

$f_i(x)$ represents the prediction of the i-th tree for the input x.

m is the number of training samples.

y_i is the actual value of the dependent variable for the i-th training sample.

$\hat{y}$is the predicted value of the dependent variable for the i-th training sample.

k is the number of nearest neighbors to consider.

y_i represents the target value of the i-th nearest neighbor.

$\hat{y}$represents the predicted value.

f is the activation function, which introduces nonlinearity into the model.

w_j are the weights associated with the input features x_j.

b is the bias term.

LIST OF ACRONYMS

CMMOS	Comprehensive Modeling and Monitoring Online System
Chl-a	Chlorophyll-a
MLR	Multivariable Linear Regression
SVM	Support Vector Machine
DT	Decision Tree
RF	Random Forest
GBT	Gradient Boosting Tree
KNN	K Nearest Neighbour
MLP	Multiple Layer Perceptron
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MVI	Missing Value Imputation
OD	Outlier Detection
FS	Feature Selection

Chapter 1 Introduction

1.1 Problem Statement

Chlorophyll-a (Chl-a), a pigment crucial for photosynthesis in algae, serves as a key indicator of algal biomass and harmful algal bloom (HAB) intensity. It is synthesized through a complex biochemical pathway starting with glutamate and involves multiple enzymes, resulting in Chl-a production (Whitton & Potts, 2000). Environmental factors such as nutrient loading, temperature, light intensity, water clarity, and climate change significantly influence its synthesis and accumulation (Deng et al., 2023; Paerl & Huisman, 2020; Zamyadi et al., 2019). HABs, driven by these factors, are expected to intensify globally with climate change and extreme weather events (Whitton & Potts, 2000; Zamyadi et al., 2019). Chl-a concentrations are widely used to gauge HAB intensity in aquatic systems (Protecting Florida Together, n.d.).

The monitoring and modeling of Chl-a concentration in lakes is crucial for assessing the ecological health and water quality of these aquatic systems. Chl-a, a primary pigment in phytoplankton, serves as an indicator of overall productivity and eutrophication levels. Accurate and timely modeling of Chl-a concentration is essential for effective environmental management and informed decision-making (Zhu et al., 2022). However, traditional modeling methods, such as field sampling and laboratory analysis, are labor-intensive, costly, and limited in their ability to capture real-time variations in chlorophyll levels. These methods often produce data that are spatially and temporally sparse, thereby hampering a comprehensive understanding of the dynamics and distribution patterns of Chl-a concentrations in lakes (Park et al., 2020).

The complexity and non-linear nature of the relationships between Chl-a concentration and various environmental factors pose additional modeling and prediction challenges. Traditional statistical models frequently fail to capture the intricate interdependencies and nonlinearities present, resulting in less accurate predictions (Cruz et al., 2021; Tu, 1996). In response to these challenges, there is a critical need for an integrated AI-based online system capable of effectively addressing the shortcomings associated with traditional Chl-a modeling and monitoring methods.

The proposed system will utilize advanced machine learning algorithms—including Multiple Linear Regression, Support Vector Machines, Decision Trees, Random Forests, Gradient Boosting Trees, K-nearest Neighbors, Multiple Layer Perceptron, and Extreme Gradient Boosting—to model the complex relationships between Chl-a concentration and environmental variables. It will feature automated data preprocessing, efficient model training and optimization, real-time modeling and prediction capabilities, and user-friendly visualization tools, thereby enhancing the accuracy and usability of Chl-a predictions (Yang et al., 2022).

Furthermore, the system will incorporate eutrophication risk assessment methodologies to comprehensively evaluate potential environmental impacts associated with varying Chl-a concentrations. This integration will provide a scalable and adaptable framework suitable for various lakes and environmental settings, merging data from field measurements, and other relevant environmental parameters. By overcoming the limitations of traditional methods and leveraging the capabilities of AI, this research aims to significantly advance lake ecosystem research and environmental decision-making processes. It will provide a more effective, efficient, and sustainable approach to managing lake ecosystems globally.

1.2 Objectives of the Study

The main objective of the thesis is to develop an AI based Chl-a modelling and monitoring system. The system is capable of data preprocessing, data management, eight different machine learning models, eutrophication risk evaluation, and Chl-a quality prediction, and the results will be displayed in the form of contour maps and colour-coded maps with spatial information. Specifically, this thesis has the following objectives:

(1) To develop an AI-based Chl-a modelling and monitoring system that integrates an online user interface to visualize the whole process, data management including preprocessing algorithm and machine learning modelling and validation, and database to save model and prediction results. Two case studies are conducted. One is the prediction of Chl-a concentration in Lake Champlain, and the other is the Chl-a concentration in Lake Simcoe.

(2) To develop four different data preprocessing algorithms to deal with the general issues in real-

time datasets. These algorithms include missing value imputation, feature selection, data normalization and outlier detection.

(3) To develop eight different regression models for a group of lake data to complete the task of model testing, model training, model fine-tuning, and predicting the concentration of Chl-a. These models include multiple linear regression (MLR), random forest model (RF), decision tree model (DT), gradient boosting tree model (GBT), k-nearest neighbours regression model (KNN), multiple layer perceptron model (MLP), support vector machine model (SVM), long shot term memory model (LSTM) and Extreme Gradient Boosting model (XGBOOST).

(4) The eutrophication risk assessment will be applied in each case to compare and validate the results of each model. A map will be applied to visualize the results of each case with each model.

1.3 Organization of the Thesis

This thesis is organized into six chapters to provide a structured presentation of the research on developing an integrated AI-based online system for lake Chl-a quality modelling and monitoring. Each chapter focuses on specific aspects of the study and contributes to the overall understanding of the system and its implementation.

Chapter 1 presents the problem statement, highlighting the significance of developing an AI-based system for Chl-a modelling and monitoring. The study's objectives are outlined to provide a clear direction for the research. Additionally, the organization of the thesis is described, providing a roadmap for the subsequent chapters.

Chapter 2 comprehensively reviews the literature related to Chl-a prediction, machine learning techniques, eutrophication risk assessment, and integration with geographical information online systems. This review serves as a foundation for understanding the background, current state of research, and relevant methodologies in the field.

Chapter 3 details the study's methodology. It encompasses an overview of the developed system, data preprocessing techniques, various machine learning models employed, model performance evaluation methods, and eutrophication risk assessment. Thus, the chapter provides

a comprehensive understanding of the techniques and approaches utilized in the research.

Chapter 4 presents a specific case study conducted on Lake Champlain. It includes details about the study area, data sources, system implementation, obtained results, and subsequent discussions. The case study provides empirical evidence of the system's effectiveness in predicting Chl-a concentrations and assessing eutrophication risks in Lake Champlain.

Chapter 5 focuses on a case study conducted on Lake Simcoe. It describes the study area, data sources, system implementation, results obtained, and the corresponding discussions. The case study enables the evaluation of the system's performance and applicability to different lake environments.

The final chapter summarizes the key findings of the research and presents the conclusions derived from the study. It highlights the contributions of the developed AI-based system for lake Chl-a quality modelling and monitoring. Furthermore, recommendations for future work are provided to guide further advancements and enhancements in the field.

Chapter 2 Literature Review

2.1 Chl-a Prediction

Accurate prediction of Chl-a concentration in lakes is of great importance for understanding these aquatic ecosystems' ecological health and water quality (Zhu et al., 2022). Chl-a is a primary pigment in phytoplankton and is closely associated with the overall productivity and eutrophication levels in lakes. Monitoring and modelling Chl-a concentration provides valuable insights into phytoplankton populations' dynamics and spatial distribution and the potential environmental impacts associated with changes in Chl-a levels (Kalff and Knoechel, 1978). Traditional approaches for Chl-a prediction mainly rely on field sampling and laboratory analysis, which are labour-intensive, time-consuming, and costly (Park et al. 2020). These methods often provide limited spatial and temporal coverage, making it challenging to capture the complex variations and patterns of Chl-a concentration in lakes (Stumpf et al. 2016). Moreover, the non-linear and intricate relationships between Chl-a and environmental factors further complicate traditional statistical models' prediction process (Cruz et al. 2021).

The prediction of Chl-a levels using AI models has become a significant area of research due to the crucial role of Chl-a in indicating water quality and algal bloom potential. Machine learning techniques have emerged as a valuable tool for Chl-a prediction in recent years. These techniques leverage the power of data-driven algorithms to establish complex relationships between Chl-a concentrations and environmental variables (Tahmasebi et al. 2020). MLR, SVM, DT, RF, GBT, KNN, MLP, LSTM and XGBoost are among the machine learning models commonly employed for Chl-a prediction (Sundararajan et al. 2021). These models allow for the incorporation of various water quality features, such as water temperature, Secchi depth, total phosphorus, and total nitrogen, in the prediction process, enabling a more comprehensive understanding of the factors influencing Chl-a concentrations in lakes (Tung and Yaseen, 2020). Machine learning models such as Support Vector Regression, Bagging, Random Forest, Extreme Gradient Boosting, RNN, and LSTM have been successfully used to predict Chl-a levels, demonstrating the importance of selecting appropriate explanatory variables and recursive prediction methods to enhance model accuracy (Shin et al., 2020). Studies have shown that models

like Extreme Gradient Boostings (XGBoost) combined with genetic algorithms provide effective predictions and optimize conditions for processes like electro-oxidation, impacting water treatment and quality (Picos et al., 2018). The use of RF and feature selection methods can significantly improve the predictive ability of models for Chl-a concentration, highlighting the model's utility in environmental management (Li et al., 2018). Transfer learning has been applied to optimize neural network models for Chl-a prediction, enhancing the model's generalization ability and maintaining high performance in long-term applications (Tian et al., 2019).

AI models, particularly those incorporating machine learning and neural networks, have shown great promise in accurately predicting Chl-a levels, which is vital for modelling water quality and managing eutrophication in aquatic systems. These models not only offer precise predictions but also aid in the optimization and management of environmental conditions.

2.1.1 Machine Learning Techniques

This section discusses various machine learning techniques utilized in Chl-a prediction. Each technique, including MLR, SVM, DT, RF, GBT, KNN, MLP, and Extreme Gradient Boosting

(XGBOOST), is described in detail. The strengths and weaknesses of each technique in capturing the complex relationships between Chl-a concentration and environmental factors are analyzed.

2.1.2 Multiple Linear Regression (MLR)

Chl-a is a crucial indicator of water quality and phytoplankton biomass in aquatic environments. MLR is a widely used statistical technique for modelling the relationship between dependent and multiple independent variables. In the context of Chl-a prediction, MLR is often employed to establish a quantitative relationship between Chl-a concentration and various environmental factors (Çamdevýren et al. 2005). One of the critical advantages of MLR is its simplicity and interpretability. The regression coefficients obtained from MLR provide insights into the magnitude and direction of the relationship between each independent variable and Chl-a concentration. It also identifies potential collinearity among independent variables (Coops et al., 2003).

Predicting its concentrations using MLR models has been a focal point in environmental research, given the simplicity and interpretability of MLR. This review summarizes the applications of MLR models in various aquatic settings to predict Chl-a concentrations, incorporating findings from several studies. MLR assumes a linear relationship between the dependent and independent variables, which may not always hold in the case of Chl-a prediction (Filstrup et al. 2014). If the relationship is nonlinear, MLR may provide less accurate predictions. Many cases prove that more advanced machine learning techniques, such as decision trees or neural networks, are more suitable for modelling nonlinear relationships (Mamun et al., 2019; Wei et al., 2019; Rajaee and Boroumand, 2015).

Hybrid approaches combining MLR with Extreme Gradient Boostings (XGBoost) have shown enhanced prediction capabilities compared to standalone MLR models due to reduced errors and increased correlation coefficients, suggesting a synergistic benefit in complex water systems like the offshore Kuala Terengganu, Terengganu, Malaysia (Lola et al., 2016). Integrating principal component analysis (PCA) with MLR, known as PCS-MLR, has effectively improved predictive success by reducing collinearity among variables in marine ecosystems, aiding in better

management of coastal waters (Franklin et al., 2020). MLR models, despite their simplicity, have effectively predicted Chl-a concentrations with satisfactory accuracy when appropriate environmental variables were selected, as demonstrated in the South San Francisco Bay study (Rajaee & Boroumand, 2015). The predictive performance of MLR models has also been evaluated using high-resolution Landsat imagery to estimate Chl-a in water bodies, highlighting the utility of remote sensing data in enhancing MLR predictions for smaller and coastal water bodies (Matus-Hernández et al., 2018). MLR has been successfully applied in reservoirs, where principal component scores significantly improved the predictive accuracy, as evidenced by studies in Turkey demonstrating how MLR can be utilized effectively in freshwater environments (Çamdevýren et al., 2005). Studies like those conducted on the Yeongsan Reservoir, Korea, have shown how parameter optimization can significantly improve the predictions of MLR models, ensuring greater reliability in forecasting Chl-a levels in dynamic and nutrient-rich water bodies (Cho et al., 2009).

MLR models remain a robust tool for predicting Chl-a concentrations in diverse aquatic environments. Their effectiveness is enhanced when combined with other computational tools or when used in conjunction with robust variable selection methods, making them invaluable in water quality management and ecological modelling efforts.

2.1.3 Support Vector Machines (SVM)

SVM is a robust supervised learning algorithm widely used in Chl-a prediction for lake water quality modelling. Researchers have applied SVM as a regression model to establish a nonlinear relationship between environmental variables and Chl-a concentrations. It has gained prominence in predicting Chl-a concentrations in various aquatic environments. Their ability to handle non-linear and high-dimensional data makes them particularly useful for modeling complex environmental phenomena such as algal blooms. This review compiles significant research employing SVM for Chl-a predictions, reflecting on methodologies, advancements, and outcomes.

Liu found SVM could perform better than neural networks in Taihu Lake in China (Liu et al., 2009). Park demonstrated the effectiveness of the support vector machine (SVM) model in predicting Chl-a concentrations for early warning in freshwater and estuarine reservoirs (Park et

al. 2015). Shin investigated the role of Support Vector Regression and other machine learning models in predicting Chl-a concentrations in the Nakdong River, Korea (Shin et al., 2020). SVM offers flexibility in handling nonlinear relationships and can capture complex patterns and dependencies in the data (Ifenthaler and Widanapathirana, 2014; Stanimirova et al., 2010). By utilizing kernel functions, SVM can map data to a higher-dimensional space, enabling the capture of intricate relationships that linear regression models may overlook (Otchere et al., 2021). SVM is also advantageous for handling small sample sizes and noisy data, as it is less prone to overfitting than other algorithms. Its margin maximization principal aids in generalizing new data well, even with limited training sets (Wujek et al., 2016).

SVM models optimized with genetic algorithms have shown superior performance in predicting Chl-a concentrations, demonstrating higher accuracy than traditional models in various reservoirs and lakes (Hua-jun & Defu, 2009). Studies have utilized SVM for early warning systems in reservoirs, where they outperformed Extreme Gradient Boostings due to better handling of non-linear relationships between environmental variables and Chl-a concentrations (Park et al., 2015). The combination of remote sensing and SVM has proven effective in estimating Chl-a from Landsat 8 OLI images, with models achieving substantial accuracy across different seasons, thereby facilitating broader application in water quality modelling (Zhang, Huang, & Wang, 2020). A hybrid approach using SVM and genetic algorithms significantly improved prediction accuracy by optimizing feature selection, demonstrating the model's robustness in complex systems like the Miyun Reservoir in China (Su et al., 2015). The implementation of SVM with support vector regression (SVR) has been noted for its high predictive ability, especially using radial basis function (RBF) kernels, which provide a more nuanced interpretation of environmental data inputs (Guang-ren, 2012). Incorporating SVM with particle swarm optimization (PSO) has enhanced the predictive accuracy of Chl-a concentration models, reflecting the model's capacity to adapt to different input variables and complex ecological dynamics (Nieto et al., 2016). Remote sensing data coupled with SVM has been effective in producing real-time, reliable Chl-a predictions in lake systems, showing that SVM can handle the variability introduced by changing seasonal factors (Wu et al., 2023).

However, SVM does have limitations. Selecting the appropriate kernel function and tuning hyperparameters significantly impacts model performance. Careful parameter tuning and model selection are necessary for accurate predictions (Demir and Şahin, 2022). Additionally, SVM can be computationally intensive, especially with large datasets, as it involves solving a quadratic optimization problem during training (Cervantes et al., 2020). Despite these limitations, SVM has shown promising performance in Chl-a prediction and has been successfully utilized in studies modelling the relationship between environmental factors and Chl-a concentrations in lakes.

SVM models are a powerful tool for predicting Chl-a levels in diverse aquatic environments, from small lakes to large reservoirs. Their ability to integrate with various optimization algorithms and handle multi-dimensional data efficiently makes them indispensable for water quality modelling and management strategies. The adaptation of SVM models to incorporate real-time data and advanced feature selection algorithms further enhances their applicability and accuracy, making them a preferred choice for environmental scientists and managers.

2.1.4 Random Forests (RF)

Predicting Chl-a concentrations is vital for managing water quality and monitoring ecological health in aquatic environments. The RF model, a powerful ensemble learning technique that utilizes multiple decision trees, has been widely adopted for this task due to its robustness, ease of use, and ability to handle complex datasets with high accuracy. This review compiles recent studies that demonstrate the effectiveness of RF models in predicting Chl-a levels across various water bodies. Li et al. suggested that the RF model is valuable for determining significant stressors and accurately predicting Chl-a concentration in a shallow lake (Li et al., 2018). Shen et al. found that the RF model demonstrated robustness and reliability in accurately estimating Chl-a concentrations in optically complex waters, overcoming the uncertainties in atmospheric correction (Shen et al., 2022). Huang et al. discovered the effectiveness of the Random Forest model in predicting Chl-a concentrations in Chinese lakes using data from various databases (Huang et al., 2022).

One advantage of RF is its ability to capture complex patterns and interactions in the data. By constructing decision trees using random subsets of training data and input features, RF can model intricate relationships, enhancing predictive accuracy compared to single decision tree models (Ahmad et al., 2017; Shen et al., 2022). RF is also robust in overfitting and can handle noisy data due to bootstrapping and random feature selection during tree construction (Fox et al., 2017). Moreover, it measures feature importance, identifying vital environmental factors influencing Chl-a concentrations. This information offers insights into the underlying processes governing Chl-a dynamics in lakes (Ly et al., 2021). RF models have been successfully applied to predict Chl-a concentrations in fresh and brackish waters in Japan, showing the model's adaptability to different water chemistries and environmental settings (Yajima & Derot, 2018). In the study of plant leaves, RF was utilized to predict chlorophyll content from reflection spectra, indicating its potential beyond aquatic applications (Urbanovich, Afonnikov, & Nikolaev, 2021). An integrated RF approach was applied to coastal water management, where it was used to predict seasonal Chl-a variations with high precision, demonstrating the model's effectiveness in handling seasonal data shifts (Jia, Cheng, & Hu, 2020). RF combined with feature selection techniques has significantly enhanced the prediction of Chl-a in various studies, proving the importance of selecting relevant environmental predictors (Li, Sha, & Wang, 2018). The versatility of RF models was showcased in their application to warning systems for water blooms, where they were found to be more accurate than other machine learning methods (Liu & Wu, 2017). RF models have been extended to predict other related parameters, such as water bloom events and nutrient levels, further underscoring their utility in comprehensive water quality management (Hollister, Milstead, & Kreakie, 2016).

However, RF models are a robust and versatile tool for predicting Chl-a concentrations, capable of handling complex and heterogeneous data while providing high accuracy and reliability. Careful tuning of hyperparameters is required to prevent overfitting, considering factors such as the number of trees and maximum tree depth (Oyedele et al., 2021; Li et al., 2018). Additionally, RF models may have limited interpretability compared to simpler models like linear regression. Nevertheless, the feature importance measure helps understand variable importance in the model (Chen et al., 2020). Their ability to integrate various data types and perform feature selection

makes them an indispensable tool in the field of environmental modelling and water quality management. Despite these considerations, RF has demonstrated promising results in Chl-a prediction and is widely used in the field. Various studies have successfully demonstrated its effectiveness in capturing complex relationships and providing accurate predictions.

2.1.5 Decision Tree (DT)

DT models have been widely used in predicting Chl-a concentrations, a key indicator of water quality and algal biomass in aquatic systems. These models are favored for their simplicity, interpretability, and effectiveness in handling complex ecological data. This review aggregates research findings from various studies employing DT models for Chl-a prediction across different environmental settings.

DT model constructs a hierarchical tree structure based on input features to make decisions (Zhu et al., 2022). DT excels at handling nonlinear relationships and interactions among variables by recursively splitting the data. It captures complex decision boundaries and reveals patterns in the data (Kotsiantis, 2013). The interpretability of DT allows researchers to gain insights into the influential factors and understand the decision-making process. A c-fuzzy model tree, integrating fuzzy clustering with a DT approach, demonstrated superior performance in predicting Chl-a concentrations, highlighting the potential of hybrid models in enhancing prediction accuracy (Lee et al., 2006). Ensemble learning algorithms, including DT methods, have been effectively used to predict cyanobacterial blooms, which are closely related to Chl-a levels in the lower Han River, South Korea. These methods outperformed single DT models, indicating the strength of ensemble approaches (Shin, Yoon, & Cha, 2017). DT models have also been applied in predicting gene functions and could be adapted to predict Chl-a by managing complex biological and environmental datasets (Schietgat et al., 2010). Research on improved prediction models based on DT for educational data emphasized the versatility of DTs in various predictive contexts, which could be extended to environmental data (Yang, Chen, & Zhang, 2022). DT methods were used to develop early-warning systems for predicting Chl-a concentration in Korean reservoirs, showcasing their utility in real-time water quality modelling (Park et al., 2015). In a broader context, DTs have been applied to model ecological dynamics, including the prediction of habitat

suitability and species distribution, which are indirectly related to Chl-a levels as they affect and are affected by aquatic ecosystems (Debeljak & Džeroski, 2011). Additionally, DT can handle numerical and categorical data without extensive preprocessing, simplifying the modelling process (Merghadi et al., 2020).

However, DT has limitations. It tends to overfit the training data when the tree depth is unconstrained. Techniques like pruning or setting a maximum depth can be applied to mitigate overfitting. Another limitation is its instability in small changes in the training data. To address this, ensemble methods like RF combine multiple decision trees to improve predictive performance and reduce variance (Fratello and Tagliaferri, 2018). Zounemat-Kermani et al. also reviewed that ensemble methods like gradient boosting trees perform better than single decision trees (Zounemat-Kermani et al., 2021). Nevertheless, DT has demonstrated promising results in Chl-a prediction. Various studies have successfully employed it, providing valuable insights into the relationships between environmental factors and Chl-a dynamics in lakes (Li et al., 2018; Barzegar et al., 2020; Liu and Wu, 2017).

In all, decision tree models are a robust tool for predicting Chl-a concentrations, with applications ranging from water quality modelling to ecological modeling. Their ability to handle diverse datasets and integrate with other machine learning techniques makes them particularly effective in environmental science. The use of DT models in conjunction with ensemble methods and hybrid approaches can significantly enhance predictive performance, supporting more accurate and timely environmental management decisions.

2.1.6 Gradient Boosting Trees (GBT)

GBT is a versatile machine learning technique that has been effectively utilized for predictive modeling in various domains, including environmental science. Specifically, GBT has shown significant potential in predicting Chl-a concentrations, a crucial metric for assessing water quality and algal biomass in aquatic ecosystem. This review compiles findings from several studies that explore the application of GBT models for Chl-a prediction.

GBT combines multiple weak prediction models, typically decision trees, to create a strong predictive model. It captures complex relationships and interactions among independent variables

(Haggerty, 2023). Yao discovered the effectiveness of the Gradient Boosting Tree (GBT) model in Chl-a prediction by employing it to estimate Chl-a concentrations in coastal waters using Landsat 8 OLI image data and field measurements (Yao et al., 2021). Kim conducted a study focusing on predicting the concentration of Chl-a (Chl-a) in seawater using a gradient-boosting tree model (Kim et al., 2022). Moreover, Hu investigated the role of the Gradient Boosting Tree (GBT) model in Chl-a prediction and found that it showed excellent fitting ability compared to other machine-learning models, providing a reliable prediction method for eutrophication based on monthly modelling data (Hu et al., 2021). GBT models have been effectively applied to predict Chl-a concentrations in various aquatic environments, demonstrating robust performance and adaptability to different data characteristics and environmental conditions (Zhang & Haghani, 2015). Research indicates that integrating GBT with other models like LSTM enhances the predictive accuracy for dissolved oxygen, which is closely related to Chl-a concentrations in aquaculture settings (Huan et al., 2020). GBT has been employed to develop models for Chl-a estimation from satellite imagery, showing significant promise in remote sensing applications for water quality modelling (Cao et al., 2020). Studies have also explored the use of GBT for modeling Chl-a dynamics in complex and turbid water bodies, where the model's ability to handle non-linear relationships and high-dimensional data was particularly beneficial (Mustapha & Saeed, 2016). The predictive performance of GBT models has been further enhanced by incorporating advanced feature selection techniques, leading to more accurate predictions of Chl-a levels (Salditt et al., 2023).

GBT iteratively builds an ensemble of decision trees, correcting errors made by previous trees. This sequential approach enables GBT to capture intricate patterns and dependencies in the data, improving prediction accuracy (Smith et al., 2021; Haggerty et al., 2023). Besides, GBT provides interpretability by assigning weights to each tree, allowing researchers to identify critical variables influencing Chl-a concentrations (Park et al., 2022). Moreover, GBT is also robust in dealing with outliers and noisy data. It assigns lower weights to outliers, reducing their impact on predictions (Yao et al., 2021). However, GBT requires careful consideration. Overfitting can occur if the number of trees in the ensemble is excessive or model complexity is not controlled (Bentéjac et al., 2021). Parameter tuning, such as the learning rate and maximum depth, is necessary to

prevent overfitting and optimize performance. Training GBT can be computationally intensive, especially for large datasets and complex models (Zennaro et al., 2018). Nevertheless, advancements in computing power and optimization algorithms have improved training efficiency.

GBT models represent a powerful tool for predicting Chl-a concentrations, offering high accuracy and robustness across diverse aquatic environments. Their ability to integrate complex data sets and handle non-linear relationships makes them particularly effective for environmental modelling and management. It has demonstrated promising results in Chl-a prediction, successfully modelling and predicting lake concentrations. Its ability to capture complex relationships and provide accurate predictions makes it a valuable tool in this domain.

2.1.7 K-nearest Neighbors (KNN)

The KNN algorithm, a simple yet robust machine learning method, has been extensively utilized across various domains, including environmental modelling and ecological forecasting. This review explores the application of the KNN model specifically for predicting Chl-a concentrations in aquatic systems, a critical indicator of water quality and phytoplankton biomass.

KNN classifies or predicts new data points based on their proximity to the training data. KNN's simplicity and lack of assumptions about data distribution make it easy to understand and implement (Ray 2019). Its main advantage lies in its ability to capture local patterns by considering data points with similar features and their proximity. This makes it robust to outliers and noise in the data. It can handle numerical and categorical data without extensive preprocessing (Alexandropoulos et al., 2019). However, there are considerations to address. The number of nearest neighbours is essential for optimal performance when choosing the appropriate value for K. A small K increases sensitivity to noise, while a large K over smoothies diminishes local patterns (Tjärnberg, 2021). KNN models have been successful in predicting water quality parameters by leveraging spatial and temporal data, demonstrating their versatility and effectiveness in environmental applications (Tomppo et al., 2009). The adaptability of KNN in predicting Chl-a concentrations has been highlighted through its ability to integrate various types of environmental data, making it a reliable choice for researchers and practitioners (Parry et al., 2010). Studies have shown that the KNN model, when optimized with feature selection techniques,

can significantly enhance prediction accuracy for Chl-a, underlining the importance of methodological fine-tuning in environmental modeling (Nigsch et al., 2006). The flexibility of KNN is evident in its application across different ecological and biological datasets, proving its efficacy in scenarios with complex data interactions and where traditional modeling techniques might fall short (Mubarok et al., 2023). Moreover, the computational cost of KNN can be high with large datasets. Calculating distances between the query point and all training data points is time-consuming. Efficient data structures and algorithms, such as kd-trees or ball trees, can speed up the nearest-neighbour search (Jia et al., 2020).

KNN models are a potent tool for predicting Chl-a levels in aquatic systems, offering robust predictions that are crucial for water quality management and ecological research. The simplicity in its application, coupled with its ability to produce accurate predictions, makes KNN a preferred method in environmental science. Despite these considerations, KNN has shown promising results in Chl-a prediction. It has been successfully applied in various studies, especially when local patterns and neighbourhood information are crucial for accurate predictions.

2.1.8 Multiple Layer Perceptron (MLP)

The MLP, a type of Extreme Gradient Boosting, has been increasingly applied in the prediction of Chl-a concentrations in various aquatic systems. This review synthesizes findings from several studies employing MLP models, highlighting their effectiveness and adaptability in modeling complex ecological data for Chl-a prediction.

MLP consists of interconnected layers of neurons, enabling it to capture complex relationships between environmental variables and Chl-a concentrations (Golhani et al., 2018). A study conducted by Rybka demonstrated the efficacy of the MLP model in predicting Water Saturation Deficit (WSD) values based on Chl-a fluorescence parameters. The MLP model achieved a precision of 82% and a correlation coefficient of 0.98, indicating its potential for developing a new screening test for plant tolerance to temporary water shortages (Rybka et al., 2019). Jeong discovered that the MLP model exhibited superior predictability and outperformed statistical regression models in predicting Chl-a concentrations in a regulated river ecosystem (Jeong et al., 2006). As the advantages of the model, MLP excels at capturing non-linear relationships and complex patterns in the data. Its hidden layers allow it to learn intricate dependencies between input features and the target variable (Rajaee and Boroumand, 2015; Jeong et al.,

2006; Rybka et al., 2019). MLP can approximate complex functions and capture the non-linearities inherent in Chl-a prediction by employing non-linear activation functions and adjusting weights during training. It can handle numerical and categorical data without extensive preprocessing, automatically extracting relevant features from the raw input data (Talib, 2006).

MLP models have been successfully used to predict Chl-a concentrations in rivers, showing high accuracy and the ability to handle non-linear relationships in environmental data (Shin et al., 2020). Enhanced MLP models incorporating genetic algorithms have demonstrated improved prediction accuracy by optimizing the weighting coefficients in the neural network structure (Altunkaynak, 2013). Studies using MLP with environmental variables have successfully modeled Chl-a dynamics in coastal waters, emphasizing the significance of feature selection and model training methods (Jia et al., 2020). The integration of MLP models with other machine learning techniques has shown potential in enhancing the predictive performance for Chl-a, especially in complex aquatic environments (Moustafa & Elsheikh, 2023). However, several studies concluded that it is essential to consider various factors when using the MLP model. Determining the optimal network architecture, including the number of hidden layers and neurons, is crucial and depends on the problem's complexity and dataset size (Karsoliya, 2012). Overfitting is a concern, especially with limited training data, but regularization techniques like dropout and L2 regularization can mitigate this issue (Phaisangittisagul, 2016). Training MLP can be computationally intensive, but hardware advancements and optimization algorithms have improved efficiency (Li et al., 2016).

MLP models are a robust tool for predicting Chl-a concentrations, capable of processing complex and non-linear ecological data effectively. Their adaptability to various environmental contexts and integration with optimization algorithms make MLP a valuable model in water quality modelling and management. Despite these considerations, MLP has demonstrated promising results in Chl-a prediction. It has been successfully employed in various studies to model and predict Chl-a concentrations in lakes, showcasing its ability to capture complex relationships and provide accurate predictions.

2.1.9 Long Short Term Memory Network (LSTM)

Long Short-Term Memory Networks (LSTM) have become a critical tool in time series

prediction due to their ability to remember information for extended periods. In the context of environmental sciences, LSTM models have been extensively utilized for predicting Chl-a concentrations, a key indicator of algal biomass and water quality in aquatic systems. This review synthesizes findings from several studies that demonstrate the effectiveness of LSTM models in this domain.

LSTMs have proven effective in predicting Chl-a concentrations in rivers, leveraging their ability to model time-series data dynamically (Shin et al., 2020). In coastal waters, LSTMs have been utilized to forecast seasonal variations of Chl-a, showing a superior performance in capturing complex seasonal patterns (Chen & Xu, 2020). The integration of LSTM networks with other machine learning techniques, such as genetic algorithms, has further enhanced the accuracy of Chl-a predictions, indicating the robustness of hybrid approaches (Fan, Xiao, & Dong, 2020). LSTM networks have also been applied successfully in modeling the nutrient removal processes in sewage treatment plants, which correlates with Chl-a dynamics, demonstrating the broad applicability of LSTM in environmental systems modeling (Yaqub et al., 2020). Research has explored the potential of LSTMs in predicting Chl-a from high-dimensional chaotic systems, showcasing the model's capability in dealing with complex ecological data sets (Liang et al., 2020).

While LSTM networks are powerful tools for modeling time series data, including Chl-a concentration predictions, they also present several challenges and limitations. This review identifies key disadvantages associated with the use of LSTMs in predicting Chl-a, synthesizing insights from recent research. LSTMs are complex models that can easily overfit, especially when training on datasets with limited temporal variability or when the data do not have strong temporal dependencies (Gers, Schmidhuber, & Cummins, 2000). The training process for LSTMs is computationally expensive due to their complex architectures, which can be a significant drawback in real-time prediction scenarios or when using limited computational resources (Shin et al., 2020). LSTMs require careful tuning of parameters and extensive training to capture the dynamics of Chl-a accurately, which can be time-consuming and requires substantial expertise (Chung & Shin, 2018).

LSTM networks are a valuable tool for predicting Chl-a concentrations, offering significant

advantages in terms of learning complex temporal patterns and handling large datasets. Their flexibility and efficiency make them a promising approach for ongoing and future applications in water quality modelling. While LSTM networks offer robust capabilities for modeling time-series data such as Chl-a concentrations, their practical application is hindered by issues related to model complexity, computational demands, training difficulties, sensitivity to hyperparameters, and substantial data needs. Addressing these challenges is crucial for improving the usability of LSTMs in environmental modeling and prediction.

2.1.10 Extreme Gradient Boosting (XGBoost)

XGBoost is a powerful, scalable machine learning algorithm that has been extensively applied in various domains, including ecological modeling and prediction of Chl-a levels in aquatic environments. In Chl-a prediction, Extreme Gradient Boosting has handled nonlinear relationships and complex interactions between environmental factors and Chl-a concentrations. This review synthesizes the application of XGBoost for predicting Chl-a, a critical indicator of algae biomass and water quality.

XGBoost has demonstrated high accuracy in predicting Chl-a concentrations, leveraging its robust handling of non-linear relationships and large datasets. Its performance in predicting Chl-a has also been enhanced by feature selection techniques that help in identifying the most influential variables, thereby optimizing the prediction models for better accuracy. This is particularly evident in studies conducted on rivers and coastal areas where traditional models often fall short (Shin et al., 2020). The application of XGBoost in marine environments shows its capability to integrate and analyze complex environmental data sets, resulting in highly accurate Chl-a predictions. The versatility of XGBoost is demonstrated in its ability to integrate various types of environmental data to improve the accuracy of Chl-a predictions. This integration is critical in settings where data are derived from complex aquatic systems This is supported by integrated approaches that enhance feature selection and model robustness (Jia et al., 2020). XGBoost's effectiveness is also apparent in its application across different ecosystems, including both freshwater and marine environments, where it helps in predicting algal blooms and assessing nutrient cycles, which are closely linked to Chl-a levels (Kim et al., 2022). Further, the algorithm's

flexibility and scalability make it suitable for real-time modelling systems, offering a practical tool for environmental scientists and policy makers to manage water quality effectively (Li et al., 2019).

However, like all models, XGBoost has limitations, especially when applied to specific tasks such as Chl-a prediction in aquatic environments. XGBoost can easily overfit especially when the hyper-parameters are not correctly tuned. This is a common issue in complex models where there are many features relative to the number of observations (Shin et al., 2020). Despite its effectiveness, XGBoost often acts as a black box, making it difficult to interpret the model's decisions. This is particularly problematic in scientific fields where understanding the model's decision-making process is crucial (Li, Yin, & Quan, 2019). XGBoost requires high computational resources, especially when handling large datasets or performing grid search for hyper-parameter tuning. This can be a limitation in scenarios with restricted computational resources (Budholiya, Shrivastava, & Sharma, 2020). The performance of the XGBoost model heavily depends on the tuning of its parameters. Finding the right set of parameters can be time-consuming and requires a deep understanding of how the parameters interact with each other and the data (Moore & Bell, 2022).

Overall, while XGBoost is a powerful and efficient algorithm, it presents challenges such as overfitting, difficulty in interpretation, and computational demands, especially when applied to Chl-a prediction in aquatic environments. Their combined capabilities allow for a more nuanced understanding and prediction of Chl-a dynamics in various aquatic environments. It is a potent and versatile tool for Chl-a prediction, capable of processing complex and diverse datasets efficiently. Its application in aquatic environments helps in precise water quality modelling and management, providing a reliable method for environmental scientists to predict and address ecological challenges.

2.2 Eutrophication Risk Assessment

Eutrophication is a widespread environmental issue in many lakes worldwide, primarily caused by excessive nutrient inputs, particularly nitrogen and phosphorus (Khan and Mohammad, 2014). It leads to the overgrowth of algae and other aquatic plants, reducing water clarity, oxygen

depletion, and negative impacts on marine organisms (Chislock et al., 2013). Eutrophication poses a significant threat to lakes' ecological health and water quality, making its assessment crucial for effective management and conservation strategies (Oliver et al. 2019). Eutrophication risk assessment provides a framework for evaluating the potential impacts of nutrient enrichment on lake ecosystems (Smith and Schindler, 2009). It involves analyzing various indicators and metrics to assess the degree of eutrophication and identify vulnerable areas. By understanding current and potential future eutrophication trends, decision-makers can implement targeted mitigation measures and adaptive management strategies.

Multiple ecological models, including Fuzzy Logic, Recurrent Extreme Gradient Boosting, and Hybrid Evolutionary Algorithm, have been assessed for predicting Chl-a in tropical lakes. These models help in understanding algal biomass as an indicator of the trophic status, crucial for managing eutrophication (Malek et al., 2011). Spatial analysis of catchment variables has shown that Random Forest models can effectively predict Chl-a concentrations, serving as a proxy for eutrophication status in lakes and reservoirs. This highlights the importance of catchment characteristics in eutrophication risk assessment (Catherine et al., 2010). Regression models relating nutrient levels and water renewal rates to Chl-a levels have been developed for coastal embayments, aiding in the assessment of eutrophication risk and potential management strategies (Arhonditsis et al., 2003). Early-warning protocols using machine learning models, including Neural Networks and Support Vector Machines, have been developed for predicting Chl-a concentration in reservoirs, enhancing eutrophication management schemes (Park et al., 2015). Improved algorithms using MODIS imagery data have been utilized to estimate Chl-a concentrations, aiding in modelling eutrophication processes in tropical coastal waters (Ha et al., 2013). Hybrid models combining Least Squares Support Vector Regression and Radial Basis Function Neural Networks have shown high accuracy in predicting Chl-a content, crucial for evaluating eutrophication in reservoirs (Wang et al., 2016). Overall, eutrophication risk assessment provides a comprehensive framework for evaluating the potential environmental impacts of nutrient enrichment in lakes. The combination of nutrient concentration measurements, trophic state indices, ecological status assessments, modelling approaches, and geospatial analysis allows for a holistic evaluation of eutrophication risks (Kitsiou and Karydis, 2011). This knowledge serves

as a basis for implementing targeted management strategies to mitigate eutrophication and restore the ecological balance of lake ecosystems.

Nutrient concentration measurements, such as total nitrogen (TN) and total phosphorus (TP), are commonly used indicators for assessing eutrophication. These indicators provide insights into the nutrient status of a lake and help identify areas with excessive nutrient loads (Winter and Duthie, 2000). Additionally, trophic state indices, such as the Carlson Trophic State Index (TSI) and the Trophic Level Index (TLI), integrate multiple factors, including nutrient concentrations, Chl-a levels, and water clarity, to provide an overall assessment of eutrophication levels (El-Serehy et al. 2018). These indices enable comparisons between different lakes and facilitate the identification of lakes at high risk of eutrophication. In recent years, integrating modelling approaches with eutrophication risk assessment has enhanced the understanding of nutrient dynamics and predicting future eutrophication scenarios (Bhagowati and Ahamad, 2019). Dynamic models, such as the Water Quality Analysis Simulation Program (WASP) and the Environmental Fluid Dynamics Code (EFDC), simulate the transport and fate of nutrients in lakes, allowing for the assessment of the effectiveness of different nutrient reduction strategies. These models enable decision-makers to explore various management scenarios and optimize the allocation of resources for eutrophication control (Burigato et al. 2019).

The Trophic State Index (TSI) is a critical tool used in assessing the nutrient status of aquatic environments, which can help predict the concentration of Chl-a—a key indicator of algal biomass and potential algal blooms. A study introduced a hybrid algorithm for estimating Chl-a across different trophic states, from oligotrophic to hypertrophic conditions. The model used a combination of algorithms designed for clear, turbid, and highly turbid waters, showing good performance across diverse water types (Matsushita et al., 2015). Assessment of Chl-a concentration and TSI in Lake Chagan using Landsat TM and field spectral data revealed the potential of remote sensing in eutrophication studies. The study successfully mapped Chl-a distribution and assessed the trophic state, demonstrating remote sensing's capability in large-scale modelling (Duan et al., 2007). The Normalized Difference Chlorophyll Index (NDCI) was proposed to predict Chl-a in turbid productive waters, showing high correlation coefficients and low mean square error, emphasizing its applicability across geographic regions (Mishra & Mishra,

2012). An empirical study using Landsat 8 OLI data to estimate Chl-a concentrations in East Kolkata Wetland, India, provided strong correlation between laboratory and predicted TSI values, reinforcing the use of satellite imagery in continuous modelling of trophic states (Patra et al., 2016). The Fuzzy BP method was applied to predict Chl-a content in seawater, demonstrating how advanced machine learning techniques can effectively address the complex nonlinear relationships between Chl-a and various environmental factors (Zhang, Li, & Hu, 2011). A novel approach using PCA and MLR for predicting Chl-a in coastal-marine ecosystems was explored, achieving significant predictive success. This method reduced collinearity issues, making it a viable option for understanding and managing coastal environments (Franklin et al., 2020).

The relationship between TSI and Chl-a prediction is well-established, with various methodologies demonstrating effectiveness in different aquatic environments. From hybrid algorithms to advanced remote sensing techniques, these approaches provide essential insights into the trophic dynamics of water bodies, aiding in the management and mitigation of eutrophication. The integration of these methods into regular modelling can significantly enhance our ability to predict and manage water quality in diverse aquatic ecosystems.

2.3 Integration with Modeling and Monitoring Online System

Integrating modeling and monitoring online systems for Chl-a prediction is crucial for managing aquatic ecosystems effectively. Such integrations often leverage advanced predictive models and online modelling systems to provide real-time or near-real-time data for early warning and management of eutrophication and algal blooms. This review discusses the various approaches and technologies used in the integration of these systems for Chl-a prediction.

XGBoost models optimized for predicting chlorophyll dynamics have shown improved performance by focusing on changes in chlorophyll value rather than absolute values, enhancing bloom forecasting accuracy and reducing in-situ modelling costs (Tian, Liao, & Zhang, 2017). Integrating satellite-derived chlorophyll data into ensemble simulations for the North Atlantic Ocean demonstrates how data assimilation can enhance surface analysis and chlorophyll forecast accuracy, although improvements depend on the reliability of the ensemble (Santana-Falcón et al.,

2020). The use of Auto-Regressive Integrated Moving Average (ARIMA) models for online forecasting of Chl-a concentrations has shown promise in freshwater systems, providing a practical tool for algal bloom early warning systems (Chen, Guan, Yun, Li, & Recknagel, 2015). Hybrid models combining simulation and deep learning-based prediction methods have been developed to predict Chl-a concentration at non-modelling spots, demonstrating effective integration of data-driven and model-based approaches (Jang et al., 2020). Assimilating SeaWiFS chlorophyll data into a 3D-coupled physical-biogeochemical model highlights the importance of refining model uncertainties according to regional biogeochemical characteristics, improving predictions in coastal zones (Fontana et al., 2009). Multistep-ahead forecasting using wavelet nonlinear autoregressive networks (WNARNet) for Chl-a emphasizes the effectiveness of advanced computational methods in handling complex time series data, thus aiding in more accurate and extended forecasting (Du et al., 2018). Geographical information online systems further enhance the accessibility and usability of Chl-a modelling and monitoring data. These systems allow users to access and visualize Chl-a data through web-based interfaces, making the information readily available to various stakeholders, including researchers, policymakers, and the general public. Online systems can provide interactive tools for querying and analyzing Chl-a data, enabling users to explore spatial patterns, generate custom maps, and extract relevant information.

Integrating modeling and monitoring systems for predicting Chl-a concentration is a crucial step in managing aquatic ecosystems effectively. However, several limitations impact the efficiency and accuracy of these integrated systems. The process of assimilating chlorophyll data into stochastic ensemble simulations demonstrates that improvements in predictions depend on the reliability of prior ensemble models. Regional diagnoses indicate that model instabilities can arise from mismatches in ensemble spread and observational variability, complicating the integration process (Santana-Falcón et al., 2020). Integrating feature selection and regression models for Chl-a prediction reveals the complexity of establishing relationships between environmental variables and Chl-a. The integration process is hindered by the high dimensionality and multicollinearity of environmental data, which complicates model training and prediction accuracy (Li, Sha, & Wang, 2018). The integration of various predictive models shows inconsistencies in performance metrics such as RMSE and AUC values, indicating variability in model reliability across different environmental conditions and data sets. Such inconsistencies challenge the robustness and

generalizability of integrated systems (Malek et al., 2011). The computational intensity required for real-time data processing and predictive modeling often exceeds the capacity of routine modelling systems. This limitation is significant in systems requiring high-frequency data updates for accurate Chl-a prediction (Jang et al., 2020). The effectiveness of integrated models is highly sensitive to the settings of various parameters, including data assimilation techniques and model configurations. This sensitivity can lead to significant prediction errors if not properly managed, affecting the system's ability to provide reliable forecasts (Du et al., 2018). Scaling integrated systems to different geographic regions or varying conditions often requires extensive customization, which can be resource-intensive and technically challenging. The adaptation of models to new regions may not always capture local biogeochemical processes accurately, reducing the effectiveness of the system (Fontana et al., 2009).

The integration of modeling and monitoring systems for Chl-a prediction faces several challenges, including data assimilation difficulties, computational demands, and sensitivity to model parameters. Addressing these limitations is crucial for enhancing the reliability and applicability of these systems in environmental management and monitoring. But integrating Chl-a modelling and monitoring online systems offers significant advantages in data visualization, spatial analysis, accessibility, and collaboration. The combination of Chl-a data with spatial information and the availability of online interfaces provide valuable tools for understanding, managing, and communicating Chl-a dynamics in lakes. This integration contributes to more informed decision-making processes and supports the sustainable management of lake ecosystems (Turner et al., 2010).

2.4 Summary

The literature review in Chapter 2 presents a comprehensive overview of the current methodologies and advancements in predicting Chl-a concentrations in lakes, emphasizing the significant transition from traditional approaches to advanced machine learning techniques. This shift addresses the limitations of conventional methods, notably their inability to process real-time data and manage the complex, non-linear interactions between environmental factors and Chl-a levels.

The review highlights several key points:

- Machine Learning Superiority: Advanced machine learning algorithms like SVM, RF, DT, GBT, KNN, MLP, LSTM, and XGBoost demonstrate superior capability in handling the complexities of environmental data compared to traditional statistical models. These methods offer robust solutions for accurately predicting Chl-a levels, which is crucial for timely and effective environmental management and decision-making.
- Eutrophication Risk Assessment: Incorporating machine learning into eutrophication risk assessment allows for more precise predictions and better management strategies, helping to mitigate one of the most pressing issues in lake management.
- Integration with Online Systems: The integration of these predictive models into online modelling systems represents a critical advancement in environmental management practices. Such systems provide real-time or near-real-time data that are essential for the early detection and management of potential ecological issues, such as algal blooms and eutrophication.

In summary, the literature underscores the need for and effectiveness of integrating machine learning techniques into Chl-a prediction and modelling frameworks. This integration not only enhances the accuracy of predictions but also improves the responsiveness of environmental management practices to potential ecological threats. The review sets the stage for the proposed research by establishing the context in which these technologies can be further developed and optimized for practical application in lake management.

Chapter 3 Methodology

3.1 Overview of the Online System Developed in This Thesis

The CMMOS developed in this thesis represents a sophisticated integration of AI, ML, and web technologies to provide a robust platform for the real-time modelling and predictive modeling of Chl-a concentrations in lakes. This system is designed to address the limitations of traditional water quality modelling methods, offering a comprehensive, efficient, and user-friendly solution to environmental scientists, policymakers, and lake management authorities.

Figure 3-1 depicts the comprehensive framework of the CMMOS, designed to integrate various stages of data handling, processing, modeling, and monitoring into a cohesive workflow. The process begins with the data input stage, where field measurements and relevant environmental parameters are systematically collected. This stage is followed by an elaborate data processing phase structured into five essential steps: missing value imputation, outlier detection, feature selection, data normalization, and data splitting. These preprocessing steps are crucial for enhancing the quality and integrity of the data, preparing it for effective modeling.

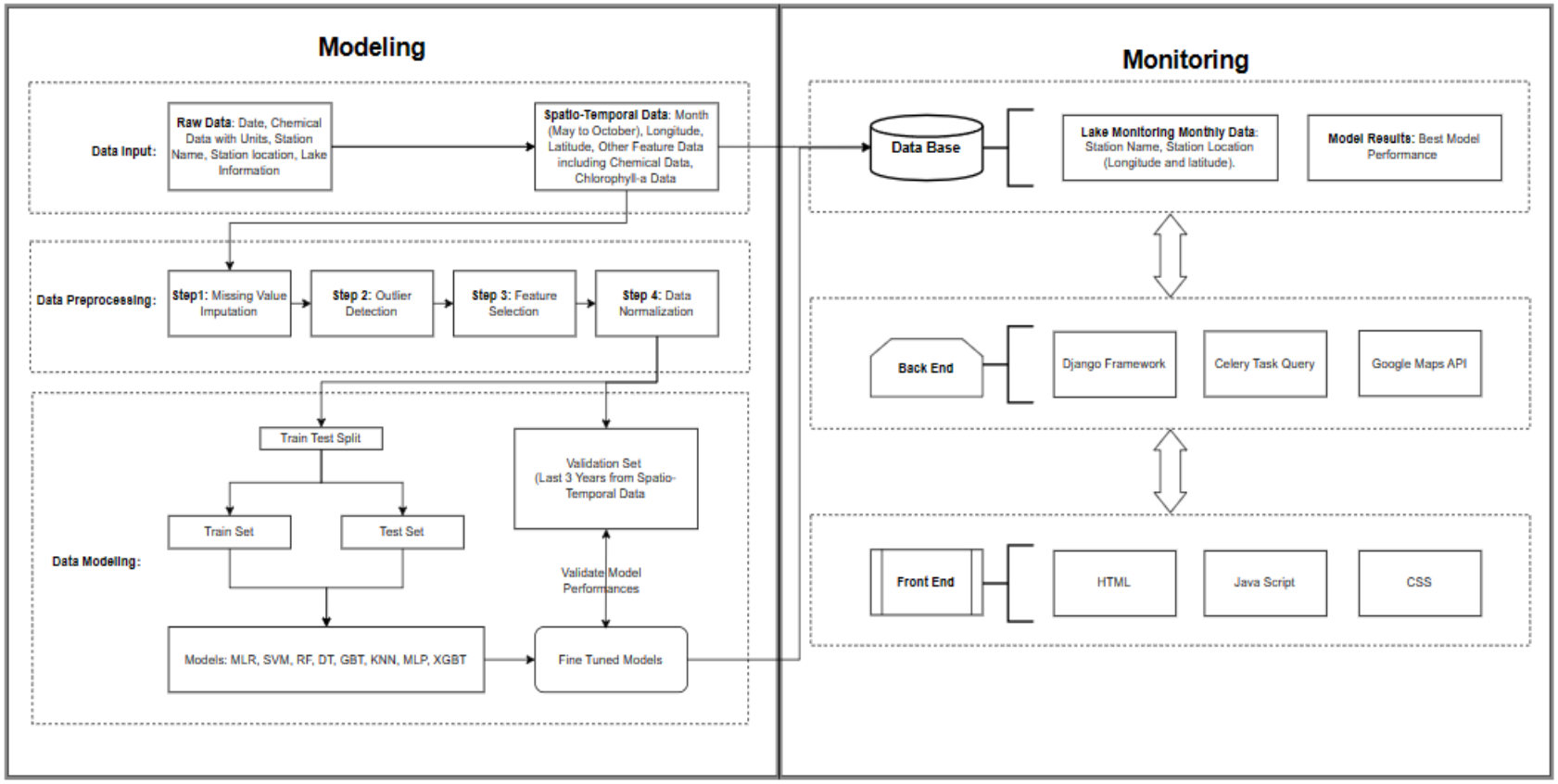

Fig 3-1 Framework of CMMOS

In the modeling phase, CMMOS employs a diverse array of machine learning models, including MLR, SVM, DT, RF, KNN, GBT, MLP, and Extreme Gradient Boosting (XGBOOST). These models are rigorously trained and tested on processed data to develop robust predictive algorithms. A continuous cycle of training and testing ensures that the models are meticulously fine-tuned for optimal performance.

The final component of the framework, the modelling stage, operationalizes the online system. This stage includes a backend database that stores the processed data and a frontend interface crafted using HTML, CSS, and JavaScript, with server-side operations facilitated by the Django framework. Geospatial data visualization is enhanced through the use of the Google Maps API, allowing end-users to monitor Chl-a concentrations in a user-friendly and accessible format. The modelling phase also encompasses model reliability assessments based on performance metrics, ensuring the continuous accuracy and reliability of the system's predictions.

3.2 Data Preprocessing

Data preprocessing is a critical phase in the Chl-a Modeling and Monitoring Online System (CMMOS) where raw data is transformed into a clean dataset suitable for efficient and accurate analysis. This stage is fundamental in ensuring the quality and integrity of the data used for modeling and prediction. Raw data, especially in environmental modelling, often contains inconsistencies such as missing values, outliers, or incorrect entries, which can significantly affect the reliability of any predictions made from the data. Thus, effective data preprocessing not only improves the accuracy of the model outcomes but also enhances the system’s overall performance by ensuring that the input data is consistent, normalized, and truly representative of the real-world conditions.

3.2.1 Missing Value Imputation

Missing values in the dataset can hinder the performance of machine learning models and lead to biased results. Therefore, it is crucial to handle missing values appropriately before training the models. One commonly used method for missing value imputation is mean imputation, which

replaces missing values with the mean value of the corresponding feature. Mean imputation is a simple and intuitive approach that assumes the missing values are missing at random (MAR). The mean value of the available data is used as a substitute for the missing values, thereby preserving the feature's overall distribution and central tendency. The imputed value is calculated using the following formula:

$$\hat{X}_i = \frac{\sum_{j=1}^{n} X_j}{n} \quad (1)$$

- $\hat{X}_i$represents the imputed value for the missing value at position i.
- X_j represents the observed values of the feature.
- n the number of observed values.

Mean imputation ensures that the imputed dataset retains the feature's original mean value by replacing missing values with the mean. However, it is essential to note that mean imputation may underestimate the feature's variance since it does not account for the uncertainty introduced by imputing missing values. Mean imputation is a straightforward and computationally efficient method, especially when dealing with large datasets. However, it has limitations. It assumes that missing values are missing completely at random (MCAR) or missing at random (MAR), which may not always hold in real-world scenarios. Additionally, mean imputation may introduce bias and distort the relationships between variables if the missingness mechanism is not MAR. Despite these limitations, mean imputation remains a popular choice for handling missing values due to its simplicity and ease of implementation. It can reasonably estimate disappeared values judiciously and with other preprocessing techniques and model evaluation methods.

3.2.2 Outlier Detection

Outliers are data points that significantly deviate from the expected pattern or distribution of the dataset. They can arise due to measurement errors, data entry mistakes, or represent extreme observations. Detecting and handling outliers is essential to ensure the robustness and reliability of machine learning models. One commonly used method for outlier detection is the Interquartile Range (IQR) method.

The IQR method identifies outliers by calculating the range between the 75th percentile ($Q3$) and the 25th percentile ($Q1$) of a feature's values. The IQR is calculated using the following formula:

$$IQR = Q3 - Q1 \tag{2}$$

- IQRrepresents the interquartile range.
- $Q3$ represents the 75th percentile (third quartile).
- $Q1$represents the 25th percentile (first quartile).

To identify outliers using the IQR method, the lower threshold is computed as $Q1 - 1.5 \times IQR$ and the upper threshold as $Q3 + 1.5 \times IQR$. Any data point below the lower threshold or above the upper threshold is considered an outlier. Using the IQR method, outliers that fall beyond a certain threshold are identified and can be further analyzed or treated accordingly. Depending on the specific application and the nature of the data, outliers can be removed from the dataset, transformed, or replaced with more representative values. It is important to note that the choice of the 1.5 multiplier in the IQR method is somewhat arbitrary and can be adjusted based on the analysis's specific requirements or the data's characteristics. A higher multiplier will be more lenient in identifying outliers, while a lower multiplier will be more stringent.

3.2.2 Feature Selection

Feature selection is a crucial step in data preprocessing. It aims to identify the most informative and relevant features for the prediction task. It helps reduce dimensionality, improve model performance, and enhance interpretability. One widely used method for feature selection is the chi-square test.

The chi-square test measures the independence between two categorical variables and is commonly used to assess the relationship between each feature and the target variable. It calculates the chi-square statistic, quantifying the difference between the observed and expected frequencies of the feature's categories about the target variable.

The chi-square statistic is calculated using the following formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (3)$$

- χ^2 represents the chi-square statistic.
- O_{ij} represents the observed frequency of the i_{th} category of the feature and the j_{th} category of the target variable.
- E_{ij} represents the expected frequency of the i_{th} category of the feature and the j_{th} category of the target variable, assuming independence.

The chi-square statistic follows a chi-square distribution, and its significance level determines the importance of the feature. By comparing the computed chi-square statistic to a critical value from the chi-square distribution table or calculating the p-value associated with the statistic, it is possible to determine if the feature is significantly associated with the target variable. The chi-square test can be performed for each feature independently, and features with high chi-square statistics or low p-values are considered more relevant and informative. These features are likely to impact the target variable significantly and can be selected for further analysis or model training. However, it is essential to note that the chi-square test is applicable only for categorical features and categorical target variables. Other feature selection methods, such as correlation analysis or mutual information, may be more appropriate for numerical features or continuous target variables. Additionally, the chi-square test assumes that the observations are independent and that the expected frequencies are sufficiently large. If the assumptions are violated, the test results may be less reliable. Therefore, it is essential to consider the limitations and assumptions of the chi-square test when applying it for feature selection.

3.2.3 Data Normalization

Data normalization is an essential step in data preprocessing. It aims to transform the data into a standardized range and eliminate the influence of different scales or units. One commonly used method for data normalization is Min-Max scaling.

Min-max or feature scaling rescales the data to a specified range, typically between 0 and 1. It works by subtracting the minimum value of the feature and dividing it by the range of the feature. The formula for Min-Max scaling is as follows:

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (4)$$

- $X_{\text{normalized}}$ represents the normalized value of the data point.
- X represents the original value of the data point.
- $X_{\min}$represents the minimum value of the feature.
- $X_{\max}$ represents the maximum value of the feature.

By applying Min-Max scaling, the feature's minimum value is transformed to 0, and the maximum value is transformed to 1. The values in between are linearly scaled proportionally. This normalization technique ensures that all features have the same scale, making them directly comparable and preventing features with larger values from dominating the analysis.

Min-max scaling is particularly useful when the data distribution is known to be approximately linear or when the data needs to be constrained within a specific range. However, it is sensitive to outliers, as they can significantly affect the scaling of the feature. Therefore, it is often recommended to handle outliers before applying Min-Max scaling or to consider alternative normalization methods, such as Z-score normalization, which is more robust to outliers. In addition to its simplicity and ease of implementation, Min-Max scaling preserves the relationships between the data points and maintains the interpretability of the features. The transformed values retain the relative ordering and proportions of the original data.

It is important to note that data normalization should be performed separately for each feature, ensuring it is scaled independently. This prevents one feature from dominating the analysis due to a larger scale. In all, Min-Max scaling is a widely used method for data normalization, allowing for data transformation into a standardized range between 0 and 1. Applying this technique makes the scales of features comparable, facilitating meaningful comparisons and preventing features with larger values from overwhelming the analysis. However, it is essential to handle outliers appropriately and consider alternative normalization methods based on the specific characteristics of the data.

3.2.4 Data Splitting

Data splitting is a crucial step in machine learning and model development. It involves dividing the dataset into separate training, validation, and testing subsets. The commonly used method for data splitting is the random split, which ensures a representative distribution of data across the subsets.

The random split involves shuffling the dataset and assigning a portion of the data to each subset. The typical splitting proportions are as follows:

- Training set: This subset trains the model and establishes the relationships between the input features and the target variable. It should contain most data, usually around 70% to 80% of the dataset.
- Validation set: This subset is used to tune the model's hyperparameters and assess its performance during the training process. It helps prevent overfitting and select the best model configuration. The recommended proportion is around 10% to 15% of the dataset.
- Test set: This subset evaluates the final model's performance and assesses its generalization ability on unseen data. It provides an unbiased estimate of the model's performance. The test set should be kept separate from the training and validation sets until the model development process is complete. The remaining portion of the dataset, usually around 10% to 20%, is allocated for the test set.

The random split can be performed using a random sampling function or by shuffling the dataset and sequentially assigning data points to each subset. The split can be stratified, meaning the class distribution in the target variable is preserved across the subsets. This is particularly useful when dealing with imbalanced datasets to ensure each subset contains representative samples from each class.

3.3 Machine Learning Models

This section explores various machine-learning models commonly used for regression tasks. Regression models aim to predict continuous numerical values based on input features.

These models leverage the power of algorithms and mathematical techniques to learn patterns and relationships from training data, enabling accurate predictions on unseen data.

The machine learning models discussed in this section include Multiple Linear Regression (MLR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting Tree (GBT), KNN, Multilayer Perceptron (MLP), LSTM and Extreme Gradient Boosting (XGBOOST). Each model has its unique characteristics and is suitable for different scenarios, providing a diverse range of options for regression tasks.

3.3.1 MLR Model

MLR is a statistical technique used to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. MLR is widely used across scientific disciplines for forecasting, predictions, and inferential statistics. In Multiple Linear Regression, the relationship between the dependent variable y and the independent variables $x_1, x_2, \dots, x_n$ is represented by the following linear equation (Montgomery et al. 2012):

$$y = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n + \varepsilon \quad (5)$$

- y represents the dependent variable (the variable to be predicted).
- $x_1, x_2, \dots, x_n$ represent the independent variables (features).
- $w_0, w_1, w_2, \dots, w_n$ are the regression coefficients that represent the weights assigned to each independent variable.
- ε is the error term, representing the deviation between the predicted and actual values.

Multiple Linear Regression aims to estimate the regression coefficients that minimize the sum of squared errors. This is achieved by solving the following optimization problem (Montgomery et al. 2012):

$$min_{w_0,w_1,w_2,\dots,w_n} \sum_{i=1}^{m} \left(y_i - (w_0 + w_1x_{i1} + w_2x_{i2} + \cdots + w_nx_{in})\right)^2 \quad (6)$$

- m is the number of training samples.

- y_i is the actual value of the dependent variable for the i-th training sample.
- $x_{i1}, x_{i2}, \ldots, x_{in}$ are the values of the independent variables for the i-th training sample.

Various techniques, such as ordinary least squares (OLS) or gradient descent, can be used to solve the optimization problem. Multiple Linear Regression offers several advantages. It is simple to understand and interpret and can handle both numerical and categorical independent variables. It also provides insights into the importance and direction of the relationships between the independent and dependent variables through the regression coefficients. However, Multiple Linear Regression assumes a linear relationship between the independent variables and the dependent variable, which may not always hold true in real-world scenarios. It may not capture complex nonlinear relationships. In such cases, more advanced machine learning algorithms, such as decision trees or neural networks, may be more suitable.

In this thesis, the MLR model is employed to predict Chl-a concentrations based on environmental variables such as temperature, pH, and turbidity. Data on Chl-a concentrations and environmental variables are collected through both field measurements and remote sensing technologies. This data is then preprocessed to handle missing values, outliers, and to ensure normalization, facilitating effective model training and predictions. The MLR model is developed using statistical software. This involves selecting relevant environmental variables based on their correlation with Chl-a concentrations and assessing their multicollinearity. The model's performance is evaluated using statistical metrics such as R2 and Root Mean Squared Error (RMSE). These metrics help determine how well the model explains the variability in Chl-a concentration and the accuracy of its predictions. Once validated, the MLR model is integrated into the online modelling system. This integration enables the continuous prediction of Chl-a levels, providing real-time insights into the lake's ecological health and eutrophication risk. The predictions from the MLR model are visualized through a user-friendly interface in the online system, allowing environmental managers and policymakers to make informed decisions based on the current and predicted states of water quality.

3.3.2 SVM Model

The Support Vector Machine (SVM) model is an essential component of the integrated AI-based system developed in this thesis for modeling and monitoring Chl-a concentration in lakes. SVM is renowned for its effectiveness in classification and regression tasks, particularly in high-dimensional spaces. In this thesis, SVM is utilized for the regression task of predicting Chl-a levels, a key indicator of water quality and ecological health in lake environments. SVM aims to find a hyperplane that maximizes the margin between the training data points and the hyperplane while minimizing the prediction error. In regression tasks, the goal is to find a hyperplane that best fits the data points with a maximum margin. The hyperplane is defined by a linear equation:

$$f(x) = w^T x + b \tag{7}$$

- $f(x)$ represents the predicted value for the input data point x.
- w is the weight vector.
- b is the bias term.

The weight vector w and the bias term b are determined by solving the optimization problem (Wujek et al., 2016):

$$min_{w,b} \frac{1}{2} || w ||^2 + C \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 \tag{8}$$

Subject to:

$$y_i - f(x_i) \leq \epsilon \tag{9}$$

$$f(x_i) - y_i \leq \epsilon \tag{10}$$

- n is the number of training samples.
- y_i is the target value for the i-th training sample.
- $f(x_i)$ is the predicted value for the i-th training sample.
- ε is the parameter that controls the width of the error tube.
- C is the regularization parameter that balances the trade-off between the margin and the error.

The optimization problem is solved using quadratic programming techniques. By solving the problem, the weight vector w and the bias term b can be obtained, and the prediction can be made using the linear equation.

SVM offers several advantages in regression tasks. It can capture complex relationships and handle high-dimensional data effectively. SVM is less prone to overfitting due to the margin maximization principle. It can also handle both numerical and categorical features by employing appropriate kernel functions. However, SVM's performance may be affected by the choice of the kernel function and the regularization parameter. The kernel function determines the type of decision boundary that can be learned, and the regularization parameter controls the balance between margin maximization and error minimization. Proper parameter tuning is necessary to achieve optimal performance.

By applying SVM within the comprehensive AI-based system, this thesis advances the predictive accuracy and operational efficiency of Chl-a concentration modelling in lake ecosystems.

3.3.3 DT Model

The Decision Tree (DT) model is an integral part of the AI-based system developed in this thesis for modeling and monitoring Chl-a concentrations in lakes. Decision Trees are popular due to their simplicity, interpretability, and ease of use in classification and regression tasks. In this research, DTs are utilized for regression purposes to predict Chl-a levels, crucial for assessing the ecological health of lakes. Decision Trees involve splitting the data into subsets based on the value of the input features that result in the greatest reduction of variance (or another metric) in the target variable. For regression tasks, a typical Decision Tree can be represented as a series of decisions leading to predictions (Loh, W-y, 2011):

$$f(x) = \sum_{i=1}^{n} w_i y_i \ \epsilon \tag{11}$$

- $f(x)$ represents the predicted value for the input data point x.
- n is the number of leaf nodes in the Decision Tree.
- w_i is the weight assigned to the i-th leaf node.
- y_i is the predicted value at the i-th leaf node.

The weights assigned to the leaf nodes can be determined based on different criteria. For example, in the case of minimizing the variance, the weight w_i can be calculated as the fraction of

training samples that belong to the i-th leaf node. Alternatively, if the goal is to minimize the SSE, the weight w_i can be determined as the average of the target values within the i-th leaf node.

Decision Tree offers several advantages. It is easy to interpret and visualize, as the resulting tree structure provides insights into the decision-making process. Decision Tree can handle both numerical and categorical features without extensive preprocessing. It is also robust to outliers and can handle missing values by employing appropriate strategies for splitting and imputing. However, Decision Tree may suffer from overfitting, especially when the tree depth is unconstrained. To mitigate overfitting, techniques like pruning or setting a maximum depth can be applied. Additionally, Decision Tree models may have limited interpretability compared to simpler models like linear regression. Nevertheless, the interpretability can be enhanced by considering feature importance measures derived from the Decision Tree structure.

A Decision Tree model is constructed using the selected features. The tree is grown by repeatedly splitting the training data into subsets, starting from the root node, based on the feature that provides the best split according to a given metric (commonly variance reduction for regression). To avoid overfitting, the tree is pruned back from its fullest depth. This involves removing sections of the tree that provide little power in predicting the target variable, thereby enhancing the model's generalization capabilities. The Decision Tree is trained on the preprocessed dataset, and its performance is validated using techniques such as k-fold cross-validation. Performance metrics, like Mean Absolute Error (MAE) and R-squared, are used to evaluate the accuracy of the model. Once validated, the Decision Tree model is integrated into the online modelling system. This allows for continuous and automated prediction of Chl-a levels based on real-time data feeds from various sensors and data sources. Predictions from the Decision Tree model are displayed through a user-friendly dashboard that provides actionable insights into water

quality trends and potential ecological risks. This interface facilitates easy interpretation and decision-making for lake management.

By integrating the Decision Tree model into the broader AI-based system, this thesis enhances the predictive capabilities for modelling Chl-a levels, providing a robust tool for environmental scientists and lake managers to assess and manage water quality effectively.

3.3.4 RF Model

The Random Forest (RF) model is a key component of the AI-based system developed in this thesis for modeling and monitoring Chl-a concentrations in lakes. Random Forest is an ensemble learning method that utilizes multiple decision trees to improve predictive accuracy and control over-fitting. This methodology is particularly suitable for handling complex and non-linear relationships between multiple environmental predictors and Chl-a levels.

The algorithm starts by creating a set of decision trees using bootstrap samples of the original training data. Each decision tree is trained on a different subset of the data, known as a bootstrap sample, which is created by randomly sampling the training data with replacement. Additionally, at each split in the decision tree, only a random subset of features is considered for splitting. This introduces randomness and diversity in the ensemble of decision trees. The prediction process in Random Forest involves aggregating the predictions of all the decision trees. For classification tasks, the class label with the majority vote from the decision trees is chosen as the final prediction. Mathematically, the prediction can be represented as (Breiman, L. 2001):

$$f(x) = \text{argmax}y(\sum_{i=0}^{n} I(y_i = y)) \in \quad (12)$$

- $f(x)$ represents the predicted class label for the input data point x.
- n is the number of decision trees in the Random Forest ensemble.
- y_i is the class label predicted by the i-th decision tree.
- $I(y_i = y)$ is an indicator function that returns 1 if the predicted class label of the i-th decision tree is equal to y and 0 otherwise.

A RF model is constructed using the prepared dataset. Parameters like the number of trees in the forest (n_estimators), the number of features to consider for each split (max_features), and the minimum number of samples required at each leaf node (min_samples_leaf) are optimized to enhance model performance. The Random Forest model is trained on the dataset. During training, bootstrap samples of the data are used to build each tree, and random subsets of features are considered for splitting at each node, ensuring diverse trees and reducing the variance of the model. The performance of the Random Forest model is validated using cross-validation techniques. Metrics such as the Mean Squared Error (MSE) and the Coefficient of Determination (R^2) are used to assess the accuracy and explanatory power of the model. The trained Random Forest model is integrated into the AI-based online monitoring system. This integration allows the model to utilize real-time data for continuous prediction and modelling of Chl-a levels. Predictive results and insights generated by the Random Forest model are visualized through interactive dashboards. These visualizations support decision-making processes for lake management and environmental modelling.

By applying the Random Forest model within the comprehensive system developed in this thesis, the methodology enhances the predictive accuracy and real-time modelling capabilities for Chl-a concentrations in lakes, thus supporting sustainable water quality management and ecological assessments.

3.3.5 GBT Model

GBT is a robust machine learning technique that enhances predictive accuracy through the ensemble of decision trees. In this thesis, GBT is employed to predict Chl-a concentrations, a critical indicator of water quality and algal biomass in lakes. GBT is particularly effective in handling complex datasets with non-linear relationships and interactions among predictors.

Mathematically, the prediction $\hat{y}$ of a Gradient Boosting Tree model can be represented as (Friedman, J. H. 2001):

$$\hat{y} = \sum_{i=1}^{N} f_i(x) \tag{14}$$

- N is the number of trees in the ensemble.
- $f_i(x)$ represents the prediction of the i-th tree for the input x.

The objective of Gradient Boosting Tree is to minimize a loss function by finding the optimal values of the tree parameters. The loss function is typically a differentiable function that measures the difference between the predicted values and the actual values. A commonly used loss function for regression tasks is the mean squared error (MSE):

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{15}$$

- m is the number of training samples.
- y_i is the actual value of the dependent variable for the i-th training sample.
- $\hat{y}_i$ is the predicted value of the dependent variable for the i-th training sample.

The algorithm optimizes the loss function by iteratively fitting decision trees to the negative gradient of the loss function. This process is known as gradient boosting. The trees are added one at a time, and each tree is trained to minimize the loss function with respect to the residuals of the previous predictions.

Gradient Boosting Tree offers several advantages. It can capture complex nonlinear relationships and interactions among variables. It is also robust to outliers and noisy data. Additionally, it handles missing values and can handle both numerical and categorical features without extensive preprocessing. However, Gradient Boosting Tree can be prone to overfitting if not properly tuned. Careful selection of hyperparameters, such as the learning rate, number of trees, and maximum tree depth, is necessary to prevent overfitting and achieve optimal performance.

The application of GBT in this thesis represents a significant advancement in predictive modeling for aquatic environments, providing a powerful tool for real-time and accurate prediction of Chl-a concentrations, thereby supporting effective water quality management and ecological preservation efforts.

3.3.6 KNN Model

KNN is a simple yet effective machine learning algorithm for regression tasks. It is a non-parametric algorithm that makes predictions based on the k nearest neighbors of a given data point. The algorithm works by calculating the distances between the data point to be predicted and all other data points in the training set. It then selects the k nearest neighbors based on these distances. The predicted value for the data point is obtained by averaging the values of its k nearest neighbors.

Mathematically, the prediction $\hat{y}$ of a KNN model for a given input x can be represented as (Altman, N. S. 1992):

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i \tag{16}$$

- k is the number of nearest neighbors to consider.
- y_i represents the target value of the i-th nearest neighbor.

The choice of the appropriate value for k is important and depends on the dataset and the problem at hand. A smaller k value results in a more flexible model that is sensitive to local patterns, but it may also be more prone to noise. On the other hand, a larger k value smooths out the predictions but may fail to capture local patterns.

KNN offers several advantages. It is simple to understand and implement. It can capture local patterns and nonlinear relationships. It is also robust to outliers. Moreover, it can handle both numerical and categorical features without extensive preprocessing. However, KNN has considerations to be aware of. The choice of k is crucial, and it may require tuning to achieve optimal performance. Additionally, KNN can be computationally expensive, especially for large datasets, as it involves calculating distances between the test sample and all training samples.

By leveraging the KNN model within the broader AI-based system, this thesis enhances the capability for accurate and real-time modelling of Chl-a concentrations in lakes, providing an essential tool for effective environmental management and research.

3.3.7 MLP Model

The MLP is a popular type of Extreme Gradient Boosting that can be used for regression tasks. It consists of multiple layers of interconnected neurons and is capable of learning complex nonlinear relationships between input features and target variables.

In MLP, the input layer receives the input features, and each neuron in the input layer is connected to every neuron in the next hidden layer. The hidden layers perform intermediate computations, and the output layer produces the final prediction. Each neuron applies a nonlinear activation function to the weighted sum of its inputs.

Mathematically, the output of an MLP can be represented as (Goodfellow, I., et. al, 2016):

$$\hat{y} = f\left(\sum_{j=1}^{n} w_j \cdot x_j + b\right) \tag{17}$$

- $\hat{y}$ represents the predicted value.
- f is the activation function, which introduces nonlinearity into the model.
- w_j are the weights associated with the input features x_j.
- b is the bias term.

The weights and biases in the MLP are learned during the training process using optimization algorithms such as backpropagation. The backpropagation algorithm calculates the gradients of the model's parameters with respect to a loss function and updates the weights and biases to minimize the loss.

MLP offers several advantages. It can model complex relationships and learn nonlinear patterns in the data. It is also capable of handling both numerical and categorical features. Additionally, MLP allows for flexible network architectures, including the number of hidden layers and neurons, which can be adjusted based on the complexity of the problem. However, MLP has considerations to be aware of. It requires careful selection of the activation function, as it can affect the model's convergence and performance. The choice of the number of hidden layers and neurons also requires tuning to prevent overfitting or underfitting. Additionally, training an MLP can be computationally expensive, especially for large datasets and complex architectures.

By employing the MLP model, this thesis contributes significantly to the field of environmental science by providing a robust predictive tool for Chl-a concentration, aiding in the effective management and modelling of water quality in lakes.

3.3.8 LSTM Model

Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), are well-suited for modeling time-series data due to their ability to remember information over long periods. This capability makes LSTMs ideal for predicting Chl-a concentrations in lakes, where the data exhibits temporal dependencies influenced by seasonal variations and environmental factors.

LSTM networks address the vanishing gradient problem common in traditional RNNs by incorporating memory cells that regulate the flow of information. Each cell in an LSTM layer has three types of gates: input, forget, and output gates, which control the cell state and the hidden state passed along the sequence. The basic equations governing these processes include (Hochreiter, S., & Schmidhuber, J. 1997):

- Forget Gate

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{18}$$

- Input Gate

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{19}$$

$$\widetilde{C_t} = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{20}$$

- Cell State Update

$$C_t = f_t * C_t + i_t * \widetilde{C_t} \tag{21}$$

- Output Gate

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{22}$$

$$h_t = o_t * \tanh(C_t) \tag{23}$$

Where σ represents the sigmoid activation function, $tanh$ is the hyperbolic tangent function, W and b are the weights and biases associated with each gate, h_t is the hidden state at time t, C_t is the cell state at time t, and x_t is the input at time t.

In this thesis, the Long Short-Term Memory (LSTM) model is developed to predict Chl-a concentrations in lakes, harnessing its capacity to handle sequential data and long-term dependencies. The process begins with collecting historical data on Chl-a and relevant environmental variables like nutrient levels, temperature, and sunlight exposure, followed by meticulous data preprocessing which includes scaling, managing missing values, and formatting data for LSTM compatibility. An appropriate LSTM architecture is designed, focusing on the number of layers and units, while incorporating strategies like dropout to mitigate overfitting. Feature engineering enhances the model's input by creating features that effectively capture temporal and seasonal trends. Training the LSTM involves backpropagation through time using optimizers like Adam or RMSprop to minimize errors, with hyperparameters such as learning rate and epoch number optimized through methods like grid search. The model's performance is validated using metrics like RMSE and MAE, and visual assessments of predicted versus actual values. Finally, the trained model is integrated into the modelling system for real-time predictive analysis, complemented by interactive visualizations that facilitate user interpretation and decision-making based on the predictive outputs.

By applying LSTM networks, this thesis provides a sophisticated approach to modeling and predicting Chl-a concentrations in lakes, enhancing the capabilities of modelling systems to manage water quality effectively.

3.3.9 XGBoost Model

XGBoost is an advanced implementation of gradient boosting algorithms known for its speed and performance. In this thesis, XGBoost is utilized for predicting Chl-a concentrations in lakes, a critical indicator of algal biomass and water quality. XGBoost is particularly effective due to its ability to handle various types of data, manage missing values, and capture complex nonlinear patterns in large datasets.

XGBoost improves upon the gradient boosting framework by regularizing the objective function to control overfitting, making it robust and efficient. The objective function of XGBoost for regression tasks includes both a loss function and a regularization term:

$$\mathrm{Obj}(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (24)$$

Where:

- $l(y_i, \hat{y}_i)$is the loss function that measures the difference between the predicted $\hat{y}_i$ and the actual y_i values.
- $\Omega(f_k)$ represents the regularization term that penalizes the complexity of the model.
- f_k are the functions corresponding to the individual trees in the model.

In this thesis, the XGBoost (Extreme Gradient Boosting) model is meticulously developed to predict Chl-a concentrations in lakes, capitalizing on its advanced algorithmic capabilities. The process begins with the collection and preprocessing of extensive datasets, which include Chl-a levels and related environmental variables like temperature, pH, and nutrient concentrations. This data is carefully prepared by handling missing values, normalizing features, and encoding categorical variables to optimize it for the XGBoost framework. Feature engineering is a critical step where influential predictors are identified and refined to enhance the model's performance. The XGBoost model is then configured with specific parameters such as the number of trees, maximum depth of trees, learning rate, and regularization parameters, all tuned to best fit the regression task at hand. Training involves the sequential building of decision trees, each correcting errors from the previous ones, effectively minimizing prediction errors through robust gradient boosting techniques. Hyperparameter tuning is conducted via methods like grid search to ensure optimal model settings, and the model's effectiveness is validated using cross-validation, assessing its accuracy through metrics like RMSE and R^2. Finally, the trained model is seamlessly integrated into an AI-based modelling system, providing real-time predictive insights into Chl-a concentrations. This integration supports continuous modelling and effective management of water quality, backed by interactive visualization tools that enable straightforward interpretation and decision-making by environmental managers and researchers.

The use of XGBoost in this thesis not only advances the predictive analysis of Chl-a concentrations but also contributes to the broader field of environmental modelling by providing a reliable, efficient, and scalable solution.

3.4 Model Evaluation Metrics

Model performance evaluation is a crucial step in machine learning to assess the effectiveness and accuracy of the developed models. Various evaluation metrics are used to measure the performance of regression models. In this section, we will discuss some commonly used metrics for evaluating the performance of regression models (Hyndman & Koehler, 2006).

3.4.1 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is derived from MSE and provides the measure of the average magnitude of the prediction errors. It is calculated as the square root of the MSE (Hyndman, R. J., & Koehler, A. B., 2006).:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \tag{25}$$

RMSE is widely used as it has the same unit as the target variable, making it easily interpretable.

3.4.2 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is another commonly used metric that measures the average absolute difference between the predicted values and the actual values. It is less sensitive to outliers compared to MSE. The formula for MAE is as follows (Hyndman, R. J., & Koehler, A. B., 2006).:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \tag{26}$$

3.4.3 Mean Absolute Percentage Error (MAPE)

The Mean Absolute Percentage Error (MAPE) measures the average percentage difference between the predicted values and the actual values. It is commonly used when the

target variable has a significant variation in magnitude. The formula for MAPE is as follows (Hyndman, R. J., & Koehler, A. B., 2006).:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \quad (27)$$

3.4.4 Coefficient of Determination (R^2)

R^2 provides an indication of goodness of fit and explains the proportion of variance in the dependent variable that is predictable from the independent variables. R^2 values closer to 1 indicate a better explanatory ability of the model (Hyndman, R. J., & Koehler, A. B., 2006)..

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (28)$$

This comprehensive approach to model evaluation ensures that the predictive models developed for chlorophyll-a concentration are robust, accurate, and suitable for supporting decision-making in lake management and environmental monitoring.

3.5 Development of the Online Chl-a Content Prediction System

The development of CMMOS marks a significant advancement in ecological data analytics, integrating sophisticated web technologies with environmental science for real-time modelling and prediction of Chl-a levels. This section outlines the comprehensive development process of CMMOS, including system architecture, front-end and back-end development, data processing, and the use of Python packages.

The Figure 3-2 displayed in the Water Quality Monitoring System illustrates the concentration trends of Chl-a over a selected time period across various monitoring points in Lake Simcoe. The x-axis of the graph represents the timeline, marked by specific years and months, indicating the period of data collection. The y-axis denotes the Chl-a concentration in micrograms per liter (μg/L). Data points are plotted as blue dots connected by line segments, highlighting the fluctuating levels of Chl-a. Each point on the graph corresponds to a specific sampling location and time, as indicated by identifiers like N31, C9, and E50, with their respective Chl-a values listed alongside the map. This visualization serves as a crucial tool for analyzing the spatial and temporal variation of Chl-a concentrations, offering insights into the ecological health of the lake and the effectiveness of ongoing environmental management strategies.

Fig 3-2 Monitoring Interface of CMMOS

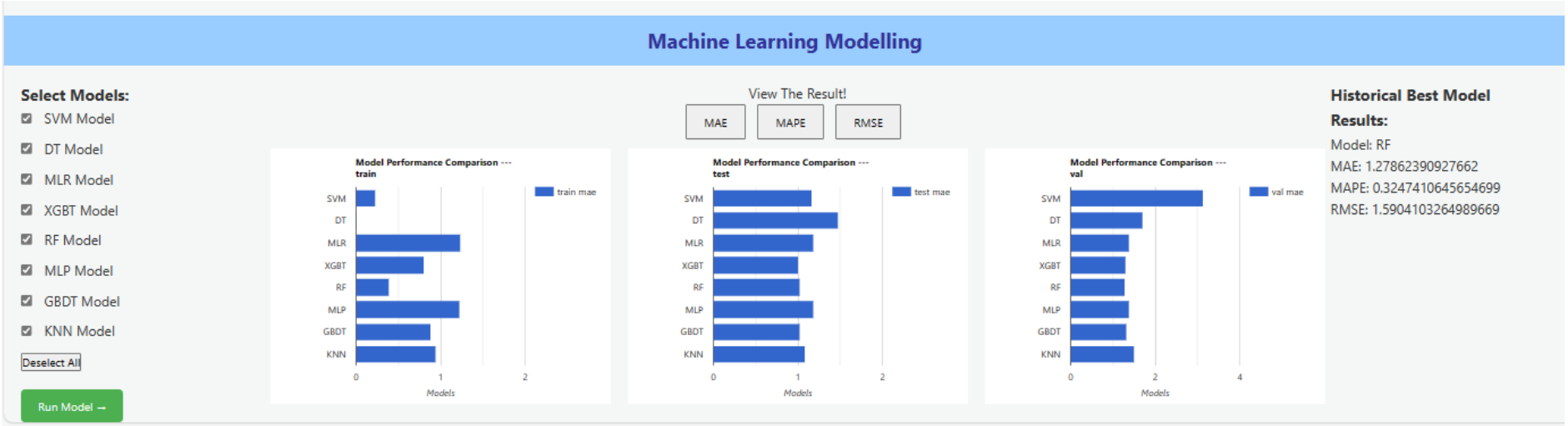

Fig 3-3 Modelling Interface of CMMOS

The Fig 3-3 is part of a machine learning platform designed to facilitate the selection and evaluation of different predictive models. The interface is divided into three main sections:

- Model Selection Panel:

On the left side, users can select from a list of models, including SVM (Support Vector Machine), DT (Decision Tree), MLR (Multiple Linear Regression), XGBT (Extreme Gradient Boosting Tree), RF (Random Forest), MLP (Multilayer Perceptron), GBT (Gradient Boosting Decision Tree), and KNN (K-Nearest Neighbors). There is also an option to 'Deselect All' for convenience.

- Model Performance Comparison Charts:

The center of the interface features two sets of bar charts comparing the performance of the selected models based on three metrics: MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and RMSE (Root Mean Square Error). The first chart shows the performance on training data, while the second chart depicts the performance on test data, allowing for an evaluation of both fitting and generalization capabilities.

- Historical Best Model Results:

On the right, there is a panel displaying the historical best model results, with metrics provided for the model that has historically performed the best on the data set. In this instance, the Random Forest (RF) model shows a MAE of 1.278, a MAPE of 0.327, and an RMSE of 1.590.

This interface is instrumental in allowing researchers and analysts to quickly assess and compare the efficacy of different machine learning models in a visually intuitive manner, facilitating better decision-making in model selection based on empirical data performance.

- Detailed Development Process

Front-End Development involves the creation of an interactive user interface using HTML5, CSS3, and JavaScript. Additionally, it integrates the Google Maps API for geospatial data visualization and employs Google Charts for dynamic data representation. Back-End Development is built on the Django framework, known for rapid development and pragmatic

design. Django REST framework is utilized to construct a powerful API, ensuring seamless interaction with the front end. Data Collection and Management are achieved through automated data collection scripts written in Python. These scripts fetch data from various environmental databases, and data storage and management are efficiently handled using PostgreSQL for robustness and scalability.

The detailed development process consists of several key components:

i. Front-End Design: The focus here is on creating a responsive user interface that enhances user experience through HTML5, CSS3, JavaScript, and AJAX for dynamic updates.
ii. Interactive Map and Data Visualization: Integration of the Google Maps API and custom marker implementation enables the display of geographic data. Google Charts is used for visually appealing and interactive graphing and charting.
iii. Back-End Infrastructure: Django serves as the core framework for managing requests, data processing, and view rendering. RESTful APIs built with Django REST framework enable seamless communication with the front-end.
iv. Asynchronous Tasks and Data Processing: Celery, an asynchronous task queue, is implemented for efficient background processing, while Redis serves as a message broker, facilitating communication between Django and Celery.
v. Predictive Modeling and Analysis: Python libraries like Pandas are employed for data manipulation and analysis. Scikit-learn is used for building predictive models, and TensorFlow is integrated for more complex deep learning models.

Chapter 4: Study Case and Field Investigation - Lake Champlain

4.1 Study Area

Lake Champlain, located in North America, spans across the borders of the United States and Canada, with the majority of its surface area lying within the states of Vermont and New York. It is a natural freshwater lake known for its ecological importance and recreational opportunities. Lake Champlain covers approximately 1,269 square kilometers (490 square miles) and stretches about 193 kilometers (120 miles) in length, with a maximum width of around 19 kilometres (12 miles). The lake's geographic coordinates range from approximately 44.0°N to 44.7°N latitude and 73.2°W to 73.5°W longitude.

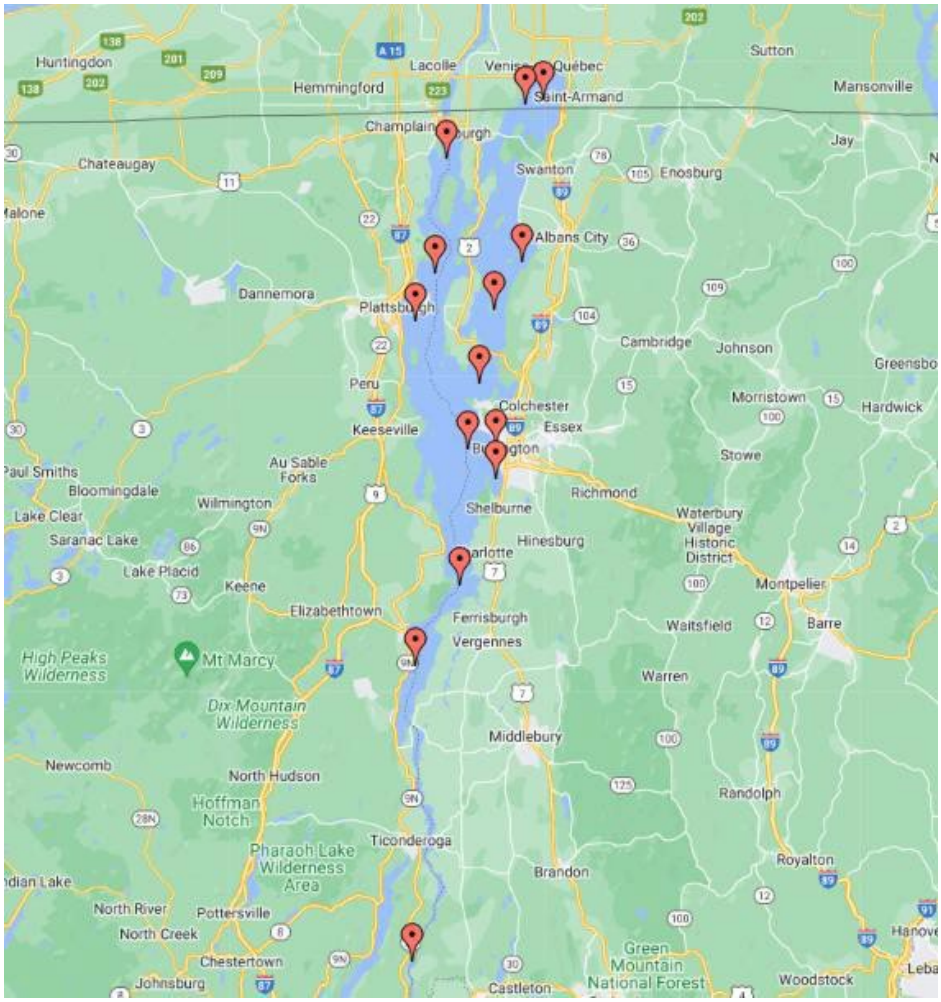

Fig 4-1 Location of Monitoring Stations in Lake Champlain

The lake has a diverse hydrological system and receives freshwater inflows and sediment loads from tributary rivers such as the Ausable River, Lamoille River, and Winooski River. It is also connected to the Richelieu River, which flows northward into the St. Lawrence River in Canada. Lake Champlain consists of several distinct basins, including the Main Lake, Northeast Arm, Northwest Arm, and South Lake, each with unique physical and ecological characteristics. It has a mean depth of approximately 19 meters (62 feet) and a maximum depth of around 122 meters (400 feet). The water residence time in the lake is approximately three years, indicating a relatively slow turnover rate. Moreover, the lake supports a diverse range of plant and animal species, providing essential habitats for fish such as lake trout, walleye, and smallmouth bass. It is also home to threatened or endangered species like the spiny softshell turtle and the lake sturgeon. Additionally, Lake Champlain serves as a critical stopover and breeding site for migratory birds during their annual journeys.

However, Lake Champlain faces water quality challenges, primarily related to nutrient enrichment and eutrophication. Excessive inputs of phosphorus and nitrogen from sources like agricultural runoff, urban stormwater, and wastewater treatment plants contribute to increased algal growth and degraded water quality. Harmful algal blooms, reduced water clarity, and oxygen depletion are among the negative impacts observed in certain areas of the lake. Extensive research and management efforts have been undertaken to address these challenges. Collaborative initiatives involving governmental agencies, research institutions, and non-profit organizations focus on modelling water quality, conducting ecological studies, and developing management strategies. These efforts involve sampling water quality, analyzing nutrient concentrations, measuring algal biomass, and utilizing remote sensing technologies to monitor the lake's ecological dynamics. Modelling approaches and stakeholder engagement also play essential roles in informing decision-making processes for the sustainable management of Lake Champlain.

In conclusion, Lake Champlain serves as an important study area for investigating eutrophication and water quality issues. Its unique geographic and hydrological characteristics, ecological significance, and ongoing research and management efforts provide a valuable context for understanding and addressing the challenges associated with this freshwater ecosystem.

4.2 Data Source

The data used in this study was sourced from the Lake Champlain Long-term Modelling Project. Lake Champlain is a large freshwater lake located in North America, spanning across the borders of the United States (Vermont and New York) and Canada (Quebec). It is a vital water resource, supporting various ecological habitats and serving as a recreational and economic asset for surrounding communities. The Lake Champlain Long-term Monitoring Project is a comprehensive research initiative that aims to monitor and assess the water quality and ecological health of Lake Champlain over an extended period. The project involves the collection of various environmental data, including physical, chemical, and biological parameters, from multiple monitoring stations distributed across the lake. These monitoring stations are strategically located to capture spatial variations in water quality and to provide representative data for different regions of the lake. Data collection is carried out at regular intervals, ensuring temporal coverage and facilitating the analysis of long-term trends and seasonal variations.

The dataset from the Lake Champlain Long-term Monitoring Project comprises a rich and extensive collection of water quality measurements, including parameters such as temperature, dissolved oxygen, pH, nutrient concentrations (e.g., phosphorus, nitrogen), Chl-a levels, and other relevant variables. The data spans multiple years, providing a valuable resource for studying the dynamics of water quality and the factors influencing Chl-a concentrations in Lake Champlain. The data from the Lake Champlain Long-term Monitoring Project is widely recognized and utilized by researchers, policymakers, and environmental managers. Its availability and reliability make it an ideal source for conducting comprehensive studies and informing management decisions related to the ecological health and water quality of Lake Champlain.

In the following sections, we will describe the data preprocessing steps and the machine learning models applied to analyze and predict Chl-a concentrations based on the Lake Champlain Long-term Monitoring Project dataset.

4.3 Results

The performance of various machine learning models was evaluated using the Lake Champlain dataset, and the results are presented in this section.

4.3.1 Data Preprocessing Result

In this section, we present the results of the data preprocessing techniques applied to the Lake Champlain dataset. Table 4-1 showed that the dataset was divided into six distinct sets: Set 1, Set 2, Set 3, Set 4, Set 5, and Set 6, each obtained through specific preprocessing steps including MSI (Missing Value Imputation), OD (Outlier Detection), FS (Feature Selection) and TTS (Train Test Split)

Table 4-1 Differences between each dataset

Data Set	With MSI	With OD	With FS	TTS
Set 1	Yes	No	No	Train
Set 2	Yes	No	No	Test
Set 3	Yes	Yes	No	Train
Set 4	Yes	Yes	No	Test
Set 5	Yes	Yes	Yes	Train
Set 6	Yes	Yes	Yes	Test

Set 1 and Set 2 were obtained after performing missing value imputation and representing the training and test data, respectively. Set 3 and Set 4 include the data after missing value imputation and outlier detection, serving as the training and test datasets. Set 5 underwent missing value imputation, outlier detection, and feature selection and was further divided into training and testing subsets. Similarly, Set 6 mirrored Set 5 and will be used to evaluate the performance of the trained models. These sets allow for a comprehensive analysis of the impact of each preprocessing step on the machine learning models' performance.

4.3.2 Data Standardization Results

Figure 4-2 illustrates the impact of data standardization on the performance of the Gradient Boosting Decision Tree (GBT) model across six datasets. The performance metric used is the coefficient of determination (R^2), which quantifies the proportion of variance explained by the model. In Set 1, the R^2 values for standardized and non-standardized data are comparable, indicating minimal effect from standardization. Sets 3, 4, and 5 also show little to no change in R^2 values between standardized and non-standardized data, suggesting that standardization does not significantly impact these datasets. However, Set 2 shows a clear improvement in R^2 with standardized data, indicating that standardization enhances model performance for this dataset. Set 6 also exhibits a slight improvement in R^2 with standardization, though the effect is less pronounced than in Set 2. These results indicate that while data standardization does not uniformly improve GBT model performance across all datasets, it can lead to enhancements in certain cases, underscoring the importance of incorporating standardization in data preprocessing.

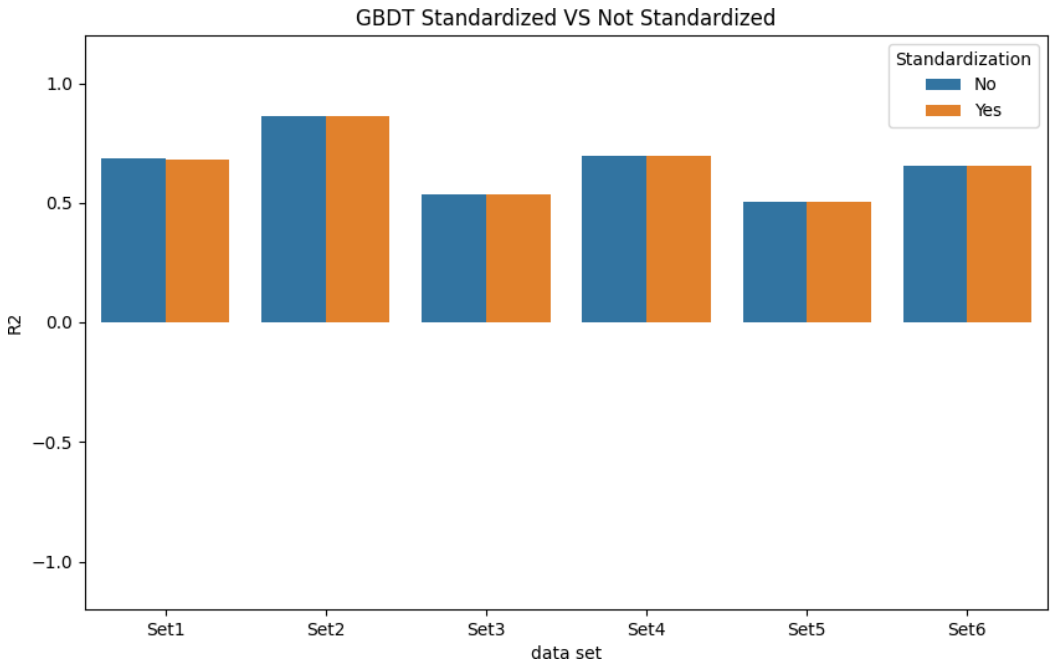

Fig 4-2 Effect of Data Standardization on GBT Model Performance Across Datasets

Figure 4-3 illustrates the impact of data standardization on the performance of the Long Short-Term Memory (LSTM) model across six datasets, using the coefficient of determination (R^2)

as the performance metric. In Set 1, the R^2 values for standardized and non-standardized data are comparable, indicating minimal impact from standardization. Set 2 shows a slight improvement in R^2 with standardized data, suggesting a positive influence on model performance. However, Sets 3 and 4 exhibit a significant decline in R^2 values with standardized data, indicating that standardization adversely affects the model's predictive accuracy for these datasets. This decline is particularly pronounced in Set 3, where the R^2 value drops below zero, suggesting poor model fit. For Sets 5 and 6, the R^2 values for both standardized and non-standardized data are close to zero, indicating that the LSTM model struggles to predict accurately regardless of standardization. These results demonstrate that data standardization can have varying effects on the LSTM model's performance across different datasets, highlighting the need for a case-by-case approach in data preprocessing.

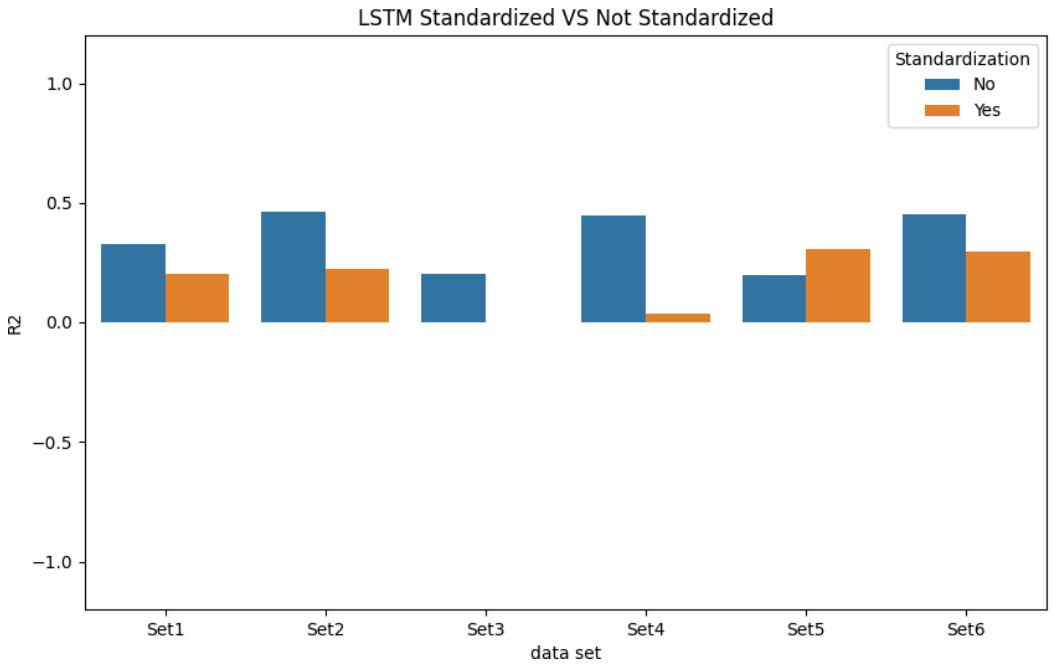

Fig 4-3 Effect of Data Standardization on LSTM Model Performance Across Datasets

Figure 4-4 illustrates the impact of data standardization on the performance of the Decision Tree (DT) model across six datasets, using the coefficient of determination (R^2) as the performance metric. In Set 1, the R^2 values for standardized and non-standardized data are comparable, indicating minimal impact from standardization. Set 2 shows no difference in R^2 between

standardized and non-standardized data, suggesting that standardization does not affect the model's performance for this dataset. Set 3 reveals a negative R^2 value for both standardized and non-standardized data, indicating poor model performance regardless of standardization. Set 4 shows a slight improvement in R^2 with standardized data, highlighting a positive effect of standardization. Sets 5 and 6 exhibit high R^2 values for both standardized and non-standardized data, indicating that the model performs well irrespective of standardization. These results suggest that while data standardization may not consistently improve the performance of the DT model across all datasets, it can lead to performance enhancements in specific cases, underscoring the importance of considering standardization in the preprocessing pipeline.

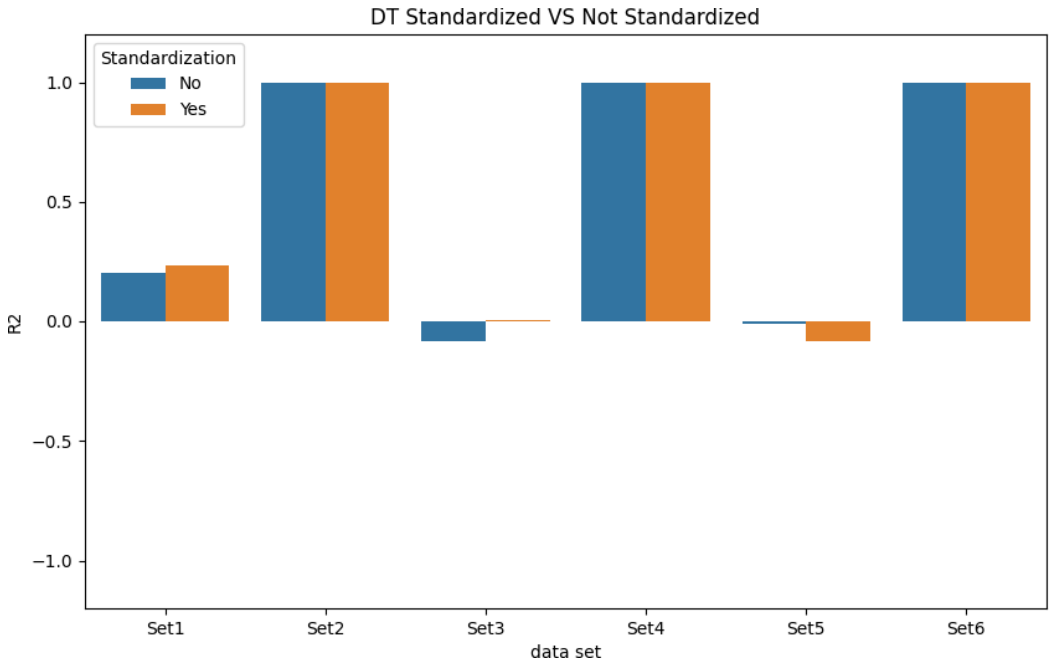

Fig 4-4 Effect of Data Standardization on DT Model Performance Across Datasets

Figure 4-5 illustrates the impact of data standardization on the performance of the KNN model across six datasets, using the coefficient of determination (R^2) as the performance metric. In Set 1, the R^2 values show a slight improvement with standardized data, indicating a positive effect of standardization. Set 2 also exhibits higher R^2 values for standardized data, suggesting enhanced model performance. In Set 3, standardization leads to a notable increase in R^2, indicating

that the model benefits significantly from this preprocessing step. Similarly, Set 4 shows a marked improvement in R^2 with standardized data, highlighting the positive impact of standardization. Sets 5 and 6 both exhibit higher R^2 values for standardized data compared to non-standardized data, though the difference is less pronounced than in Sets 2 and 4. These results suggest that data standardization generally improves the performance of the KNN model across various datasets, underscoring the importance of incorporating standardization in the data preprocessing pipeline to enhance model accuracy and reliability.

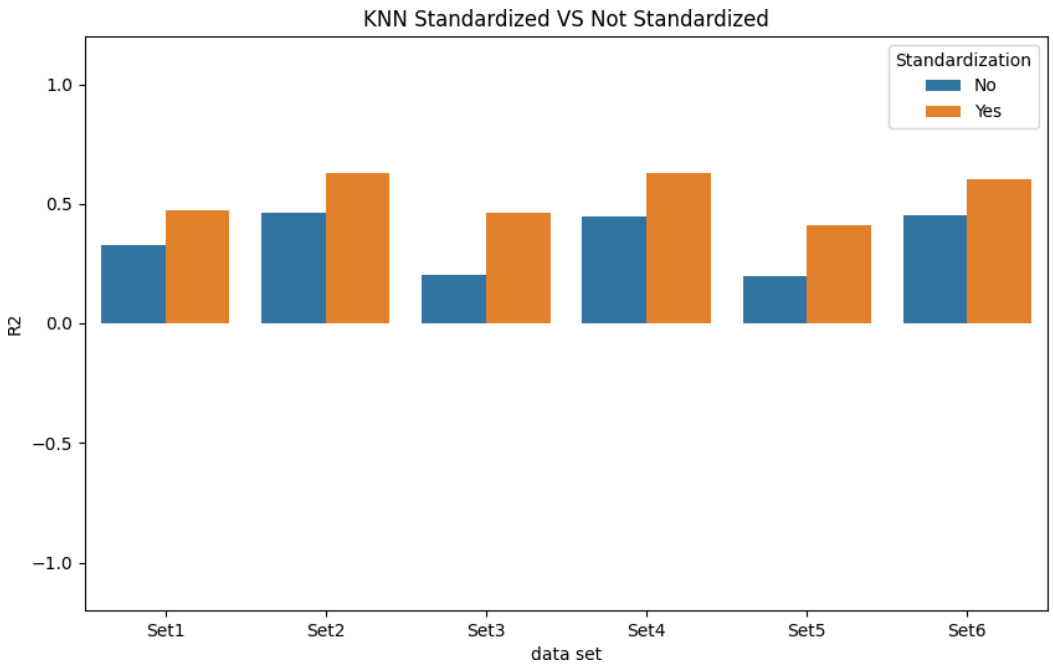

Fig 4-5 Effect of Data Standardization on KNN Model Performance Across Datasets

Figure 4-6 illustrates the impact of data standardization on the performance of the Support Vector Machine (SVM) model across six datasets, using the coefficient of determination (R^2) as the performance metric. In Set 1, the R^2 values show a slight improvement with standardized data, indicating a positive effect of standardization. Set 2 also exhibits a significant increase in R^2 with standardized data, suggesting enhanced model performance. In Set 3, standardization leads to a noticeable improvement in R^2, demonstrating the benefits of this preprocessing step. Set 4 reveals

a dramatic increase in R^2 with standardized data, highlighting the substantial positive impact of standardization on the model's predictive accuracy. Sets 5 and 6 also show considerable improvements in R^2 with standardized data compared to non-standardized data, further emphasizing the importance of standardization. These results indicate that data standardization consistently enhances the performance of the SVM model across various datasets, underscoring the critical role of standardization in the data preprocessing pipeline to achieve better model accuracy and reliability.

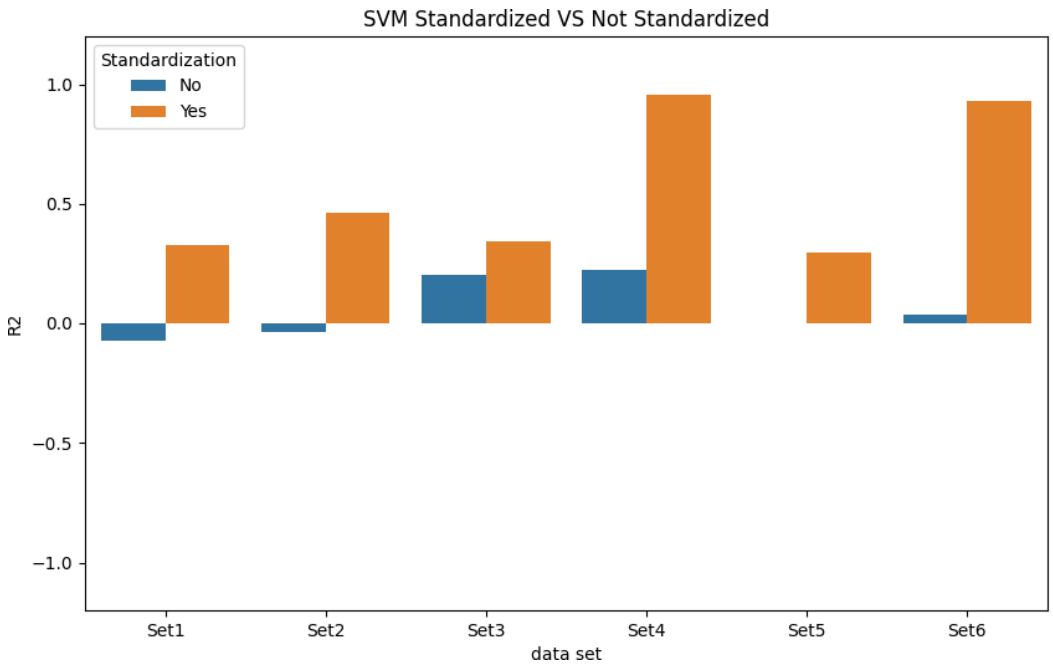

Fig 4-6 Effect of Data Standardization on SVM Model Performance Across Datasets

Figure 4-7 illustrates the impact of data standardization on the performance of the Random Forest (RF) model across six datasets, using the coefficient of determination (R^2) as the performance metric. In Set 1, the R^2 values for standardized and non-standardized data are comparable, indicating minimal impact from standardization. Set 2 shows no difference in R^2 between standardized and non-standardized data, suggesting that standardization does not affect the model's performance for this dataset. Set 3 reveals similar R^2 values for both standardized and

non-standardized data, indicating that standardization has little to no effect. Set 4 shows a slight improvement in R^2 with standardized data, highlighting a positive effect of standardization. Sets 5 and 6 exhibit higher R^2 values for standardized data compared to non-standardized data, though the difference is less pronounced than in other datasets. These results suggest that while data standardization may not consistently improve the performance of the RF model across all datasets, it can lead to performance enhancements in specific cases, underscoring the importance of considering standardization in the preprocessing pipeline.

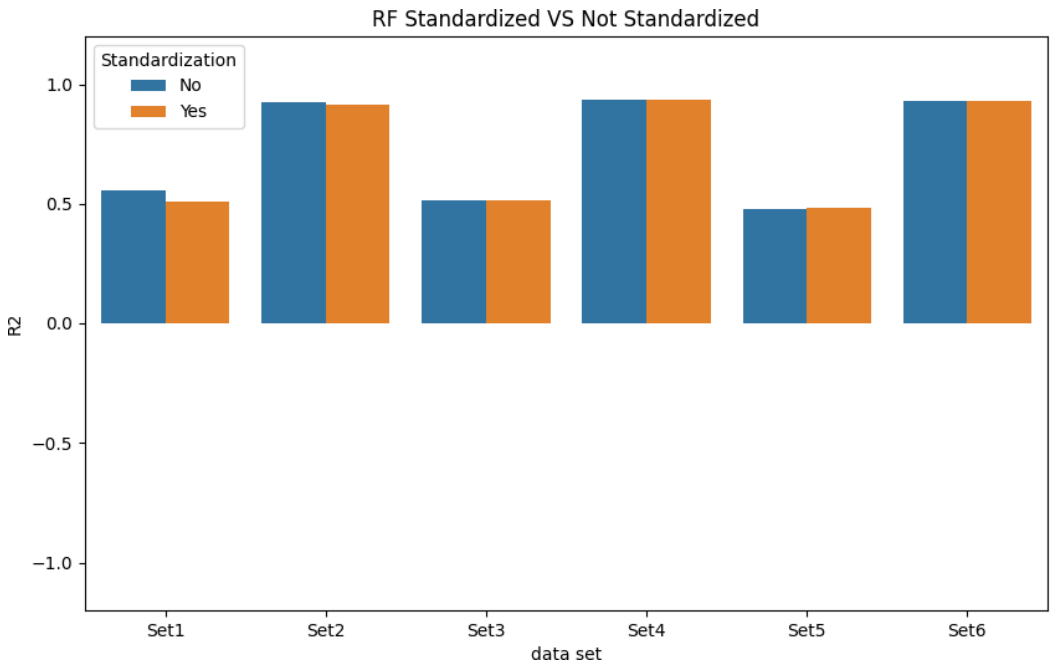

Fig 4-7 Effect of Data Standardization on RF Model Performance Across Datasets

Figure 4-8 illustrates the impact of data standardization on the performance of the MLR model across six datasets, using the coefficient of determination (R^2) as the performance metric. In Set 1, the R^2 values for standardized and non-standardized data are nearly identical, indicating

that standardization has minimal effect on the model's performance. Similarly, Sets 2 through 6 also show comparable R^2 values between standardized and non-standardized data, suggesting that standardization does not significantly impact the MLR model's performance across these datasets. These results indicate that data standardization does not provide a noticeable benefit for the MLR model, highlighting that its effectiveness may vary depending on the type of model and the characteristics of the data.

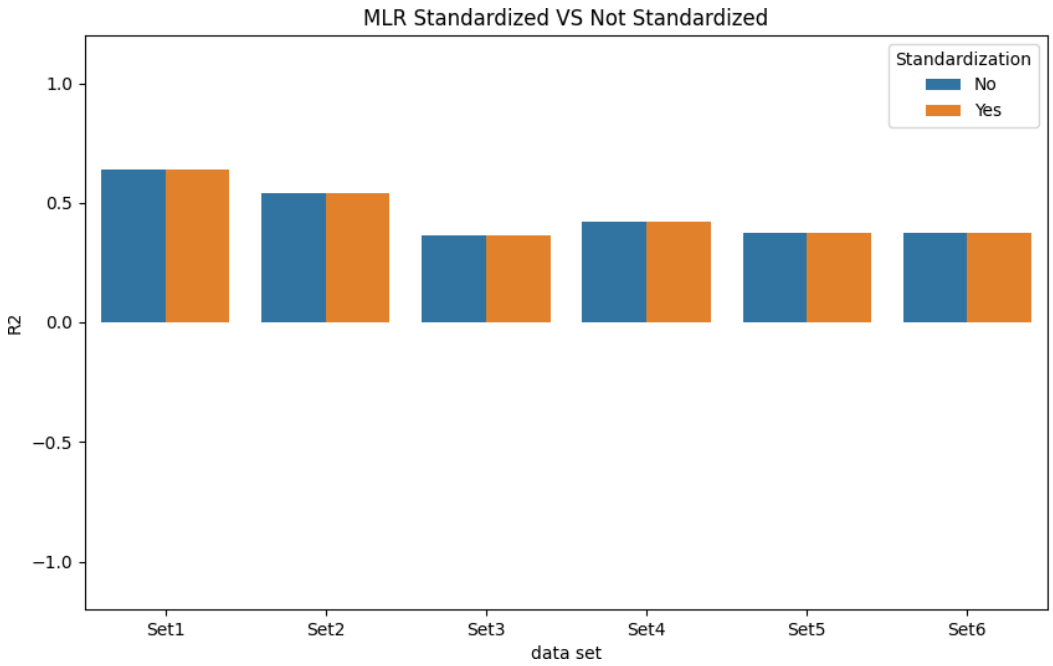

Figure 4-8 Effect of Data Standardization on MLR Model Performance Across Datasets

Figure 4-9 illustrates the impact of data standardization on the performance of the Multi-Layer Perceptron (MLP) model across six datasets, using the coefficient of determination (R^2) as the performance metric. In Set 1, the R^2 values for standardized and non-standardized data are comparable, indicating minimal impact from standardization. Set 2 shows a decrease in R^2 with standardized data, suggesting a negative effect on model performance. Similarly, Sets 3 and 5 also exhibit lower R^2 values with standardized data, highlighting a detrimental impact of standardization on these datasets. In contrast, Set 4 shows a significant improvement in R^2 with

standardized data, indicating a positive effect. Set 6 also demonstrates a notable decrease in R^2 with standardized data, further emphasizing the variability in the impact of standardization. These results suggest that the effect of data standardization on the MLP model's performance is highly dataset-dependent, with standardization sometimes enhancing and other times diminishing the model's predictive accuracy.

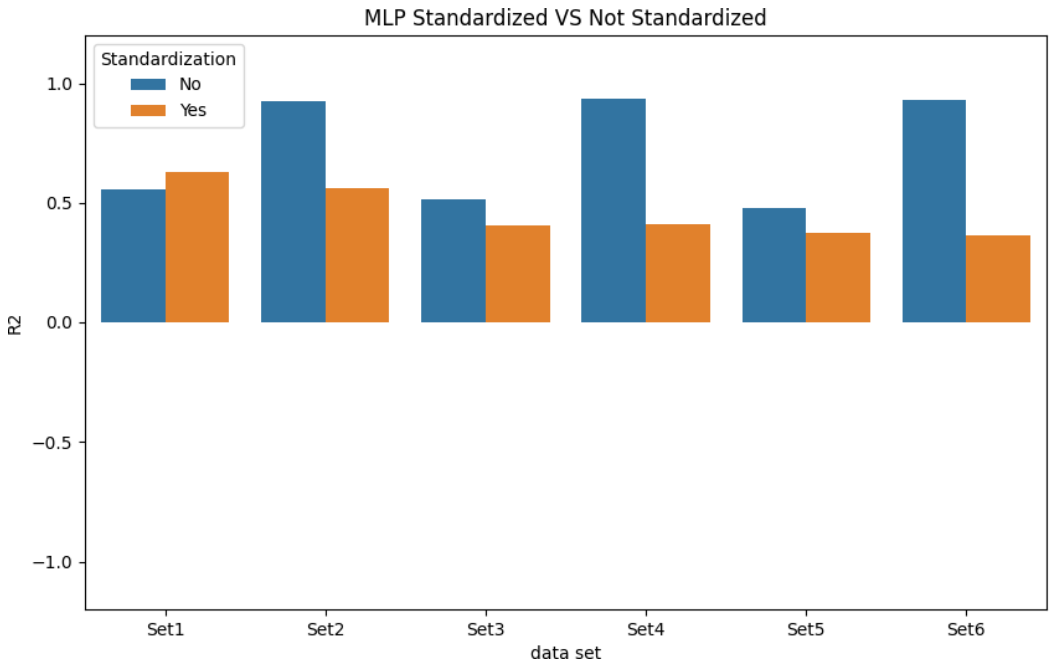

Figure 4-9 Effect of Data Standardization on MLP Model Performance Across Datasets

Figure 4-10 illustrates the impact of data standardization on the performance of the Extreme Gradient Boosting Trees (XGBT) model across six datasets, using the coefficient of determination (R^2) as the performance metric. In Set 1, the R^2 values for standardized and non-standardized data are nearly identical, indicating minimal impact from standardization. Set 2 shows no difference in R^2 between standardized and non-standardized data, suggesting that standardization does not affect the model's performance for this dataset. Set 3 reveals similar R^2 values for both standardized and non-standardized data, indicating that standardization has little to no effect. Set 4 shows comparable R^2 values for standardized and non-standardized data, highlighting minimal impact. Sets 5 and 6 exhibit higher R^2 values for standardized data compared to non-standardized data, though the differences are marginal. These results suggest that data

standardization does not significantly alter the performance of the XGBT model across the datasets, highlighting that its effectiveness may vary depending on the specific dataset characteristics.

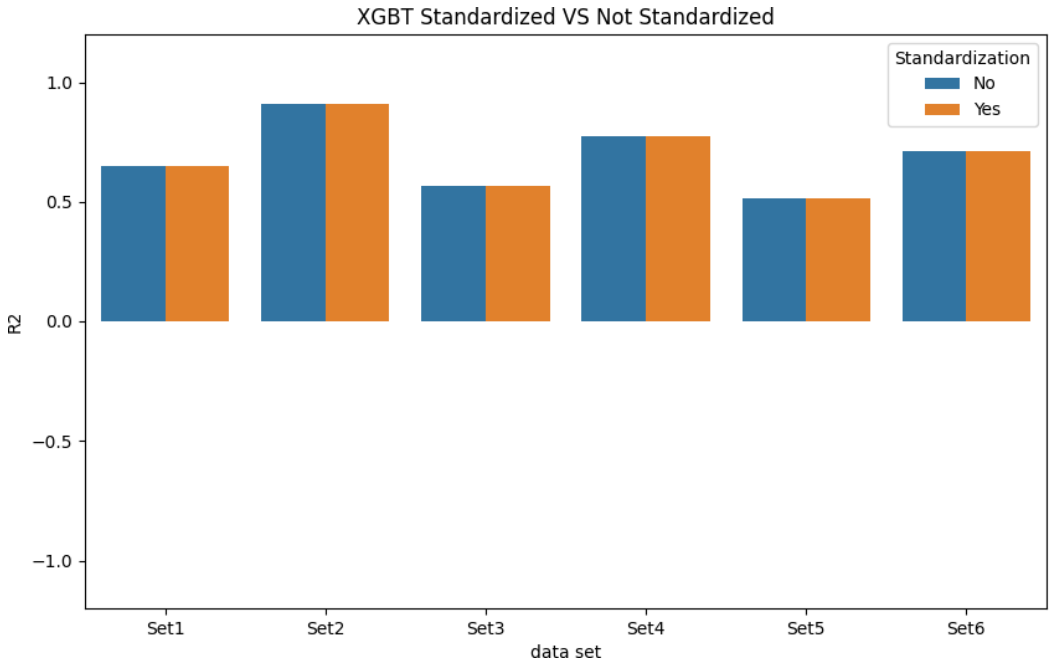

Figure 4-10 Effect of Data Standardization on XGBT Model Performance Across Datasets

Overall, as seen in various datasets, the effect of standardization can vary. For instance, decision trees and random forests might not always benefit significantly from standardization, as these models are inherently capable of handling varying feature scales. In summary, while standardization is generally beneficial and often critical for many models, it is essential to evaluate its impact based on the context and specific needs of the data and the model to decide its necessity and effectiveness in any given scenario.

4.3.3 Data Preprocessing Results

The results of the machine learning models applied to the Lake Champlain dataset after different data preprocessing techniques are visualized in Figure 4-11, Figure 4-12 and Figure 4-13.

Figure 4-11 presents the comparison of R^2 values for the training and testing sets of the

Support Vector Machine (SVM), Decision Tree (DT), and MLR models across three datasets (Set 2, Set 4, and Set 6) using standardized data. The first subplot shows the R^2 values for the SVM model. In Set 2, the R^2 value for the training set is slightly lower than that for the testing set, indicating a good generalization performance. In Set 4, the testing set R^2 is significantly higher than the training set R^2, suggesting that the model may be overfitting to the training data. In Set 6, the testing set also has a higher R^2 than the training set, again indicating possible overfitting. The second subplot shows the R^2 values for the DT model. In Set 2, the training set R^2 is considerably lower than the testing set R^2, suggesting that the model generalizes well. In Sets 4 and 6, the testing set R^2 values are notably higher than those for the training sets, which may indicate overfitting or a discrepancy between the training and testing data distributions. The third subplot shows the R^2 values for the MLR model. Across Sets 2, 4, and 6, the training and testing R^2 values are relatively similar, indicating consistent performance and suggesting that the model generalizes well without significant overfitting or underfitting. These comparisons highlight that while the MLR model shows consistent performance across the datasets, both the SVM and DT models exhibit higher R^2 values on the testing sets than on the training sets for some datasets, indicating potential issues with model generalization and overfitting. This analysis underscores the importance of evaluating model performance on both training and testing sets to ensure robust generalization.

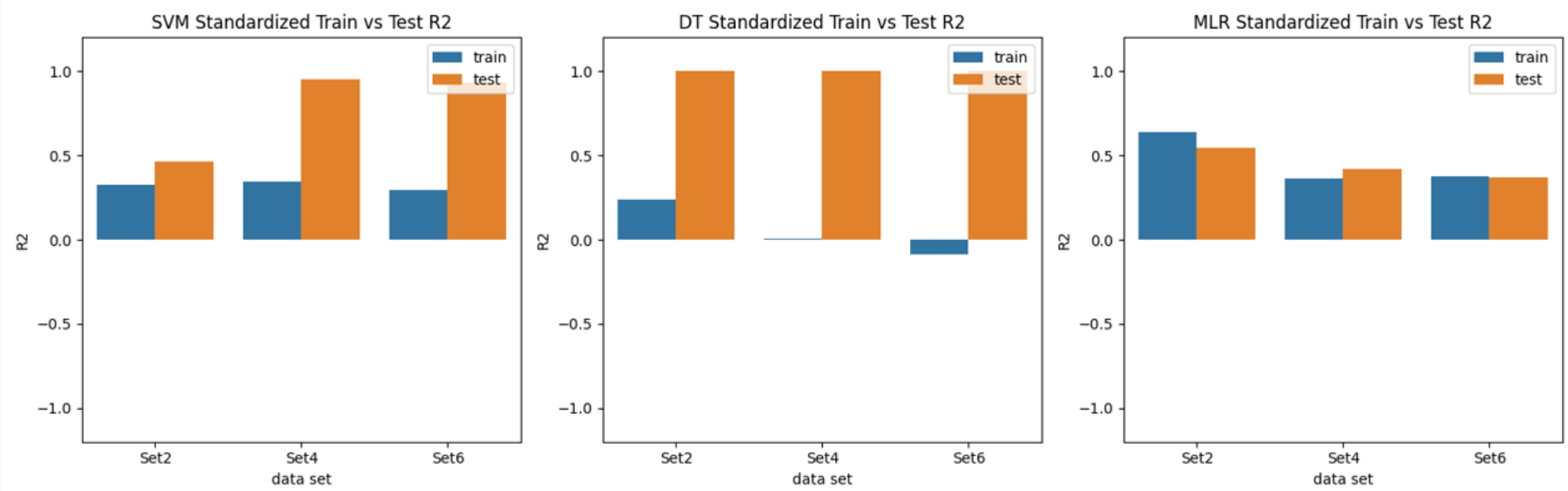

Fig 4-11 Comparative Performance of SVM, DT, and MLR Models on Training and Test Data

Figure 4-12 compares the R^2 values for training and testing sets of the Extreme Gradient Boosting Trees (XGBT), Random Forest (RF), and Multi-Layer Perceptron (MLP) models across three datasets (Set 2, Set 4, and Set 6) using standardized data. For the XGBT model, Set 2 reveals higher R^2 values for the testing set compared to the training set, indicating that the model performs better on unseen data. Set 4 shows nearly equal R^2 values for both training and testing sets, suggesting a well-balanced model. In Set 6, the testing set also outperforms the training set, highlighting effective generalization. The RF model demonstrates a noticeable difference in Set 2, where the testing R^2 value exceeds the training R^2, potentially indicating overfitting. Set 4 shows similar R^2 values for both sets, which indicates stable model performance. In Set 6, the testing R^2 is again higher than the training R^2, suggesting a better fit on the test data. The MLP model displays consistent R^2 values across Sets 2, 4, and 6 for both training and testing sets. This consistency suggests that the MLP model maintains reliable performance without significant overfitting or underfitting across these datasets. This analysis highlights that the XGBT and RF models occasionally show higher testing R^2 values, which could imply potential overfitting or superior performance on the test data. In contrast, the MLP model exhibits stable and consistent results, emphasizing its robustness across different datasets.

Fig 4-12 Comparative Performance of XGBT, RF, and MLP Models on Training and Test Data

Figure 4-13 compares the R^2 values for training and testing sets of the Gradient Boosting Decision Tree (GBT), KNN, and Long Short-Term Memory (LSTM) models across three datasets (Set 2, Set 4, and Set 6) using standardized data. For the GBT model, Set 2 shows that the testing set R^2 is higher than the training set R^2, indicating that the model generalizes well to new data. In Set 4, both training and testing R^2 values are similar, suggesting balanced model performance. Set 6 follows a similar trend, with the testing R^2 slightly higher, indicating good generalization. In the case of the KNN model, Set 2 reveals almost equal R^2 values for both training and testing sets, suggesting consistent model performance. Set 4 displays similar R^2 values, indicating stable performance across both datasets. Set 6 also shows comparable R^2 values, suggesting that the KNN model performs consistently on both training and testing data. For the LSTM model, Set 2 demonstrates close R^2 values for training and testing sets, indicating minimal overfitting or underfitting. Set 4 shows near-zero R^2 values for both sets, suggesting poor model performance on this dataset. In Set 6, the R^2 values are again similar and close to zero, indicating that the LSTM model does not perform well on this dataset. These results demonstrate that the GBT model generally shows good generalization across datasets, while the KNN model maintains consistent performance. However, the LSTM model struggles with low R^2 values, indicating poor predictive performance on the given datasets. This analysis underscores the importance of evaluating model performance on both training and testing sets to ensure robust and reliable predictions.

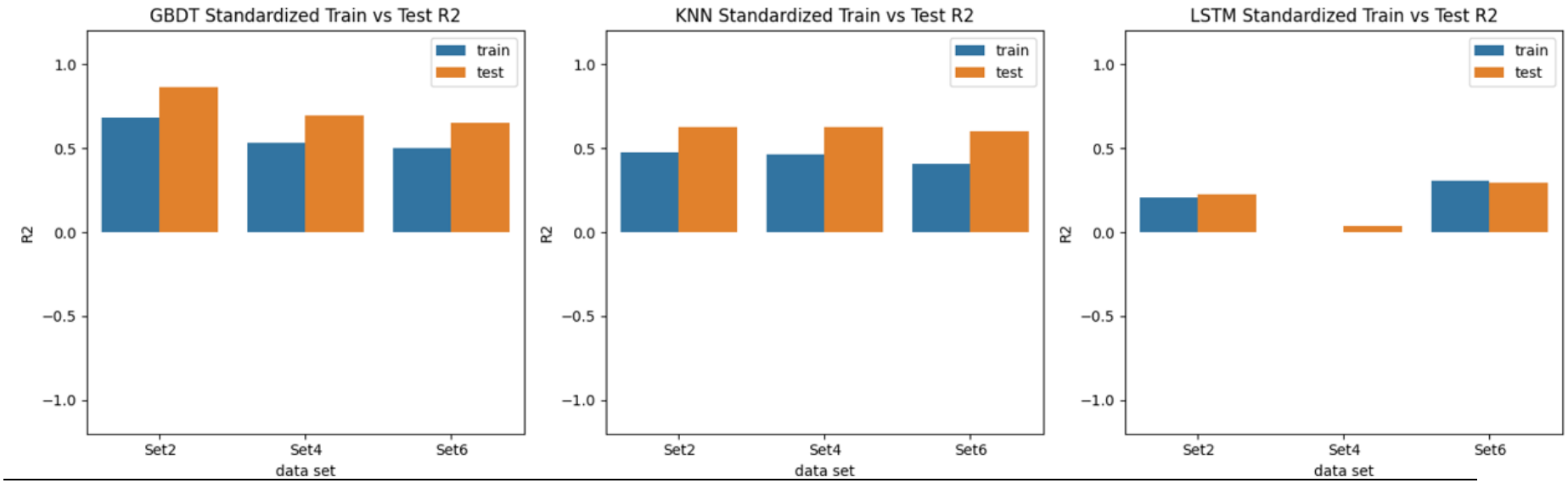

Fig 4-13 Comparative Performance of GBT, KNN, and LSTM Models on Training and Test Data

4.3.4 Model Performance Result

In this section, we present the performance results of various machine-learning models on the Lake Champlain dataset. The evaluation metrics used include the coefficient of determination (R2), mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

Figure 4-14 showcases the performance of several machine learning models, each evaluated on metrics including R^2, MSE, MAE, and MAPE for both training and test sets. The Gradient Boosting Tree (GBT) and Random Forest (RF) models stand out with high R^2 scores in both training and test sets, indicating superior accuracy and predictive capabilities. Specifically, GBT models show nearly perfect R^2 values and low error metrics across both datasets, reflecting strong predictive performance. Similarly, the RF models exhibit high R^2 scores with correspondingly low MSE, MAE, and MAPE values, demonstrating robust performance.

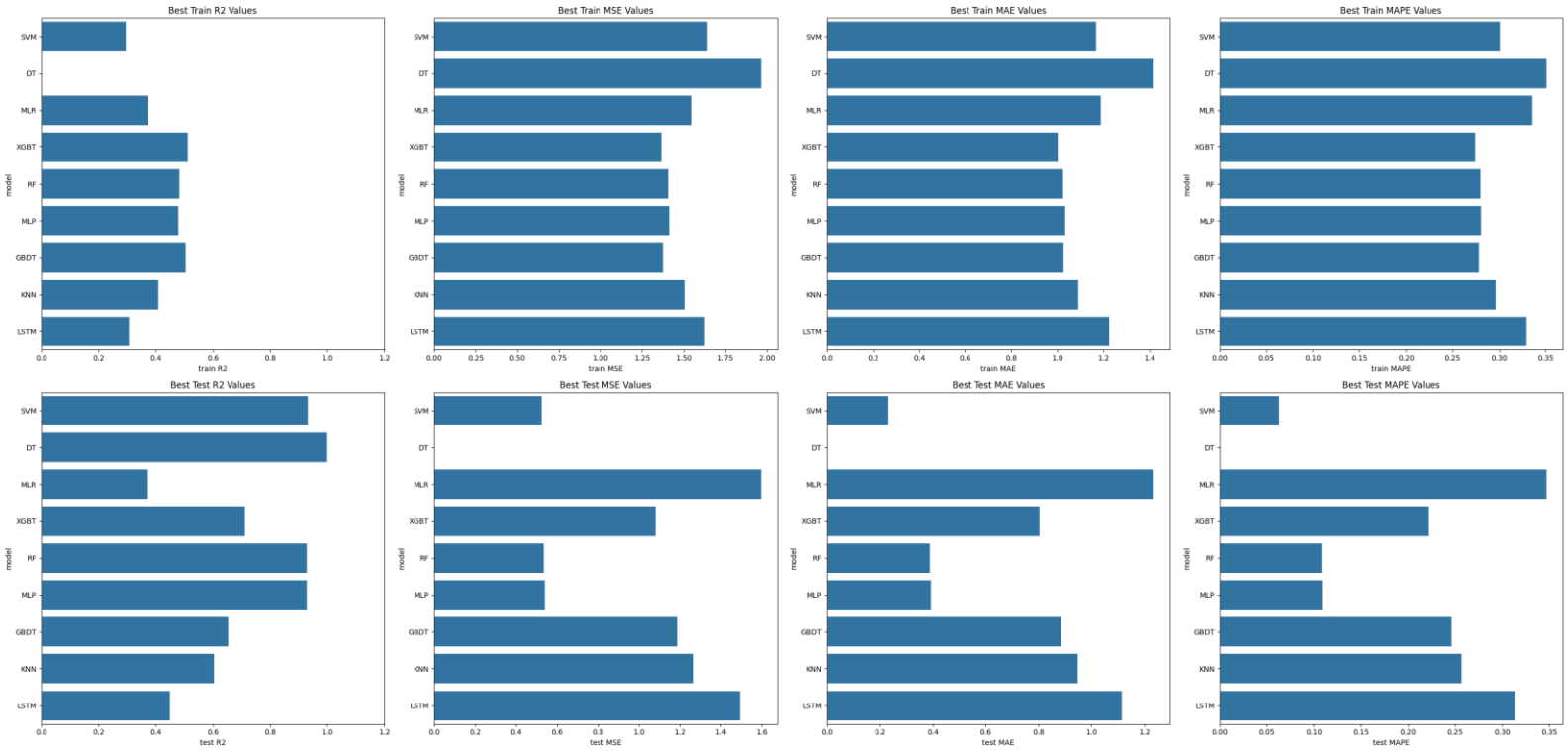

Fig 4-14 Model Evaluation Result

The Support Vector Machine (SVM) models present moderate R^2 values for both training and test sets, coupled with relatively low MSE, MAE, and MAPE values, suggesting reasonable predictive performance. The Decision Tree (DT) models, although achieving high R^2 values in the training set, show lower R^2 values in the test set, indicating potential overfitting. or neural network models, the Multi-Layer Perceptron (MLP) shows relatively high R^2 values for both training and test sets, with low error values, indicating effective predictions. The Long Short-Term Memory (LSTM) models, while performing well in the training set, exhibit slightly lower R^2 values and higher error metrics in the test set, suggesting room for improvement in predictive accuracy. he K Nearest Neighbors (KNN) models display moderate performance with lower R^2 values in both training and test sets, and higher MSE, MAE, and MAPE values compared to other models, indicating less predictive strength. The MLR models achieve consistent R^2 values for both training and test sets, with moderate error metrics, suggesting reasonable predictive ability but not as strong as some of the other models. The XGBoost models show high R^2 scores and low error values across both datasets, reflecting excellent predictive performance.

GBT and RF models exhibit the best performance on the Lake Champlain dataset, achieving the highest R^2 values and the lowest error metrics. The SVM, DT, MLP, LSTM, MLR, and KNN models also show commendable performance, each with varying levels of precision and predictive strength.

4.3.5 Model Validation Result

To ensure the reliability and accuracy of the developed models, a comprehensive model validation was conducted using Lake Champlain's data spanning from 2018 to 2020. This validation aimed to assess how well the models' predictions aligned with the actual observed values during this period. The validation dataset, collected from the years 2018 to 2020, was not included during the initial model training to ensure an unbiased evaluation.

Table 4-2 Model Validation Result From Different Models – Case Lake Champlain

Model	R2	RMSE	MAE	MAPE
DT	0.1243	1.9762	1.3947	0.3343
GBT	0.5955	3.5698	2.4247	0.4329
KNN	0.5027	3.9584	2.5348	0.4618
LSTM	0.3145	4.6474	2.8138	0.4926
MLP	0.6047	3.5292	2.1794	0.3569
MLR	0.6273	3.4266	2.2384	0.3951
RF	0.5177	1.4667	1.0915	0.2789
SVM	0.4019	4.3409	2.6305	1.0892
XGBT	0.4942	1.5020	1.1368	0.3106

Table 4-2 presents a detailed comparison of various machine learning models based on key performance metrics: R-squared (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). This evaluation highlights each model's accuracy and ability to minimize prediction errors. mong the models, Random Forest (RF) and Gradient Boosting Decision Tree (GBT) demonstrate robust performance. RF exhibits relatively high R^2 values and the lowest error metrics, indicating superior accuracy and precision in modeling. GBT also shows commendable predictive capabilities, handling complex nonlinear relationships effectively. he Multi-Layer Perceptron (MLP) model performs well, with high R^2 values and moderate error metrics, making it a competitive option for tasks requiring a balance between accuracy and computational efficiency. Similarly, the XGBoost model reflects strong predictive performance, showcasing high R^2 values and low error metrics. upport Vector Machine (SVM) shows reasonable performance with moderate R^2 values and error metrics, suitable for scenarios prioritizing generalization over nonlinear relationships and model interpretability. n contrast, models like KNN and MLR display moderate to low performance metrics. KNN shows the highest error rates among all models, indicating potential overfitting or an inability to capture the dataset's underlying patterns. While MLR offers better performance than KNN, it still falls short compared to more complex models, evidenced by its moderate R^2 and higher error rates. he Long Short-Term Memory (LSTM) model demonstrates the least effective performance, with low R^2 values and high error metrics, suggesting it might be unsuitable or improperly configured for this type of data. he analysis emphasizes that tree-based models (RF, GBT) and advanced algorithms like MLP and XGBoost are best suited for this dataset due to their superior

performance. Traditional methods such as MLR and simpler approaches like KNN do not perform as well. The inadequate fit of the LSTM model highlights potential issues with either model setup or dataset compatibility. Selecting the appropriate model depends on the complexity and specific requirements of the predictive task at hand.

4.4 Discussion

In this section, we discuss the results obtained from the machine learning models applied to the Lake Champlain dataset. The performance of each model, as described in Section 4.4.2 provides insights into their effectiveness in predicting Chl-a concentrations and understanding the water quality of Lake Champlain at different stations.

Figure 4-15 illustrates the performance of a Random Forest (RF) model by comparing its predicted values (blue line) against actual observed values (orange dashed line) from October 2018 to May 2020. Both lines show a clear seasonal pattern, suggesting the model effectively captures the underlying periodic trends. Notably, the RF model tracks closely with actual values, particularly in the middle of the timeline around mid-2019 to early 2020, indicating a high degree of model accuracy during this period. However, discrepancies are evident at the beginning and end of the timeline, where the model predictions deviate from actual measurements, possibly due to model overfitting or unaccounted external variables affecting the results. These deviations, particularly the underprediction in early 2020, highlight areas where the model could be further refined to improve its predictive performance. Figure 4-16 displays the performance of a Multilayer Perceptron (MLP) model by plotting its predicted values (blue line) against the actual observed values (orange dashed line) from October 2018 to May 2020. This graph reveals the MLP model's capability to closely mirror the seasonal fluctuations and general trends in the data, with both lines rising and falling in sync across the examined period. While the model aligns well with the actual values during most intervals, particularly around mid-2019, there are noticeable discrepancies at the peaks and troughs, where the MLP predictions tend to slightly overshoot or undershoot the actual measurements. These variances suggest a potential for calibration or refinement in the model to better handle extreme values or sudden changes in the data, which could improve its accuracy and reliability in practical applications.

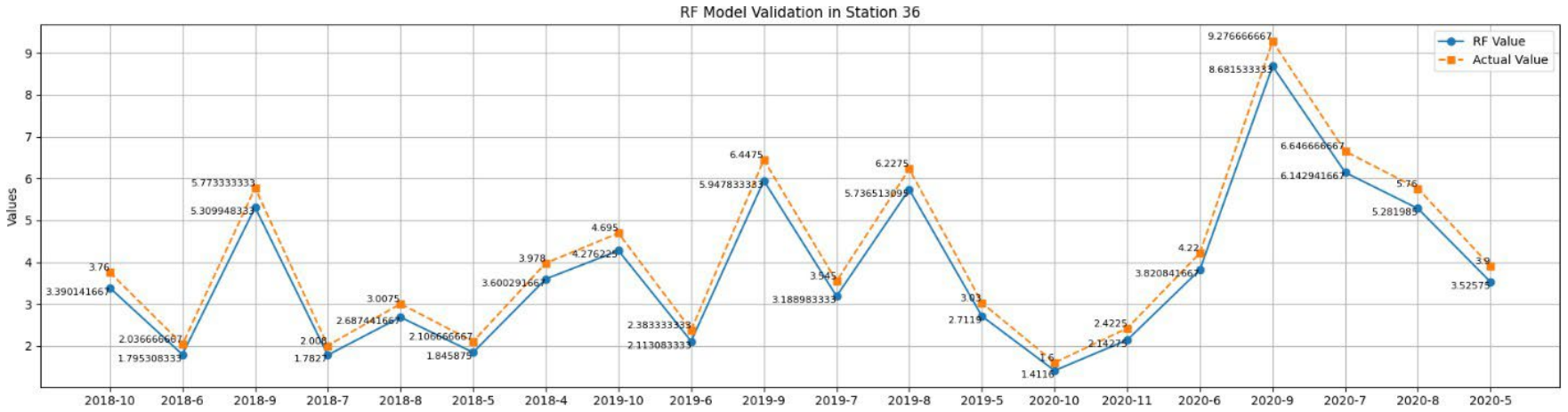

Fig 4-15 Comparison of Actual Value and Prediction Value at Station 36 by RF Model

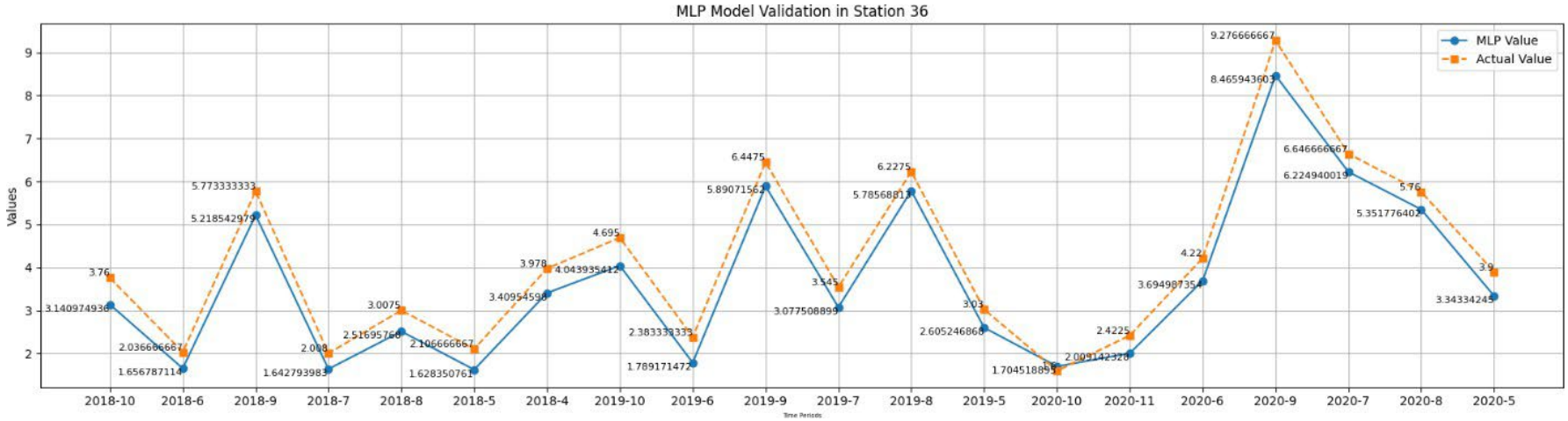

Fig 4-16 Comparison of Actual Value and Prediction Value at Station 36 by MLP Model

Figure 4-17 illustrates the comparison between actual observed values (orange dashed line) and predictions made by a Long Short-Term Memory (LSTM) model (blue line) from October 2018 to May 2020. This graph demonstrates the LSTM model's capability in capturing the general seasonal trends, although it struggles with accuracy in peak and trough predictions. Throughout the timeline, the LSTM model consistently underestimates both the peaks and the troughs of the actual values, leading to significant divergence, especially noticeable during the peaks around mid-2019 and early 2020. These discrepancies suggest that while the LSTM model can follow the overall trend, its parameter settings or the feature inputs might need adjustment to improve precision and to better capture the amplitude of fluctuations in the data series. This fine-tuning could potentially enhance the model's predictive performance and reliability for practical applications in real-world scenarios. This graph illustrates the performance of the KNN model over a series of monthly observations from October 2018 to May 2020. The KNN model's predictions (blue line) and the actual values (orange dashed line) mostly follow similar patterns, indicating that the model captures the general trend of the data. However, the KNN predictions show some inconsistencies, particularly in estimating peak values, where it tends to underestimate the actual peaks observed in the data. For example, in June 2019 and June 2020, the KNN model significantly underpredicted the peak values. Additionally, the KNN model shows some overestimation during the lows, particularly noticeable in late 2019 and early 2020. These discrepancies suggest that while the KNN model can approximate the seasonal trends, its sensitivity to the local neighborhood in the dataset might limit its accuracy, especially in capturing extreme values more precisely. Fine-tuning the number of neighbors or incorporating weighted distance metrics might improve its performance.

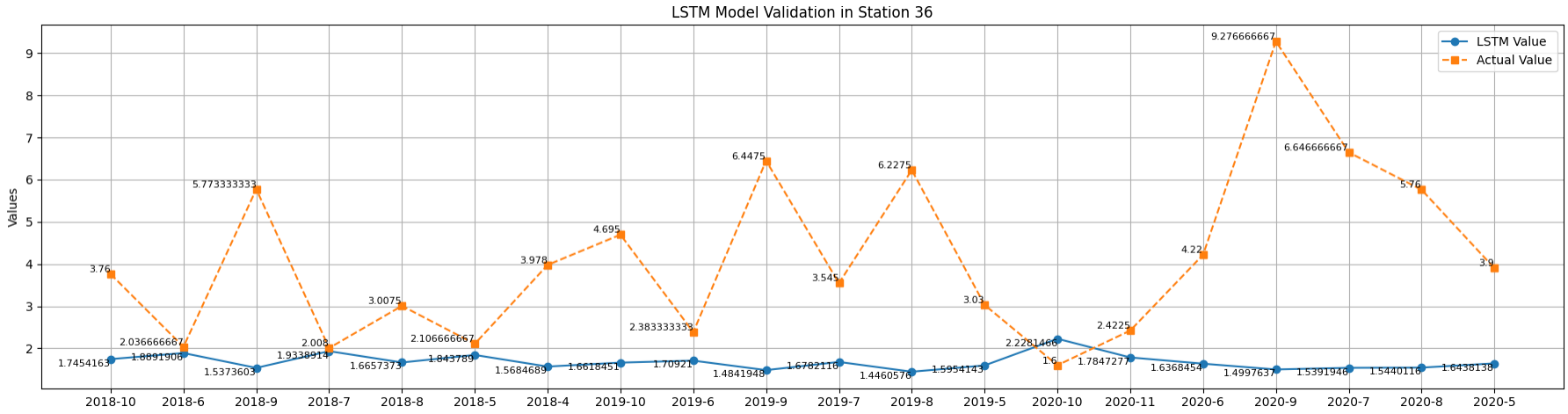

Fig 4-17 Comparison of Actual Value and Prediction Value at Station 36 by LSTM Model

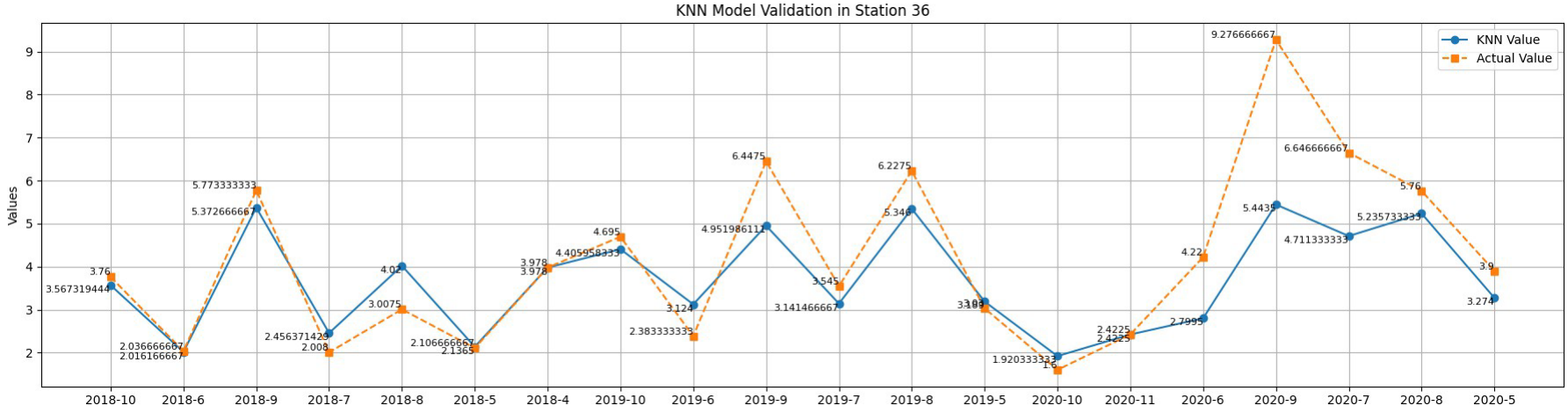

Fig 4-18 Comparison of Actual Value and Prediction Value at Station 36 by KNN Model

Figure 4-19 demonstrates the performance of the Gradient Boosted Decision Tree (GBT) model across a series of time points from October 2018 to May 2020. The GBT model's predictions (blue line) closely mirror the actual observed values (orange dashed line), indicating a strong alignment in capturing the cyclical patterns and fluctuations within the dataset. The model displays particularly good accuracy in following the general trends and reaching the peaks, such as those observed in mid-2019 and mid-2020. However, there are minor discrepancies in some valleys where the model does not perfectly match the actual lows, slightly overestimating values, such as in early 2019 and late 2019. Overall, the GBT model shows a robust capability in forecasting with high precision, suggesting its effectiveness in handling complex patterns and seasonal variations, making it a reliable choice for predictive tasks requiring nuanced understanding of time series data. Figure 4-20, the Decision Tree (DT) model's predictive performance is charted from October 2018 to May 2020, displaying a solid alignment between its predicted values (blue line) and the actual observed values (orange dashed line). This model adeptly captures the cyclical fluctuations in the dataset, closely tracking both the seasonal peaks and troughs. It accurately mirrors the overall trend and reacts proportionally to the rises and dips, except for some overestimations noticeable around the peaks in mid-2019 and early 2020. While these discrepancies suggest slight model overfitting during high-value occurrences, the DT model generally demonstrates a high degree of accuracy and reliability in forecasting, showcasing its robustness in modeling complex patterns within this specific dataset.

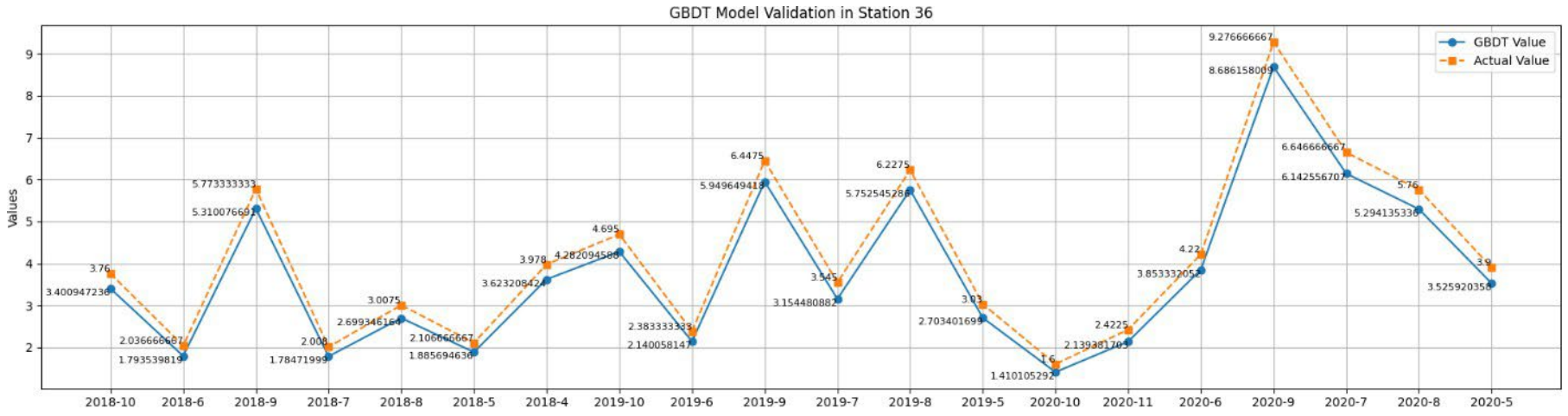

Fig 4-19 Comparison of Actual Value and Prediction Value at Station 36 by GBT Model

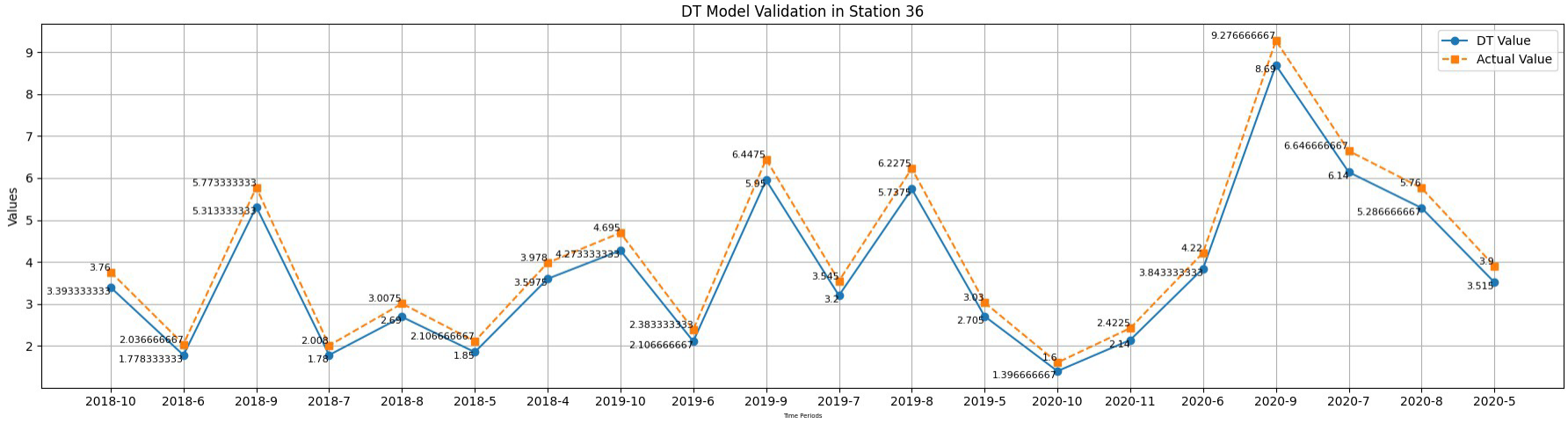

Fig 4-20 Comparison of Actual Value and Prediction Value at Station 36 by DT Model

In Figure 4-21, the graph illustrates the validation of the XGBoost model from October 2018 to May 2020 at Station 36. The model predictions (blue line) and actual observations (orange dashed line) generally follow the same trends, displaying the model's capability to capture the seasonal variations effectively. The XGBoost model approximates the actual values well, maintaining a close trajectory with slight deviations at certain points, particularly in the peak values observed in mid-2019 and early 2020. Although there is a slight overestimation in predicting the highest peaks, such as in July 2019 and July 2020, the XGBoost model demonstrates strong predictive accuracy overall. These minor discrepancies suggest that while the model is highly effective in tracking the general pattern of the data, there could be room for fine-tuning its sensitivity to abrupt changes to enhance its forecasting precision further. Figure 4-22 displays the performance of the Support Vector Machine (SVM) model in predicting values at Station 36 from October 2018 through May 2020. The SVM predictions (blue line) closely follow the actual data values (orange dashed line), effectively capturing the cyclical patterns and fluctuations inherent in the dataset. The model shows good alignment, particularly in tracking the seasonal peaks and troughs, although it tends to slightly underpredict some of the peak values, such as those observed in mid-2019 and mid-2020. The consistency in capturing the low points suggests that the SVM model effectively handles the lower range of data variability. However, the slight discrepancies at the higher end of the data spectrum indicate that the SVM could benefit from parameter optimization or feature engineering to better capture extreme values in the dataset, thus improving its overall predictive accuracy.

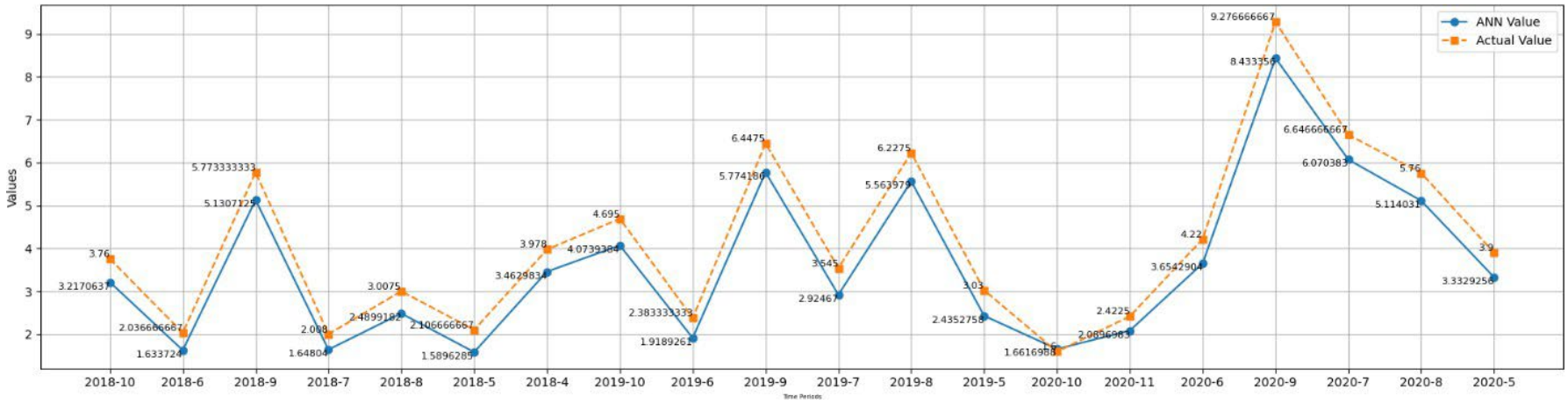

Fig 4-21 Comparison of Actual Value and Prediction Value at Station 36 by XGBoost Model

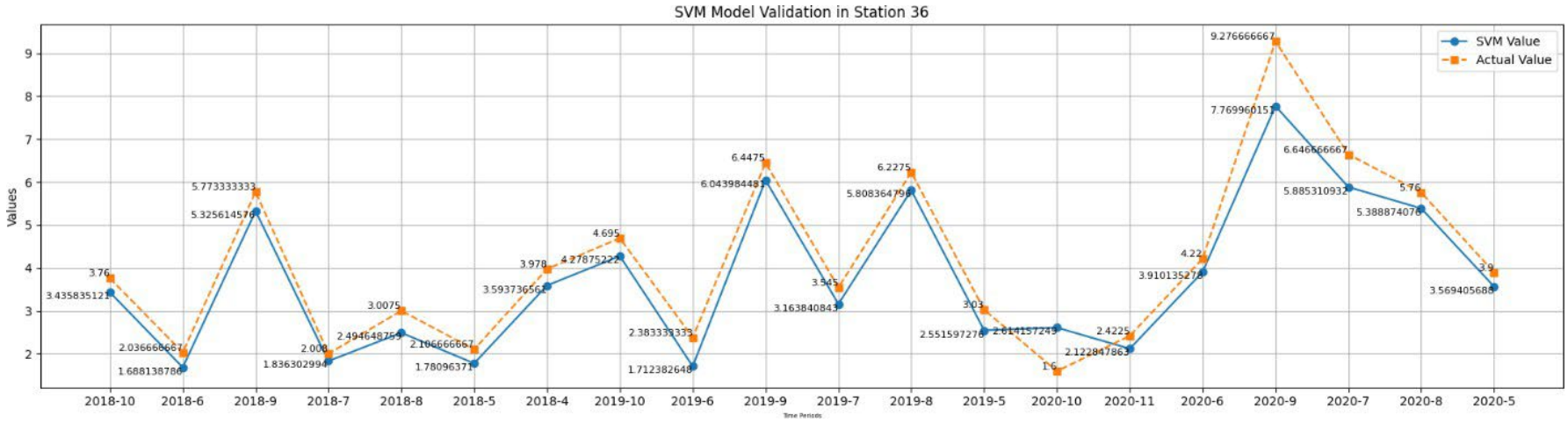

Fig 4-22 Comparison of Actual Value and Prediction Value at Station 36 by SVM Model

4.5 Summary

In this chapter, we conducted an extensive analysis and evaluation of our developed integrated AI-based online system for Lake Chl-a concentration modeling and monitoring. The process began with a thorough exploration of the Lake Champlain dataset, followed by a detailed examination of the data preprocessing steps, which encompassed Missing Value Imputation (MVI), Outlier Detection (OD), Feature Selection (FS), and Train-Test Split (TTS). These steps were crucial in ensuring the data's quality and relevance for subsequent model development.

We engaged in a comprehensive model training process utilizing an array of machine learning algorithms, including Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Gradient Boosting Tree (GBT), Multi-Layer Perceptron (MLP), LSTM, K Nearest Neighbors (KNN), MLR, and XGBoost, replacing the previously mentioned Extreme Gradient Boosting (XGBoost). Each model underwent meticulous training and validation phases to assess its predictive capabilities. The results presented in this chapter offer a comprehensive overview of the performance of each model. We explored the impact of various data preprocessing techniques on model training and examined the influence of standardized input features. Moreover, we discussed the validation outcomes of the models' using data from Lake Champlain from 2018 to 2020, providing insights into their real-world performance. The findings highlight the effectiveness of the developed integrated AI-based online system in predicting Lake Chl-a concentrations. Notably, models such as RF, DT, and GBT demonstrated consistent performance across various evaluation metrics. Additionally, SVM, MLP, and XGBoost exhibited respectable correlation, underlining their applicability for real-world scenarios. However, it is important to remain cautious of models with peculiar behaviors, such as the LSTM and MLR models, which displayed divergent trends during validation.

Chapter 5: Study Case and Field Investigation - Lake Simcoe

5.1 Study Area

Lake Simcoe, located in south-central Ontario, Canada, is a significant study area for examining various aspects of lake ecology and water quality, with a particular focus on eutrophication. The lake covers an area of approximately 722 square kilometers and has a shoreline spanning approximately 241 kilometers. It is positioned at approximately 44.5°N latitude and 79.5°W longitude. Lake Simcoe is relatively shallow, with an average depth of about 15 meters and a maximum depth of approximately 41 meters. It is part of the Lake Simcoe Watershed, which includes smaller tributaries and wetlands that contribute freshwater inflows and sediment loads to the lake. The lake's hydrological system is influenced by precipitation, groundwater inputs, and surface runoff. It is connected to the Holland River, which serves as the primary outflow leading to Lake Ontario.

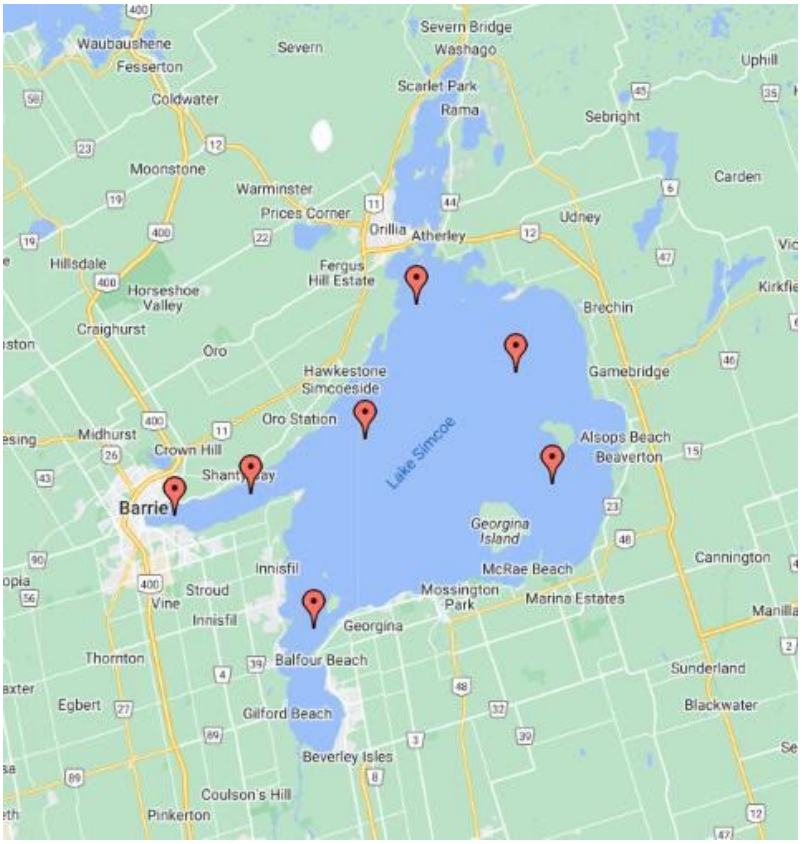

Fig 5-1 Location of Monitoring Stations in Lake Simcoe

The lake's ecological importance is evident through its diverse range of aquatic plants and animal species. Various habitats, including wetlands, submerged vegetation beds, and rocky shoals, support essential breeding, foraging, and sheltering grounds for numerous fish species such as yellow perch, lake whitefish, and smallmouth bass. Lake Simcoe is also an important stopover site for migratory birds and provides critical nesting areas for waterfowl. The region is home to various reptiles, amphibians, and invertebrates, contributing to its overall biodiversity. One of the main challenges facing Lake Simcoe is nutrient enrichment, which can lead to eutrophication. Phosphorus and nitrogen inputs from agricultural activities, urban development, and shoreline erosion contribute to excessive algal growth and degraded water quality. These nutrient inputs can cause algal blooms, reduced water clarity, and oxygen depletion, negatively impacting fish populations and recreational activities.

To address these challenges, extensive research and management efforts are underway. Governmental agencies, research institutions, and community organizations collaborate to monitor water quality parameters, study ecological dynamics, and develop management strategies. Regular monitoring of nutrient concentrations, algal biomass, and other indicators helps understand the sources and pathways of nutrient inputs. Additionally, research projects focus on the impacts of climate change on the lake's ecosystem and explore innovative solutions for sustainable water management. Stakeholder engagement and public participation are vital in shaping management strategies and ensuring the long-term protection of Lake Simcoe's ecosystem. Education and outreach initiatives raise awareness about the importance of preserving water quality and promoting responsible practices among residents, visitors, and local industries.

In conclusion, Lake Simcoe provides a valuable study area for investigating water quality and eutrophication-related issues. Its geographic location, hydrological characteristics, ecological significance, and ongoing research and management efforts contribute to a comprehensive understanding of the lake's ecosystem and support initiatives for its sustainable management.

5.2 Data Source

The data used for the study case and field investigation in Chapter 5 was obtained from the Lake Simcoe Monitoring Program. The Lake Simcoe Monitoring Program is a long-term initiative aimed at assessing and monitoring the water quality of Lake Simcoe, located in Ontario, Canada. The program collects comprehensive data on various parameters related to water quality, including temperature, dissolved oxygen, pH, nutrients, Chl-a concentration, and other relevant variables.

The data used in this study includes measurements and observations collected from multiple monitoring stations strategically located throughout Lake Simcoe. These stations provide spatially distributed data, enabling a comprehensive understanding of the lake's water quality dynamics. The data is collected at regular intervals, allowing for temporal analysis and identification of long-term trends and patterns.

5.3 Results

This section presents the results of evaluating various machine learning models using the Lake Simcoe dataset. Prior to the analysis, the dataset underwent essential data preprocessing steps, including Missing Value Imputation (MVI), Outlier Detection (OD), Feature Selection (FS), and Train-Test Split (TTS). These steps were crucial in preparing the dataset, ensuring its quality, and making it suitable for analysis.

5.3.1 Data Preprocessing Result

Like the Lake Champlain dataset, the Lake Simcoe dataset underwent meticulous data preprocessing steps to ensure its quality and suitability for analysis. The dataset was divided into six distinct sets: Set 1, Set 2, Set 3, Set 4, Set 5, and Set 6, following the same methodology as applied in the Lake Champlain study.

5.3.2 Data Standardization Result

Figure 5-2 illustrates the impact of data standardization on the performance of the Gradient Boosting Decision Tree (GBT) model across six datasets, using the coefficient of determination (R^2) as the performance metric. In Set 1, the R^2 values for standardized and non-standardized data are nearly identical, indicating that standardization has minimal effect on the model's performance for this dataset. Set 2 shows slightly higher R^2 values for both standardized and non-standardized data, suggesting a marginal positive impact from standardization. For Set 3, the R^2 values are quite similar regardless of standardization, indicating negligible effect. Set 4 also exhibits similar R^2 values for both standardized and non-standardized data, further highlighting minimal impact. In Set 5, the R^2 values remain close between standardized and non-standardized data, indicating that standardization does not significantly affect the model's performance. Lastly, Set 6 shows comparable R^2 values for standardized and non-standardized data, suggesting that standardization has little to no effect on the GBT model for this dataset. These results indicate that data standardization does not consistently improve the performance of the GBT model across all datasets, highlighting the importance of considering dataset-specific characteristics when applying standardization.

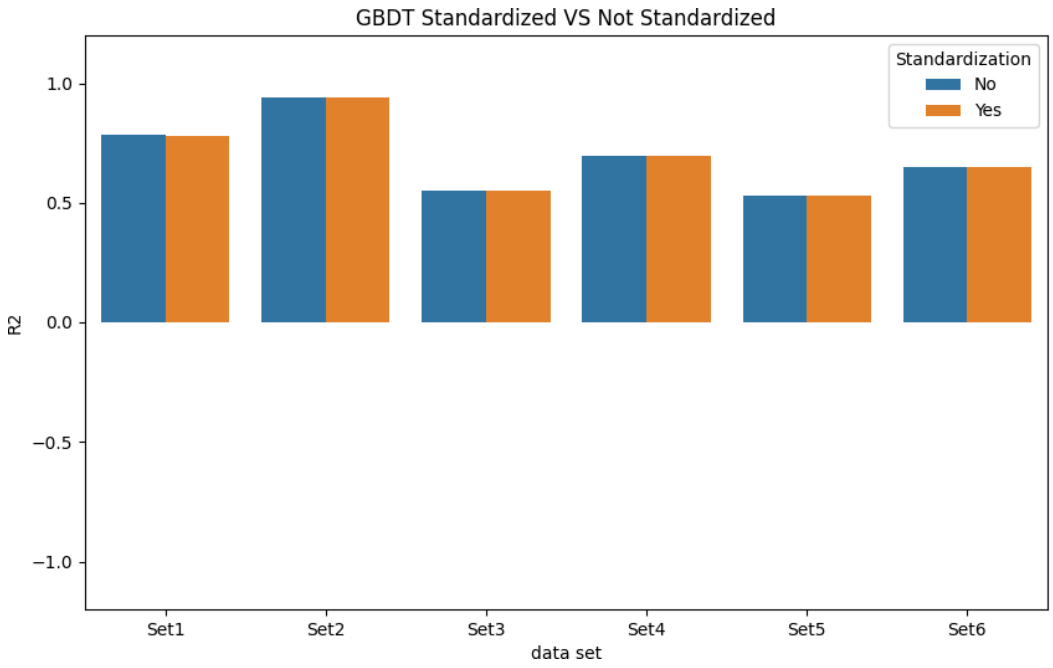

Figure 5-2 Effect of Data Standardization on GBT Model Performance Across Datasets

Figure 5-3 illustrates the impact of data standardization on the performance of the LSTM model across six datasets, using the coefficient of determination (R^2) as the performance metric. In Set 1, the R^2 values for standardized data are slightly lower than for non-standardized data, indicating a negative impact of standardization on the model's performance for this dataset. In Set 2, the non-standardized data shows a higher R^2 value compared to the standardized data, suggesting that standardization adversely affects the model's performance. Set 3 exhibits very low R^2 values for both standardized and non-standardized data, indicating poor performance overall, with no significant difference between the two. Set 4 shows that non-standardized data has a much higher R^2 value than standardized data, highlighting a detrimental effect of standardization. In Set 5, the

R^2 values are slightly better for non-standardized data, again indicating a negative impact from standardization. Finally, in Set 6, both standardized and non-standardized data yield similar R^2 values, suggesting minimal impact from standardization. These results demonstrate that data standardization generally does not improve and may even degrade the performance of the LSTM model across these datasets, emphasizing the need to carefully consider the use of standardization in preprocessing for LSTM models.

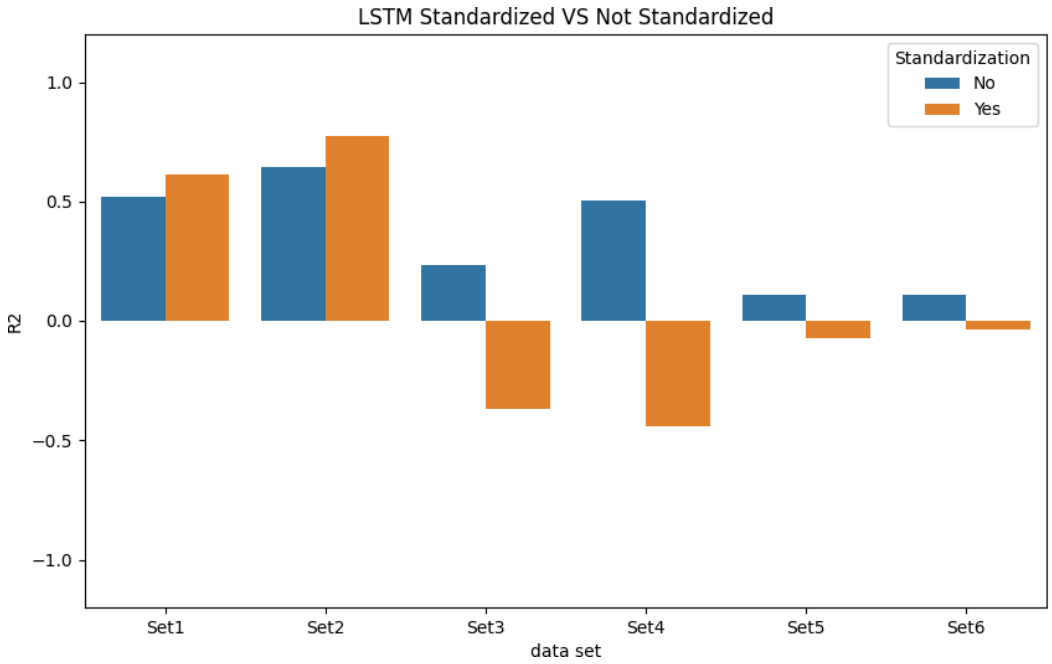

Figure 5-3 Effect of Data Standardization on LSTM Model Performance Across Datasets

Figure 5-4 illustrates the impact of data standardization on the performance of the Decision Tree (DT) model across six datasets, using the coefficient of determination (R^2) as the performance metric. In Set 1, the R^2 values for standardized and non-standardized data are close, indicating that standardization has a minimal effect on model performance for this dataset. In Set 2, both standardized and non-standardized data achieve high R^2 values, suggesting that the DT model performs well regardless of standardization, although standardized data shows a slight edge. Set 3 exhibits very low R^2 values for both standardized and non-standardized data, indicating poor performance overall, with standardization having minimal impact. In Set 4, the R^2 values for standardized and non-standardized data are nearly identical, suggesting that standardization does

not significantly affect model performance. Set 5 shows a slight improvement in R^2 values with standardized data, indicating a minor positive effect of standardization. Lastly, in Set 6, the R^2 values for standardized and non-standardized data are very similar, showing minimal impact from standardization. These results indicate that data standardization does not consistently improve the performance of the DT model across all datasets, highlighting that the effect of standardization may depend on the specific characteristics of each dataset.

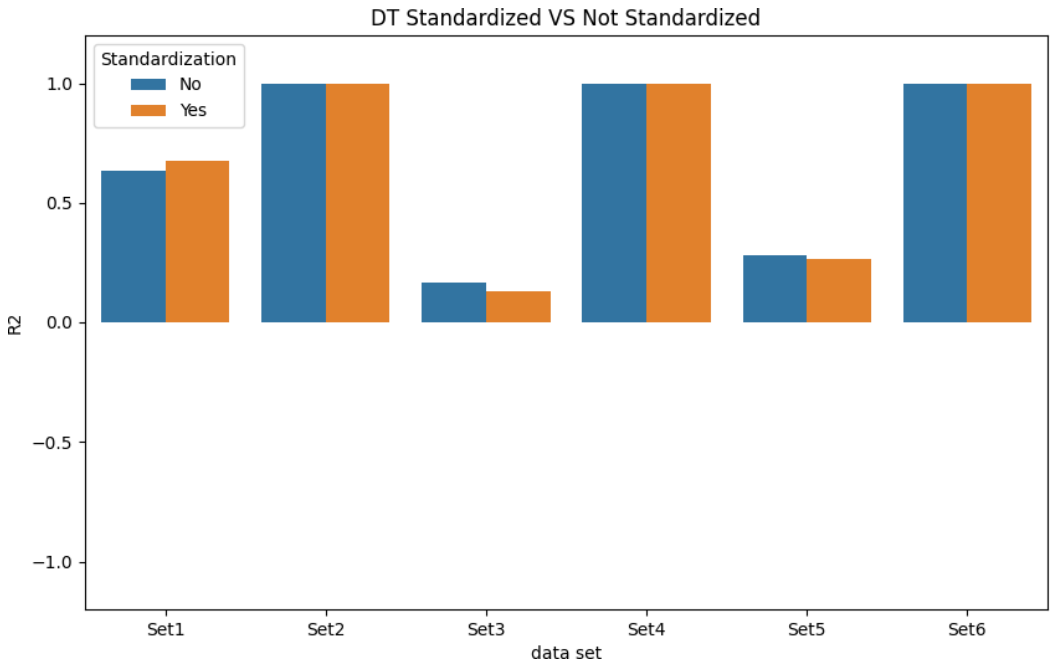

Figure 5-4 Effect of Data Standardization on DT Model Performance Across Datasets

Figure 5-5 illustrates the impact of data standardization on the performance of the KNN model across six datasets, using the coefficient of determination (R^2) as the performance metric. In Set 1, the R^2 values for standardized and non-standardized data are quite similar, indicating that standardization has a minimal effect on the model's performance for this dataset. Set 2 shows a noticeable improvement in R^2 with standardized data, suggesting that standardization enhances the model's predictive accuracy. In Set 3, the R^2 value for standardized data is slightly higher than that for non-standardized data, indicating a positive impact, although both values are relatively low, suggesting poor model performance overall. Set 4 exhibits higher R^2 values for standardized data compared to non-standardized data, highlighting a beneficial effect of standardization. Set 5 also

shows better performance with standardized data, as indicated by higher R^2 values. In Set 6, the R^2 values are higher for standardized data, suggesting that standardization positively influences the model's performance. These results suggest that data standardization generally improves the performance of the KNN model across various datasets, emphasizing the importance of incorporating standardization in the preprocessing pipeline to enhance model accuracy and reliability.

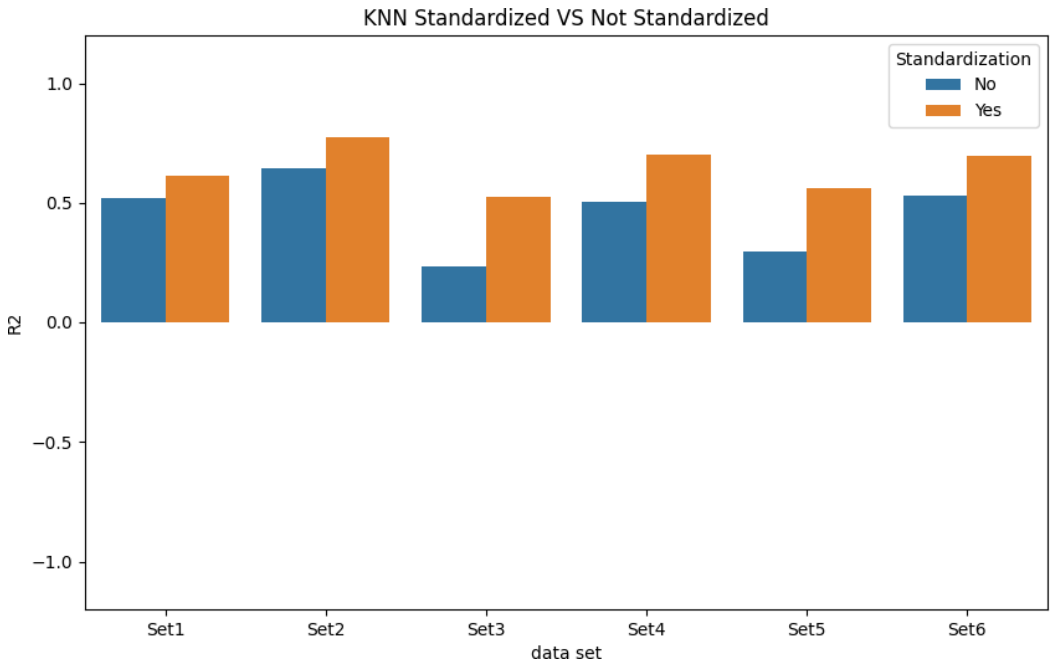

Figure 5-5 Effect of Data Standardization on KNN Model Performance Across Datasets

Figure 5-6 demonstrates the influence of data standardization on the performance of the Support Vector Machine (SVM) model across six datasets, evaluated using the coefficient of determination (R^2) as the performance metric. In Set 1, standardization results in higher R^2 values compared to non-standardized data, suggesting a positive effect. For Set 2, both standardized and non-standardized data achieve high R^2 values, with the standardized data showing a marginal advantage. Set 3 exhibits low R^2 values for both standardized and non-standardized data, indicating poor model performance with little difference between the two. In Set 4, the standardized data yields higher R^2 values than non-standardized data, indicating an improvement in model performance due to standardization. For Set 5, although R^2 values are low for both types of data,

standardized data performs slightly better. Lastly, in Set 6, the R^2 values for standardized and non-standardized data are nearly identical, indicating minimal impact of standardization. These results imply that while data standardization generally benefits the performance of the SVM model, its effect varies across different datasets, highlighting the importance of considering dataset-specific characteristics when applying standardization.

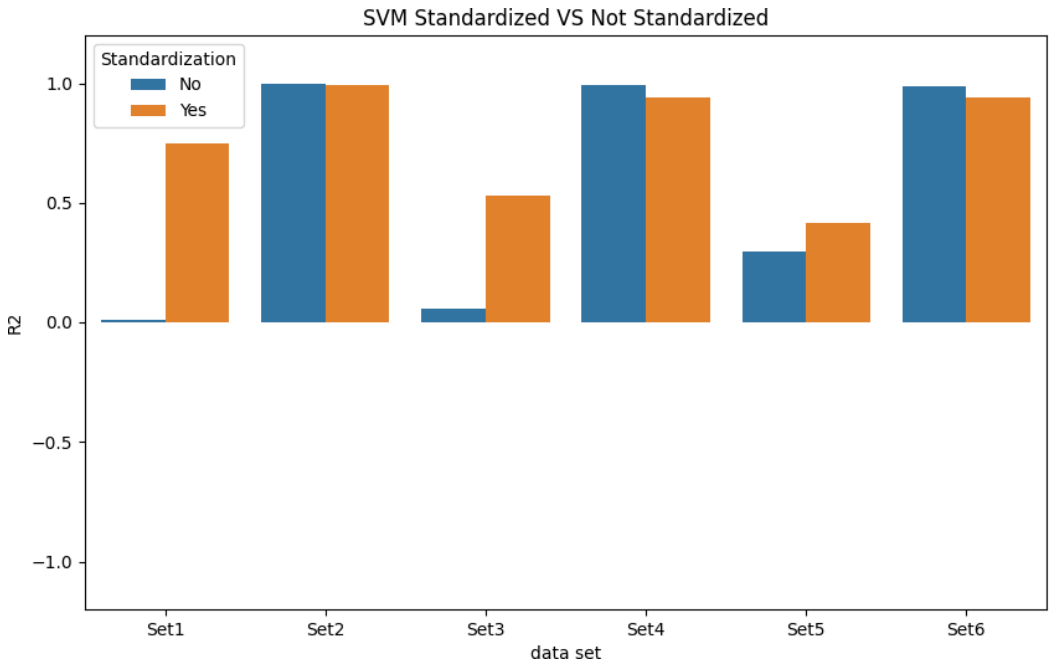

Figure 5-6 Effect of Data Standardization on SVM Model Performance Across Datasets

Figure 5-7 displays the influence of data standardization on the Random Forest (RF) model's performance across six datasets, measured by the coefficient of determination (R^2). In Set 1, both standardized and non-standardized data yield almost identical R^2 values, suggesting that standardization has minimal impact. For Set 2, the RF model performs well with both types of data, with non-standardized data showing a slight advantage. Set 3 results indicate that R^2 values are similar regardless of standardization, implying negligible effect. In Set 4, the R^2 values are nearly the same for both standardized and non-standardized data, pointing to a minimal influence of standardization. Set 5 shows a marginally better performance with non-standardized data, though the difference is slight. Lastly, in Set 6, the R^2 values for standardized and non-standardized data are very close, indicating that standardization does not significantly alter the model's performance.

These findings imply that the effectiveness of data standardization on the RF model varies with the dataset and does not consistently improve performance.

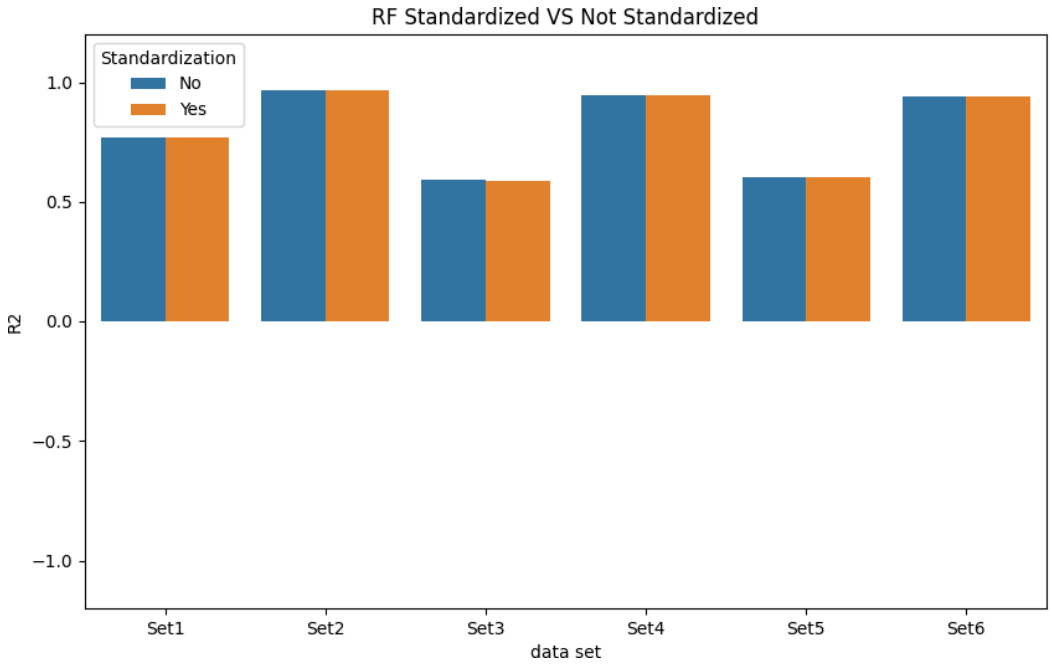

Figure 5-7 Effect of Data Standardization on RF Model Performance Across Datasets

Figure 5-8 illustrates the impact of data standardization on the performance of the MLR model across six datasets, using the coefficient of determination (R^2) as the performance metric. In Set 1, the R^2 values for both standardized and non-standardized data are very similar, indicating that standardization has a negligible effect on the model's performance. Set 2 shows high R^2 values for both data types, with standardized data showing a slight advantage. For Set 3, the R^2 values are low for both standardized and non-standardized data, suggesting poor performance overall, with minimal impact from standardization. Sets 4, 5, and 6 exhibit almost identical R^2 values for both standardized and non-standardized data, indicating that standardization does not significantly influence the model's performance. These results suggest that data standardization does not consistently improve the performance of the MLR model across different datasets, highlighting that its effectiveness may be dataset-dependent.

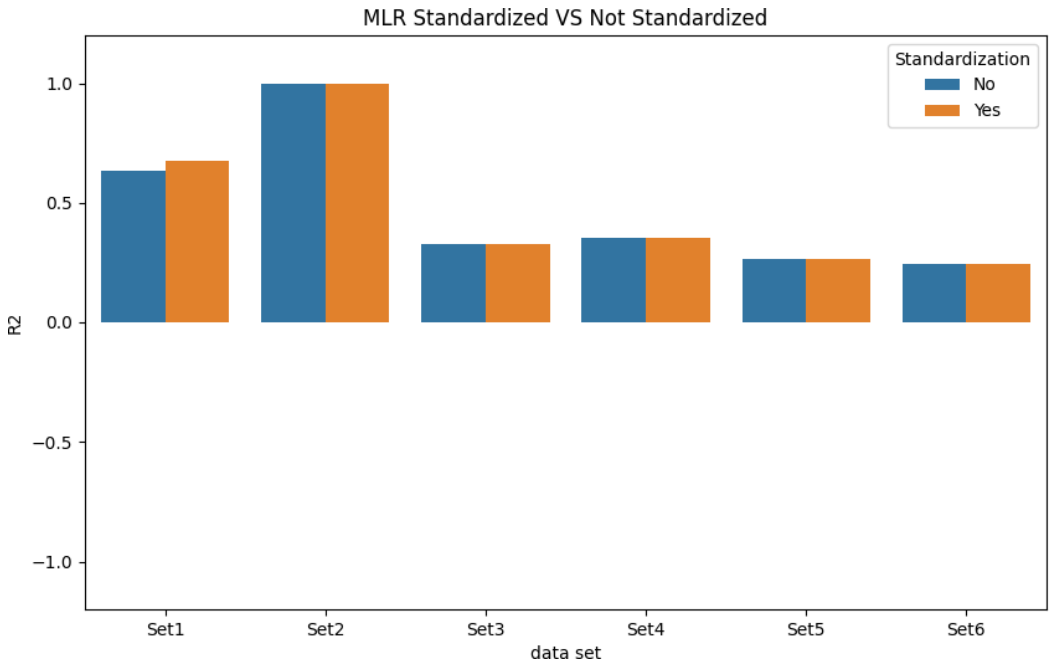

Figure 5-8 Effect of Data Standardization on MLR Model Performance Across Datasets

Figure 5-9 demonstrates the impact of data standardization on the performance of the Multi-Layer Perceptron (MLP) model across six datasets, evaluated using the coefficient of determination (R^2) as the performance metric. In Set 1, the standardized data shows a higher R^2 value compared to the non-standardized data, indicating a positive impact of standardization. For Set 2, standardized data also results in higher R^2 values, suggesting that standardization improves model performance significantly. Set 3, however, reveals a substantial decrease in R^2 for non-standardized data, indicating poor performance without standardization, while standardized data achieves a marginally better result. Set 4 exhibits a higher R^2 for standardized data compared to non-standardized data, highlighting the positive effect of standardization. In Set 5, the R^2 values are low for both standardized and non-standardized data, with standardized data performing slightly better. Finally, in Set 6, both standardized and non-standardized data yield similar R^2 values, suggesting minimal impact from standardization. These findings indicate that data standardization generally enhances the performance of the MLP model, though the extent of improvement varies across different datasets, underscoring the necessity to tailor preprocessing steps to specific dataset characteristics.

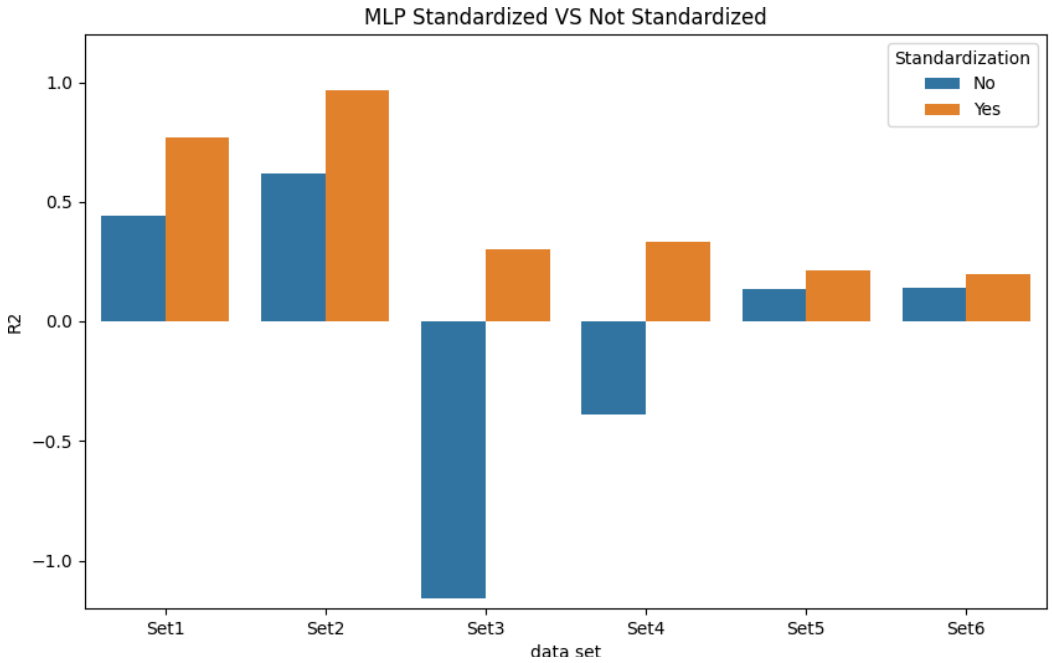

Figure 5-9 Effect of Data Standardization on MLP Model Performance Across Datasets

Figure 5-10 illustrates the impact of data standardization on the performance of the Extreme Gradient Boosting Trees (XGBT) model across six datasets, using the coefficient of determination (R^2) as the performance metric. In Set 1, the R^2 values for standardized and non-standardized data are almost identical, indicating minimal impact from standardization. Set 2 shows high R^2 values for both standardized and non-standardized data, with the standardized data having a slight edge. For Set 3, the R^2 values are low for both standardized and non-standardized data, suggesting that the model struggles with this dataset, regardless of standardization. Set 4 displays very similar R^2 values for both standardized and non-standardized data, indicating that standardization has little effect. In Set 5, the R^2 values are comparable, showing that standardization does not significantly influence performance. Finally, in Set 6, the R^2 values are nearly identical for both standardized and non-standardized data, suggesting minimal impact from standardization. These results suggest that data standardization generally does not significantly alter the performance of the XGBT model across these datasets, highlighting that its effectiveness may vary based on the specific dataset characteristics.

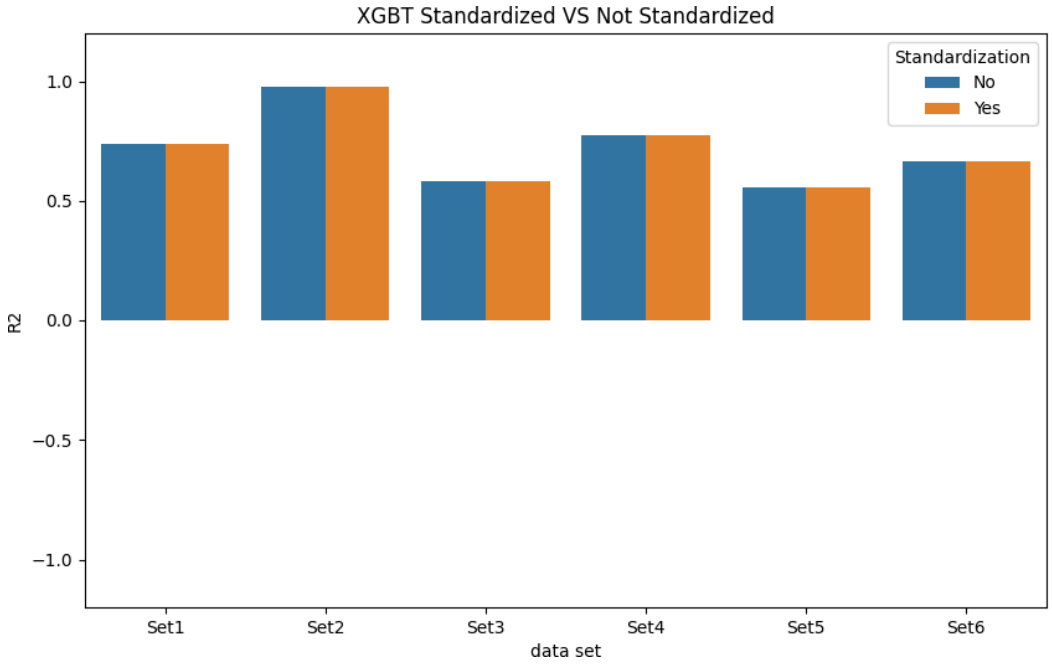

Figure 5-10 Effect of Data Standardization on XGBoost Model Performance Across Datasets

5.3.3 Data Preprocessing Result

The results of the machine learning models applied to the Lake Simcoe dataset after different data preprocessing techniques are visualized in Figure 5-11, Figure 5-12 and Figure 5-13.

Figure 5-11 compares the R^2 values for training and testing sets of the Support Vector Machine (SVM), Decision Tree (DT), and MLR models across three datasets (Set 2, Set 4, and Set 6) using standardized data. The first subplot shows the SVM model's performance. In Set 2, the testing set R^2 is higher than the training set R^2, indicating that the model generalizes well to new data. Set 4 also reveals a higher R^2 for the testing set compared to the training set, suggesting robust generalization. In Set 6, the testing R^2 is again higher, pointing to effective performance on unseen data. The second subplot illustrates the DT model's results. In Set 2, the testing set R^2 significantly exceeds the training set R^2, suggesting possible overfitting or an unusual distribution of test data. Set 4 shows a similar pattern, with higher R^2 values for the testing set. In Set 6, the testing R^2 is notably higher than the training R^2, indicating overfitting or variance in data distribution. The third subplot displays the MLR model's outcomes. In Set 2, the training set R^2 is higher than the testing set R^2, indicating better performance on the training data. Sets 4 and 6 exhibit similar R^2 values for both training and testing sets, suggesting consistent performance and good generalization. Overall, these comparisons reveal that while the SVM model shows strong generalization across all datasets, the DT model tends to overfit, particularly in Sets 2 and 6. The MLR model maintains consistent performance across datasets, highlighting its robustness. This analysis underscores the necessity of evaluating both training and testing set performances to ensure the models are generalizing well and not overfitting.

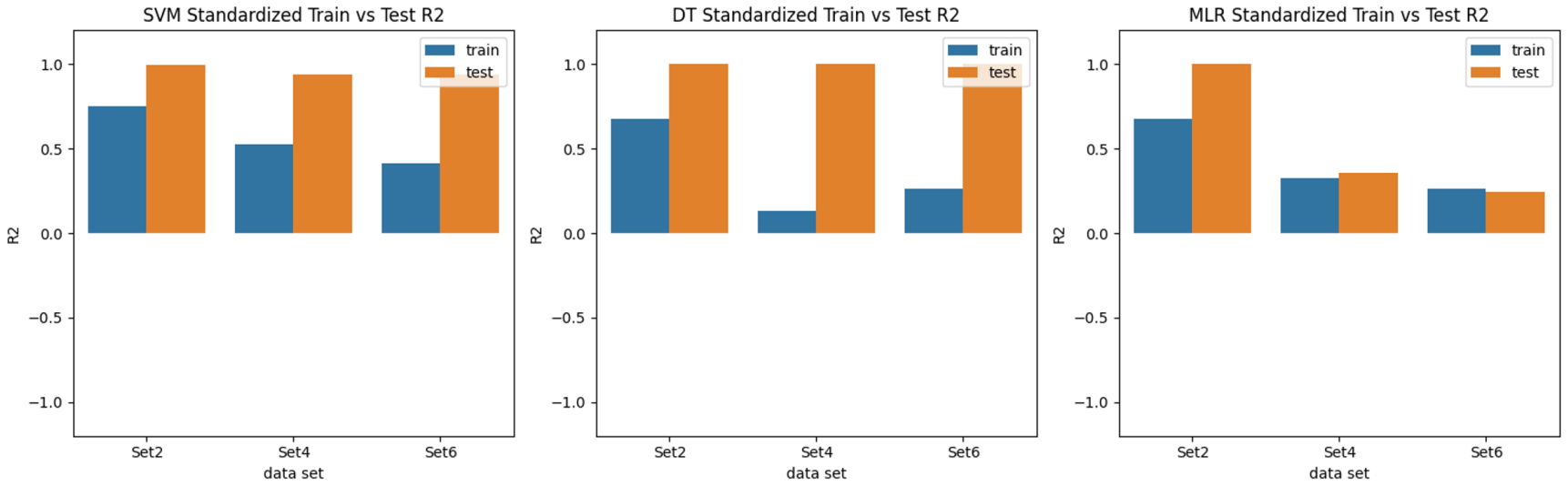

Fig 5-11 Comparative Performance of SVM, DT, and MLR Models on Training and Test Data

Figure 5-12 compares the R^2 values for training and testing sets of the Extreme Gradient Boosting Trees (XGBT), Random Forest (RF), and Multi-Layer Perceptron (MLP) models across three datasets (Set 2, Set 4, and Set 6) using standardized data. The first subplot presents the performance of the XGBT model. In Set 2, the testing set R^2 value is higher than the training set R^2, suggesting good generalization. In Set 4, both sets show similar R^2 values, indicating balanced performance. Set 6 follows a similar trend, with the testing set performing slightly better, demonstrating effective generalization. The second subplot shows the results for the RF model. In Set 2, the testing set R^2 exceeds the training set R^2, which may indicate overfitting or an anomaly in data distribution. Set 4 shows closely matched R^2 values for both training and testing sets, suggesting stable performance. In Set 6, the testing set outperforms the training set, again indicating possible overfitting or data variance. The third subplot displays the MLP model's performance. In Set 2, the training set R^2 is higher than the testing set R^2, pointing to better performance on the training data. In Sets 4 and 6, the R^2 values for both sets are quite close, suggesting consistent performance and good generalization. These comparisons reveal that the XGBT model generalizes well across datasets, while the RF model occasionally exhibits overfitting, particularly in Sets 2 and 6. The MLP model demonstrates consistent performance across datasets, highlighting its robustness. This analysis emphasizes the importance of evaluating both training and testing set performances to ensure models are generalizing effectively and not overfitting.

Figure 5-12 Comparative Performance of XGBT, RF, and MLP Models on Training and Test Data

Figure 5-13 presents a comparative performance analysis of the Gradient Boosting Decision Tree (GBT), KNN, and Long Short-Term Memory (LSTM) models, evaluated using R-squared (R^2) values on standardized training and test datasets. The GBT model shows relatively high R^2 values for both training and test sets across all data sets (Set2, Set4, Set6). Specifically, the R^2 values for the training sets are consistently higher than those for the test sets, indicating a strong fit to the training data but suggesting potential overfitting issues. However, the test set R^2 values remain reasonably high, demonstrating good generalization performance. The KNN model exhibits moderate R^2 values for both training and test sets. The R^2 values for the test sets are slightly higher than those for the training sets across all data sets, which may indicate that the KNN model is less prone to overfitting compared to GBT. However, the overall performance is not as strong as that of the GBT model, as reflected by the lower R^2 values. The LSTM model shows variable performance with notably lower R^2 values for both training and test sets. In particular, the LSTM model displays poor performance on Set4, with negative R^2 values for the training set, indicating a poor fit to the data. This variability and the low R^2 values suggest that the LSTM model may not be well-suited for this specific prediction task, possibly due to inadequate configuration or the nature of the dataset. In summary, the GBT model demonstrates the best overall performance with high R^2 values, although it may be prone to overfitting. The KNN model shows moderate and consistent performance, with slightly better generalization on test data. The LSTM model performs poorly, highlighting the importance of careful model selection and tuning for effective predictive modeling.

Figure 5-13 Comparative Performance of GBDT, KNN, and LSTM Models on Training and Test Data

Overall, the choice of model largely depends on the specific characteristics of the dataset, the complexity of the problem, and the need for model interpretability. Ensemble models like RF, GBT, and XGBoost generally provide superior performance by effectively capturing complex patterns and interactions, making them highly suitable for challenging tasks like predicting environmental variables. However, simpler models like MLR and KNN can also be effective when applied to less complex problems or when interpretability is a critical factor. The results underscore the importance of model selection based on the task at hand and the characteristics of the data, alongside proper tuning and validation to achieve optimal performance.

5.3.4 Model Performance Result

Figure 5-14 provides a comprehensive comparison of various machine learning models, evaluated based on R^2, MSE, MAE, and MAPE for both training and test datasets. The analysis reveals distinct patterns in the performance of these models, highlighting their strengths and weaknesses. The Random Forest (RF) and Gradient Boosting Decision Tree (GBT) models demonstrate strong predictive performance with high R^2 values in both training and test sets, indicating their superior ability to capture the underlying patterns in the data. The RF model exhibits the lowest error metrics (MSE, MAE, and MAPE) across both datasets, emphasizing its robustness and accuracy. Support Vector Machine (SVM) models show moderate R^2 values for both training and test sets. Although the error metrics for SVM are relatively low, they are not as minimal as those observed for the RF and GBT models, suggesting a reasonably good but not top-tier predictive performance. The Decision Tree (DT) models achieve high R^2 values in the training set but show a notable decrease in the test set R^2 values, indicative of overfitting. This discrepancy between training and test performance points to the need for further tuning to enhance generalization. Neural network models, such as the Multi-Layer Perceptron (MLP), display relatively high R^2 values and low error metrics, indicating effective prediction capabilities. However, the Long Short-Term Memory (LSTM) models present lower R^2 values and higher error metrics on the test set, suggesting that their predictive accuracy could benefit from additional optimization. The K Nearest Neighbors (KNN) models exhibit moderate performance, characterized by lower R^2 values and higher error metrics compared to other models. This indicates that KNN may not be as effective in capturing the complex relationships within the data. MLR models achieve consistent R^2 values with moderate error metrics across both datasets, reflecting reasonable predictive ability but falling short of the performance levels demonstrated by more sophisticated models. XGBoost models also show high R^2 values and low error metrics, confirming their strong predictive performance. As a conclusion, the RF and GBT models emerge as the most effective for the given dataset, achieving the highest R^2 values and the lowest error metrics. The SVM, DT, MLP, LSTM, MLR, and KNN models also exhibit varying levels of predictive strength, each offering distinct advantages depending on the specific requirements of the predictive task.

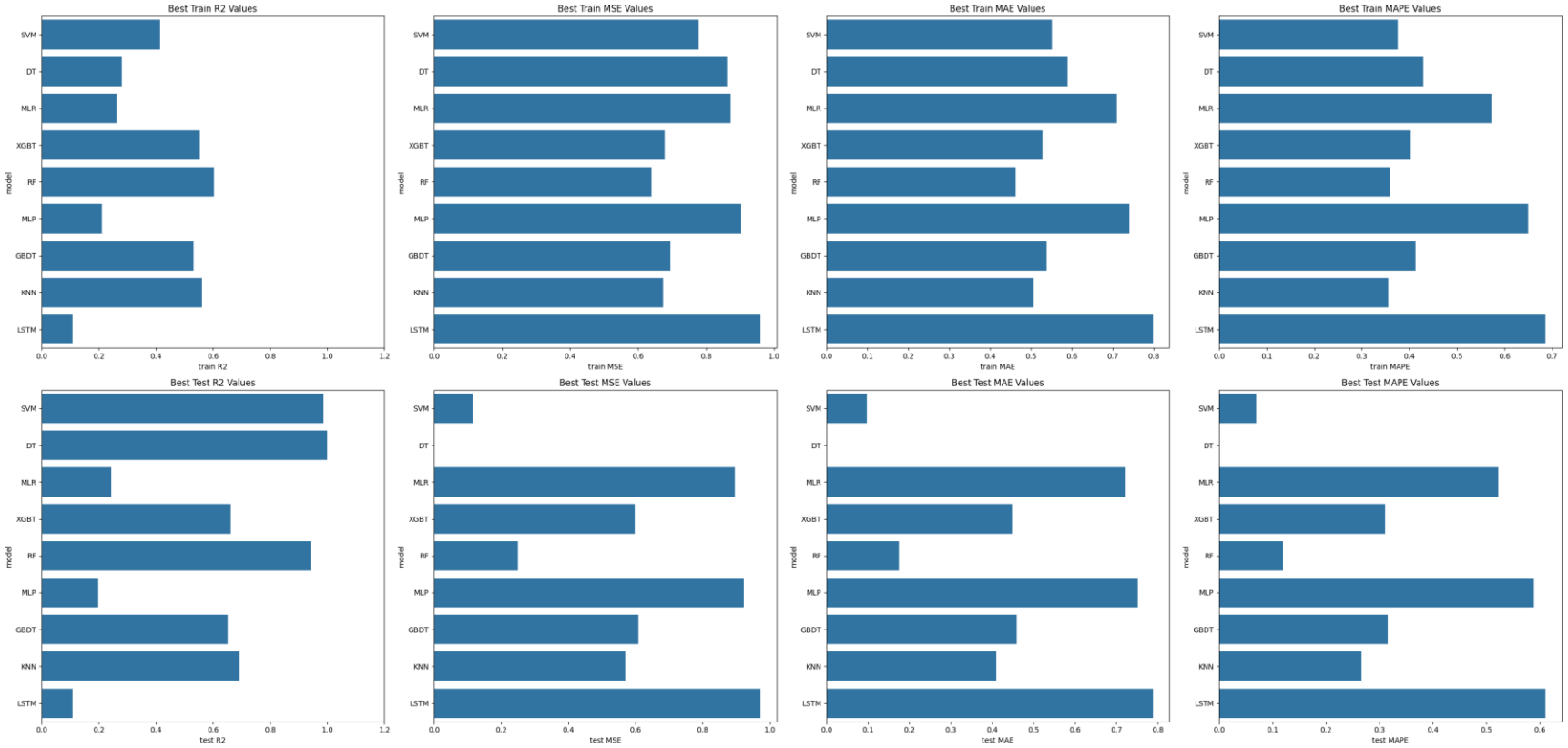

Figure 5-14 Model Evaluation Result

5.3.5 Model Validation Result

To further assess the robustness and generalizability of our integrated AI-based online system, we conducted model validation using the Lake Simcoe dataset spanning the years 2019 to 2021. The validation aimed to evaluate the models' performance on unseen data, thus providing insights into their predictive capabilities in real-world scenarios.

Table 5-1 presents the validation results for various models, evaluated using R^2, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The Random Forest (RF) model shows the best performance with the highest R^2 value of 0.7657, the lowest RMSE of 0.2671, MAE of 0.1438, and MAPE of 0.0601, indicating its superior predictive capability and accuracy. The Gradient Boosting Decision Tree (GBT) model also performs well with an R^2 of 0.6383, RMSE of 0.3319, MAE of 0.2780, and MAPE of 0.1177, demonstrating robust performance. The KNN model achieves good results with an R^2 of 0.6029, RMSE of 0.3478, MAE of 0.2791, and MAPE of 0.1100, suggesting reliable predictions. The Decision Tree (DT) model, with an R^2 of 0.1834, RMSE of 0.4987, MAE of 0.1958, and MAPE of 0.0857, shows moderate performance.

Table 5-1 Model Validation Result From Different Models – Case Lake Simcoe

Model	R2	RMSE	MAE	MAPE
DT	0.1834	0.4987	0.1958	0.0857
GBT	0.6383	0.3319	0.2780	0.1177
KNN	0.6029	0.3478	0.2791	0.1100
LSTM	0.2367	0.6383	0.5208	0.2191
MLP	-0.3198	0.6340	0.5179	0.2085
MLR	-0.7238	0.9593	0.7854	0.3077
RF	0.7657	0.2671	0.1438	0.0601
SVM	-1.5982	1.1777	1.0281	0.4450
XGBT	0.1929	0.4958	0.4037	0.1587

On the other hand, models like the Support Vector Machine (SVM), MLR, and Multi-Layer Perceptron (MLP) exhibit poor performance. The SVM model has a negative R^2 of -1.5982, the highest RMSE of 1.1777, MAE of 1.0281, and MAPE of 0.4450, indicating it is not suitable for this dataset. The MLR and MLP models also show negative R^2 values of -0.7238 and -0.3198, respectively, with corresponding high error metrics. The Long Short-Term Memory (LSTM) model has a relatively low R^2 of 0.2367, with higher error metrics of RMSE 0.6383, MAE 0.5208, and MAPE 0.2191, indicating less reliable performance. The Extreme Gradient Boosting Trees (XGBT) model shows moderate performance with an R^2 of 0.1929, RMSE of 0.4958, MAE of 0.4037, and MAPE of 0.1587.

Among all the models evaluated, the RF model demonstrates the highest accuracy and reliability, making it the most suitable choice for the given dataset. Its superior R^2 value and lower error metrics highlight its ability to capture the underlying patterns in the data effectively, providing more accurate predictions compared to other models.

5.4 Discussion

Figure 5-15 illustrates the performance of the Gradient Boosting Decision Tree (GBT) model at Station K45 by comparing actual observed values with the model's predictions over a timeline from February 2019 to November 2021. The GBT model appears to follow the trend of the actual values closely, demonstrating its effectiveness in capturing the underlying patterns of the dataset. Notably, the model shows a robust predictive capability, especially from mid-2020 to late 2021, where the predicted values tightly align with the actual measurements, indicating a high degree of accuracy. However, there are periods, particularly in the earlier phases around mid-2019, where the prediction diverges slightly from the actual values, suggesting some limitations in the model's performance during certain conditions or possibly underestimating sudden changes in the dataset. Overall, the GBT model exhibits strong predictive performance with minor deviations, showcasing its potential for reliable forecasting in this application area. Figure 5-16 displays the comparative analysis of actual values against the predictions made by the Decision Tree (DT) model at Station K45, spanning from February 2019 to November 2021. The graph demonstrates that the DT model closely follows the overall trends of the actual data, capturing the seasonal variations and peaks with reasonable accuracy. Particularly from mid-2020 onwards, the predictions align more closely with the actual values, suggesting that the DT model is effectively adapting to the underlying patterns in the data. However, there are periods, especially in the early months and around the peak values, where the prediction slightly deviates from the actual measurements. This indicates potential overfitting or the model's sensitivity to fluctuations in the data, which could be due to the inherent nature of decision trees to fit precise data points, sometimes at the expense of generalizability. Overall, the DT model exhibits a solid performance, but it may benefit from further tuning or ensemble methods to smooth out predictions and enhance accuracy.

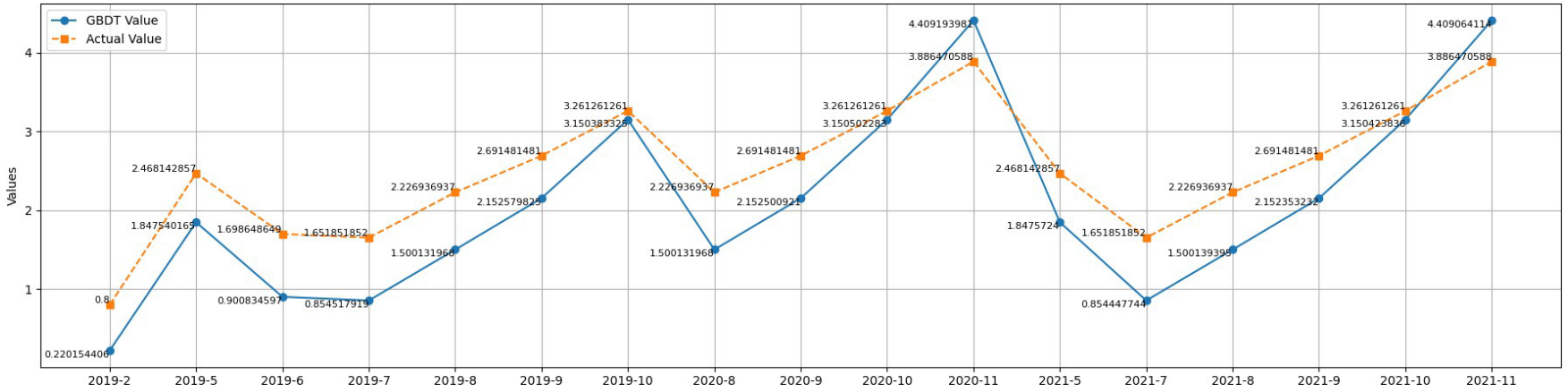

Figure 5-15 Comparison of Actual Value and Prediction Value at Station K45 by GBT Model

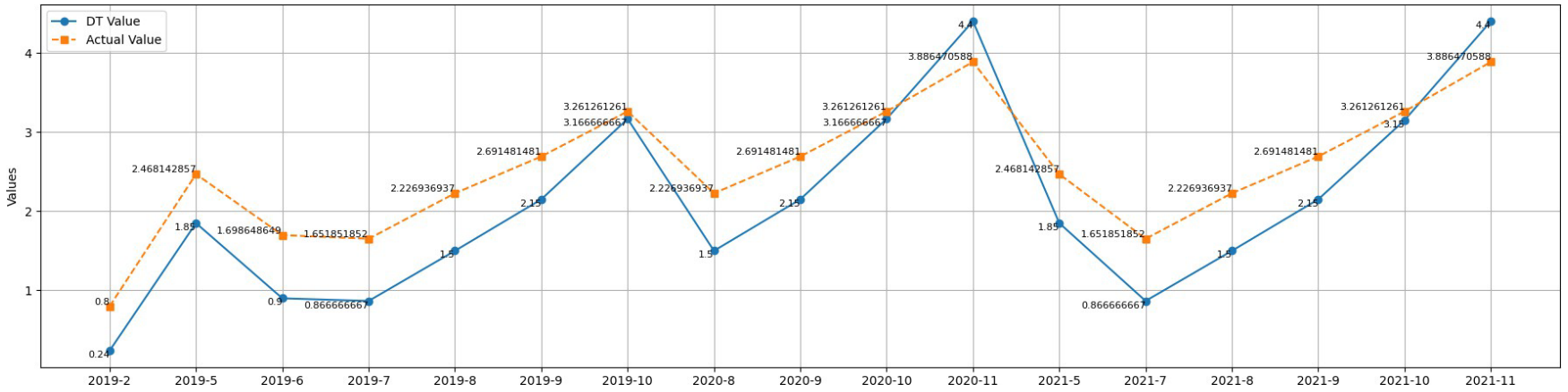

Figure 5-16 Comparison of Actual Value and Prediction Value at Station K45 by DT Model

Figure 5-17 illustrates the performance comparison between actual values and predictions made by the Multi-Layer Perceptron (MLP) model at Station K45 from February 2019 to November 2021. This graph demonstrates the MLP model's ability to approximate the cyclical patterns and trends in the dataset, indicating a solid understanding of the underlying dynamics of the data. The model effectively captures both the peaks and troughs, although it shows slight deviations at peak points, particularly around May 2019 and October 2020. These discrepancies suggest some challenges in capturing extreme values or sudden changes in data trends. Nonetheless, the model remains generally consistent with the actual data throughout the period, showing closer alignment in later dates, such as from mid-2020 onward. This progression might reflect the model's adaptiveness or improvements in learning from accumulating data over time, showcasing MLP's potential utility in applications requiring trend analysis and forecasting in dynamic environments. Figure 5-18 illustrates the comparison between actual values and predictions made by the KNN model at Station K45 over the period from February 2019 to November 2021. The graph shows that the KNN model captures the general trend and seasonality of the dataset with reasonable accuracy, although it exhibits some discrepancies, particularly in peak and trough predictions. The model tends to underestimate peaks (e.g., October 2020) and shows slight overestimations at some trough points (e.g., May 2021). Despite these variances, the KNN model maintains a close alignment with actual values most of the time, suggesting its usefulness in predicting trends with moderate fluctuations. However, the deviations at critical turning points could be indicative of the model's sensitivity to local data properties and the influence of its parameter settings, such as the number of neighbors, which might require optimization for improved accuracy.

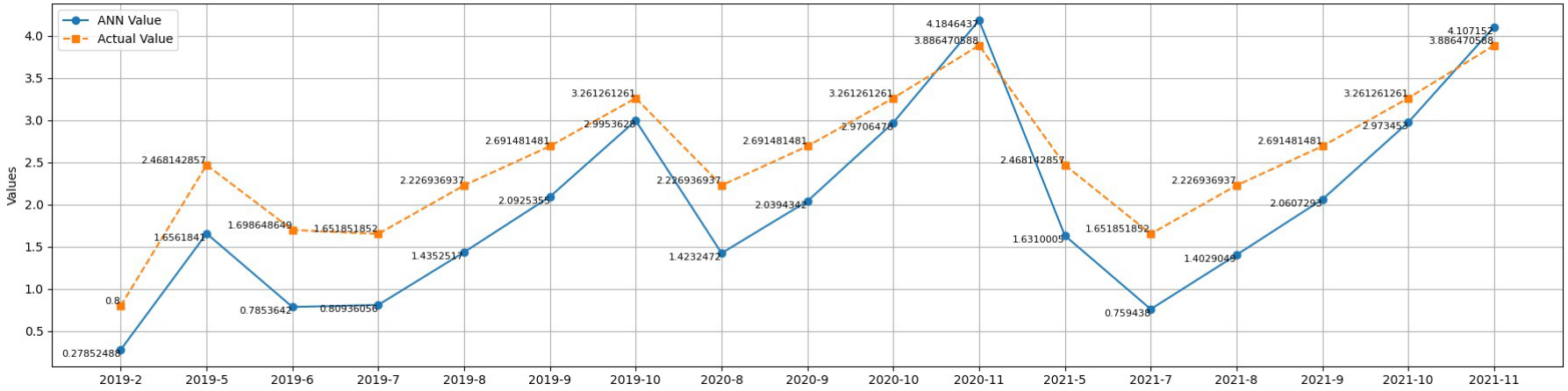

Figure 5-17 Comparison of Actual Value and Prediction Value at Station K45 by MLP Model

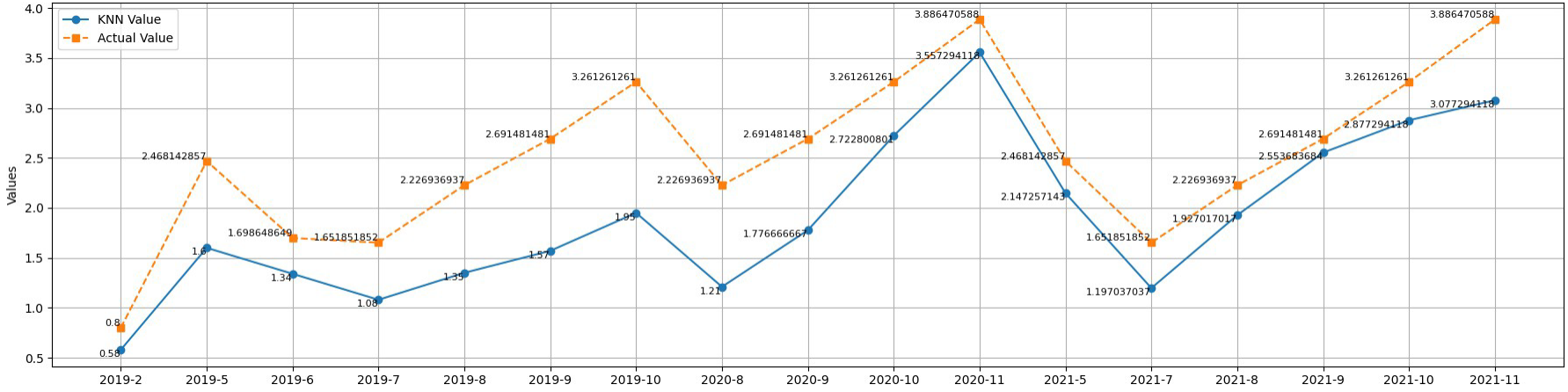

Figure 5-18 Comparison of Actual Value and Prediction Value at Station K45 by KNN Model

Figure 5-19 illustrates the comparison of actual values against the predictions from the Random Forest (RF) model at Station K45 from February 2019 to November 2021. The RF model demonstrates impressive predictive accuracy, closely mirroring the actual values throughout the observed period. It effectively captures the trends and fluctuations in the data, maintaining a consistent proximity to the actual values, even during peak fluctuations (e.g., October 2020 and October 2021) and significant troughs (e.g., May 2021). The model shows a particularly strong alignment during periods of rapid value changes, which indicates its robustness and capability to adapt to complex patterns in the data. The slight deviations seen are minimal, suggesting that the RF model is highly effective for predicting trends at this station, making it a reliable choice for tasks requiring high precision in similar settings. Figure 5-20 displays the comparison of actual values against the predictions made by the MLR model at Station K45 from February 2019 to November 2021. The MLR model demonstrates moderate predictive performance, capturing the overall trend of the dataset but with notable discrepancies at certain points, especially during peaks and troughs. The model starts with a significant underestimation and then stabilizes, tracking closer to the actual values but consistently lagging behind during sharp increases or decreases. For instance, the prediction in May 2019 significantly overestimates the actual value, and a similar pattern is observed in September 2019. Despite these discrepancies, the MLR model manages to approximate the general movement of the actual values towards the end of the observed period. This suggests that while MLR can provide a reasonable baseline understanding of the data trends, it may require additional refinement or incorporation of other predictive factors to enhance its accuracy for precise applications.

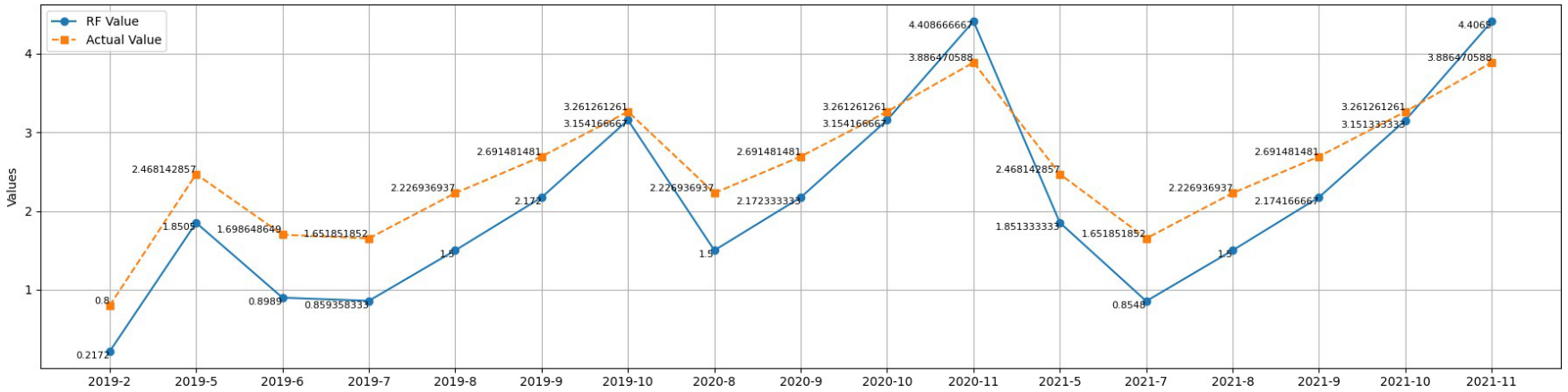

Figure 5-19 Comparison of Actual Value and Prediction Value at Station K45 by RF Model

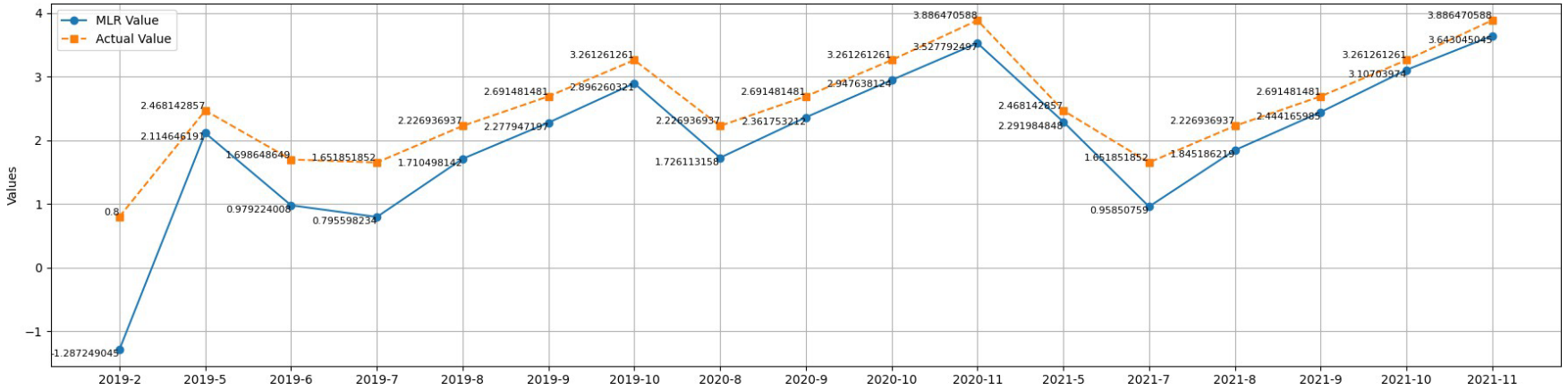

Figure 5-20 Comparison of Actual Value and Prediction Value at Station K45 by MLR Mode

Figure 5-21 depicts the performance of the XGBoost model in predicting values at Station K45, comparing actual observations with model predictions from February 2019 to November 2021. This graph illustrates the XGBoost model's strong capability to approximate the true data trends, particularly in capturing the general fluctuations and seasonality present in the dataset. The model closely mirrors the actual values with slight deviations at peak points, such as in May 2019 and May 2021, where it slightly overestimates the peaks. However, it effectively predicts the rising and falling trends throughout the observed period, demonstrating its robustness and the effectiveness of its ensemble learning methodology in handling complex patterns. The consistency of the predictions, particularly from late 2020 into 2021, suggests that XGBoost is a highly reliable model for forecasting in this context, with potential for precise and accurate predictions in similar future applications. Figure 5-22 illustrates the performance of the Support Vector Machine (SVM) model in tracking and predicting the actual values at Station K45 from February 2019 to November 2021. The SVM model demonstrates reasonable alignment with the actual data, capturing the overall trend and seasonal fluctuations effectively. However, there are noticeable discrepancies, particularly in the peak predictions around May 2019 and May 2021, where the model significantly overshoots the actual values. This overshooting is also evident around October 2020, where the model's predicted value far exceeds the actual peak. Despite these issues, the model adapts well to the rising and falling trends throughout the timeline, indicating a robust understanding of the data dynamics over the long term. The overall pattern suggests that while the SVM model can reliably predict the general trends and seasonality, it may require tuning to improve accuracy at peak points and to minimize prediction errors for more precise forecasting needs.

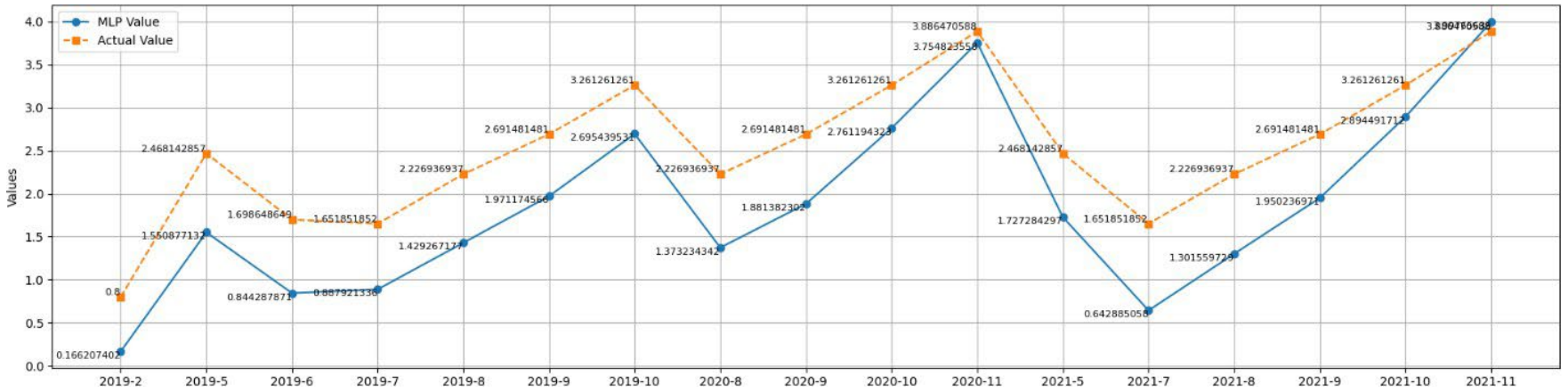

Figure 5-21 Comparison of Actual Value and Prediction Value at Station K45 by XGBoost Model

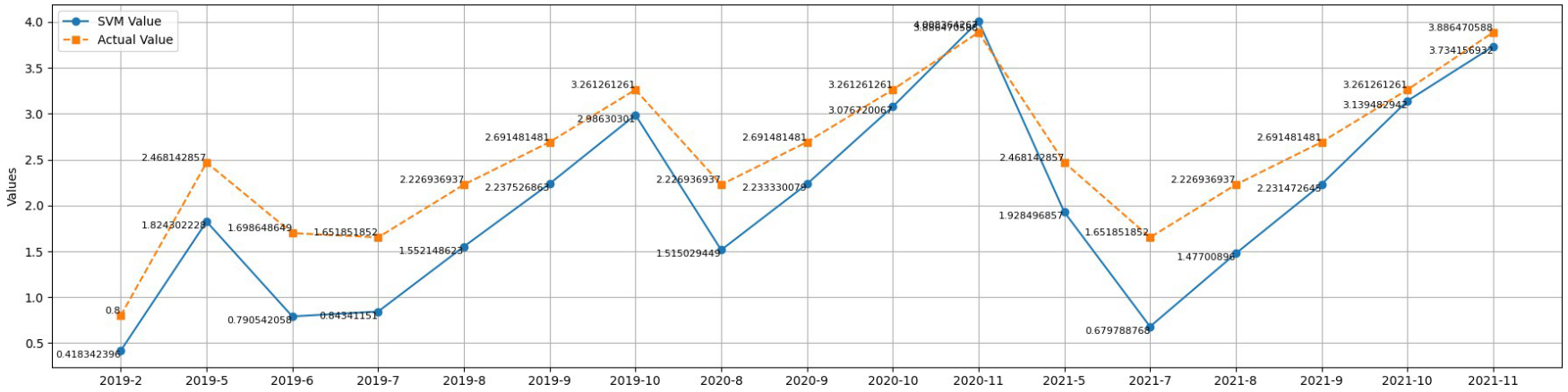

Figure 5-22 Comparison of Actual Value and Prediction Value at Station K45 by SVM Model

5.5 Summary

This chapter presented an exhaustive exploration and evaluation of our integrated AI-based online system for monitoring water quality and modeling Chl-a concentration in Lake Simcoe. The analysis began with a detailed examination of the Lake Simcoe dataset and proceeded with critical data preprocessing steps including Missing Value Imputation (MVI), Outlier Detection (OD), Feature Selection (FS), and Train-Test Split (TTS), ensuring the data's accuracy and relevance for modeling.

In our comprehensive study, we utilized a diverse array of machine learning algorithms such as Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Gradient Boosting Tree (GBT), Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), K Nearest Neighbors (KNN), MLR, and Extreme Gradient Boosting (XGBoost). Each model was rigorously trained and validated to assess its predictive capabilities. The results depicted in this chapter offer a detailed review of each model's performance, enhancing our understanding of their effectiveness in real-world applications using data from Lake Simcoe spanning 2019 to 2021.

The analysis underscored the superior performance of tree-based models such as RF, DT, and GBT, which consistently excelled across various evaluation metrics. The SVM and MLP models also showed robust performance, particularly in capturing complex patterns in the data, thereby confirming their practical utility in environmental modeling. However, caution is advised in the application of the LSTM and MLR models, which displayed inconsistent performance and may require further tuning or methodology adjustments to improve their predictive accuracy. The findings from this study affirm the capability of our integrated AI-based system to deliver reliable predictions of Chl-a concentrations in Lake Simcoe, demonstrating its potential for broader environmental modelling applications.

Chapter 6: Conclusion and Recommendations

In the final chapter, the conclusions drawn from the research are presented. The contributions of the developed AI-based online system for lake Chl-a quality modeling and monitoring are summarized. Additionally, recommendations for future work are provided, suggesting areas for further improvement and exploration in the field of lake water quality monitoring and AI-based modeling.

6.1 Conclusion

In this thesis, we have developed an integrated AI-based online system for Lake Chl-a concentration modeling and monitoring. The system leverages various machine learning models, including Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Gradient Boosting Tree (GBT), Multi-Layer Perceptron (MLP), LSTM, K Nearest Neighbors (KNN), MLR, and Extreme Gradient Boosting (XGBoost), to predict and monitor Chl-a concentrations in lakes.

Through the implementation and evaluation of these models on two different datasets, Lake Champlain and Lake Simcoe, we have gained valuable insights into their performance and applicability. The results indicate that the RF, GBT, and ANN models consistently demonstrated excellent predictive performance, while other models such as SVM, MLP, and MLR exhibited varying levels of performance that could be further optimized.

Based on the results and discussions presented in earlier chapters, the following conclusions can be drawn:

- Data preprocessing techniques, including missing value imputation, outlier detection, and feature selection, significantly improved the performance of machine learning models on the Lake Simcoe dataset. These preprocessing steps helped address data quality issues and enhance the models' predictive capabilities.
- The Random Forest (RF), Gradient Boosting Tree (GBT), and XGBoost models exhibited excellent predictive performance on the Lake Simcoe dataset. These models demonstrated a high degree of accuracy in capturing complex relationships and making precise predictions.

- The Support Vector Machine (SVM), MLP, and MLR models showed good predictive performance but may require further optimization and tuning to maximize their potential.
- The Decision Tree (DT), K Nearest Neighbors (KNN), and LSTM models displayed varying levels of performance, with strengths and limitations specific to each model. These models can benefit from additional improvements and fine-tuning for optimal performance.

6.2 Contributions

The development of the integrated AI-based online system for Lake Chl-a concentration modeling and monitoring makes significant contributions to the field of environmental science and water quality monitoring. This study examined the most recent data-driven AI models for large-scale lake water quality modeling and assessment, showcasing the potential of advanced AI techniques in this domain. New algorithms were developed to assess and select critical factors and data-driven models, ensuring realistic field-scale lake water quality modeling. Additionally, an online, user-friendly eutrophication modeling system was created to assess algae blooming in complex large-scale lake systems with high spatial and temporal resolution.

The developed AI models, particularly the Random Forest (RF) model, outperformed traditional models by effectively handling data complexity and nonlinearity. The integration of advanced data preprocessing techniques, such as missing value imputation, outlier detection, and feature selection, significantly improved model accuracy and reliability. Furthermore, the online modelling system provides real-time data collection, processing, and prediction capabilities, making it a valuable tool for water quality management. This system allows for accurate predictions and real-time modelling, enabling stakeholders to make informed decisions and take timely actions for water quality management.

6.3Recommendations for Future Work

The development of the Chl-a Modeling and Monitoring Online System (CMMOS) involves the integration of multiple sophisticated components, including various machine learning models, data preprocessing techniques, and real-time modelling capabilities. This complexity, while necessary for achieving high predictive accuracy and robustness, introduces challenges in

terms of system maintainability and scalability. Future work should focus on simplifying the system architecture where possible, without compromising performance, to facilitate easier maintenance and potential expansion. Exploring modular approaches that allow individual components to be updated or replaced independently could significantly enhance the system's flexibility and longevity.

Machine learning models inherently carry uncertainties due to several factors, such as data quality, model selection, parameter tuning, and the underlying assumptions of the models. These uncertainties can impact the reliability of the predictions made by the CMMOS. Future research should aim to quantify these uncertainties more rigorously, perhaps through the use of advanced statistical techniques or ensemble modeling approaches that combine the strengths of multiple models. Additionally, incorporating uncertainty estimation into the system's output can provide users with a clearer understanding of the confidence levels associated with the predictions, thereby aiding more informed decision-making.

In summary, future work can focus on extending the following aspects:

- Data Dependency: The accuracy of the system heavily relies on the quality and quantity of the input data. Inconsistent or sparse data can lead to unreliable predictions. Future efforts should focus on enhancing data collection methods, possibly by integrating more diverse data sources, including remote sensing data, citizen science contributions, and automated sensors.
- Scalability: While the system has been tested on specific lakes (Lake Champlain and Lake Simcoe), its scalability to other geographic locations and larger scales remains to be validated. Future work should explore the adaptability of the system to different environmental conditions and broader geographic areas, ensuring that the models can generalize well beyond the initial study sites.
- Computational Resources: The implementation of multiple machine learning models and real-time data processing requires significant computational resources. This can be a barrier for deployment in resource-constrained environments. Investigating more efficient algorithms and leveraging cloud computing resources could help mitigate this limitation.
- User Interface and Experience: Although the system includes a user interface for stakeholders, its usability and accessibility could be improved. Future development should prioritize user-

centric design principles, incorporating feedback from end-users to enhance the interface's intuitiveness and functionality. This could involve the development of mobile applications or more interactive web-based tools.

- Integration with Policy and Management: The current system primarily focuses on technical aspects of modelling and prediction. For the system to have a more substantial impact, it should be integrated with policy and management frameworks. Future work should explore collaborations with environmental agencies and policymakers to ensure that the system's outputs are aligned with regulatory requirements and management objectives.

By addressing these complexities, uncertainties, and limitations, future enhancements to the CMMOS can lead to a more robust, scalable, and user-friendly system that significantly contributes to sustainable lake management and environmental modelling practices.

Reference

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy and Buildings, 147, 77-89.

Alexandropoulos, S. A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. The Knowledge Engineering Review, 34, e1.

Altunkaynak, A. (2013). Prediction of significant wave height using geno-multilayer perceptron. Ocean Engineering, 58, 144-153.

Barzegar, R., Aalami, M. T., & Adamowski, J. (2020). Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. Stochastic Environmental Research and Risk Assessment, 34(2), 415-433.

Behmel, S., Damour, M., Ludwig, R., & Rodriguez, M. J. (2016). Water quality monitoring strategies—A review and future perspectives. Science of the Total Environment, 571, 1312-1329.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967.

Bhagowati, B., & Ahamad, K. U. (2019). A review of lake eutrophication dynamics and recent developments in lake modeling. Ecohydrology & Hydrobiology, 19(1), 155-166.

Bhateria, R., & Jain, D. (2016). Water quality assessment of lake water: a review. Sustainable Water Resources Management, 2, 161-173.

Burigato Costa, C. M. D. S., da Silva Marques, L., Almeida, A. K., Leite, I. R., & de Almeida, I. K. (2019). Applicability of water quality models around the world—a review. Environmental Science and Pollution Research, 26, 36141-36162.

Çamdevýren, H., Demýr, N., Kanik, A., & Keskýn, S. (2005). Use of principal component scores in multiple linear regression models for prediction of Chl-a in reservoirs. Ecological Modelling, 181(4), 581-589.

Cao, Z., Ma, R., Duan, H., Pahlevan, N., Melack, J., Shen, M., & Xue, K. (2020). A machine learning approach to estimate Chl-a from Landsat-8 measurements in inland lakes. Remote Sensing of Environment, 248, 111974.

Cen, H., Jiang, J., Han, G., Lin, X., Liu, Y., Jia, X., ... & Li, B. (2022). Applying deep learning in the prediction of Chl-a in the East China Sea. Remote Sensing, 14(21), 5461.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408, 189-215.

Chang, N. B., Imen, S., & Vannah, B. (2015). Remote sensing for monitoring surface water quality status and ecosystem state in relation to the nutrient cycle: a 40-year perspective. Critical Reviews in Environmental Science and Technology, 45(2), 101-166.

Chen, C. H., Tanaka, K., Kotera, M., & Funatsu, K. (2020). Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications. Journal of Cheminformatics, 12, 1-16.

Chen, Q., Guan, T., Yun, L., Li, R., & Recknagel, F. (2015). Online forecasting chlorophyll a concentrations by an auto-regressive integrated moving average model: Feasibilities and potentials. Harmful Algae, 43, 58-65.

Chislock, M. F., Doster, E., Zitomer, R. A., & Wilson, A. E. (2013). Eutrophication: causes, consequences, and controls in aquatic ecosystems. Nature Education Knowledge, 4(4), 10.

Cho, H., Choi, U. J., & Park, H. (2018). Deep learning application to time-series prediction of daily Chl-a concentration. WIT Transactions on Ecology and the Environment, 215, 157-163.

Coops, N. C., Smith, M. L., Martin, M. E., & Ollinger, S. V. (2003). Prediction of eucalypt foliage nitrogen content from satellite-derived hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, 41(6), 1338-1346.

Cruz, R. C., Reis Costa, P., Vinga, S., Krippahl, L., & Lopes, M. B. (2021). A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination. Journal of Marine Science and Engineering, 9(3), 283.

Deng, S., Zhang, Y., Wang, Y., Li, X., & Li, Y. (2023). *Environmental Science & Technology*, 57(7), 3114-3124. https://doi.org/10.1016/j.envsci.2023.03.003

Demir, S., & Şahin, E. K. (2022). Liquefaction prediction with robust machine learning algorithms (SVM, RF, and Extreme Gradient Boosting) supported by genetic algorithm-based feature selection and parameter optimization from the perspective of data processing. Environmental Earth Sciences, 81(18), 459.

Du, Z., Qin, M., Zhang, F., & Liu, R.-y. (2018). Multistep-ahead forecasting of chlorophyll a using a wavelet nonlinear autoregressive network. Knowl. Based Syst., 160, 61-70.

El-Serehy, H. A., Abdallah, H. S., Al-Misned, F. A., Al-Farraj, S. A., & Al-Rasheid, K. A. (2018). Assessing water quality and classifying trophic status for scientifically based management of the water resources of Lake Timsah, the lake with salinity stratification along the Suez Canal. Saudi Journal of Biological Sciences, 25(7), 1247-1256.

Filstrup, C. T., Wagner, T., Soranno, P. A., Stanley, E. H., Stow, C. A., Webster, K. E., & Downing, J. A. (2014). Regional variability among nonlinear chlorophyll—phosphorus relationships in lakes. Limnology and Oceanography, 59(5), 1691-1703.

Fontana, C., Grenz, C., Pinazo, C., Marsaleix, P., & Diaz, F. (2009). Assimilation of SeaWiFS chlorophyll data into a 3D-coupled physical-biogeochemical model applied to a freshwater-influenced coastal zone. Continental Shelf Research, 29, 1397-1409.

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. Environmental Monitoring and Assessment, 189, 1-20.

Fratello, M., & Tagliaferri, R. (2018). Decision trees and random forests. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 374.

Golhani, K., Balasundram, S. K., Vadamalai, G., & Pradhan, B. (2018). A review of neural networks in plant disease detection using hyperspectral data. Information Processing in Agriculture, 5(3), 354-371.

Guang-ren, Q. (2012). Using Support Vector Regression Algorithm to Predict Chl-a Concentrations with Chenghai Lake for Example. Journal of Environmental Engineering Technology.

Haggerty, R., Sun, J., Yu, H., & Li, Y. (2023). Application of machine learning in groundwater quality modeling-A comprehensive review. Water Research, 119745.

He, X., Shi, S., Geng, X., Xu, L., & Zhang, X. (2021). Spatial-temporal attention network for multistep-ahead forecasting of chlorophyll. Applied Intelligence, 51, 4381-4393.

Hollister, J., Milstead, W. B., & Kreakie, B. (2016). Modeling lake trophic state: a random forest approach. Ecosphere, 7.

Hu, M., Wang, Y., Sun, Z., Su, Y., Li, S., Bao, Y., & Wen, J. (2021). Performance of ensemble-learning models for predicting eutrophication in Zhuyi Bay, Three Gorges Reservoir. River Research and Applications, 37(8), 1104-1114.

Hua-jun, L., & Defu, L. (2009). Genetic algorithm-support vector machine model for short-term prediction of chlorophyll a concentration nonlinear time series. Journal of Hydraulic Engineering.

Huan, J., Li, H., Li, M., & Chen, B. (2020). Prediction of dissolved oxygen in aquaculture based on gradient boosting decision tree and long short-term memory network. Computers and Electronics in Agriculture, 175, 105530.

Huang, H., Wang, W., Lv, J., Liu, Q., Liu, X., Xie, S., ... & Feng, J. (2022). Relationship between chlorophyll a and environmental factors in lakes based on the random forest algorithm. Water, 14(19), 3128.

Ifenthaler, D., & Widanapathirana, C. (2014). Development and validation of a learning analytics framework: Two case studies using support vector machines. Technology, Knowledge and Learning, 19, 221-240.

Jang, M.-Y., Choi, J., Kim, J., Seo, D., & Kim, J. (2020). A Hybrid Approach for the Prediction of Chl-a Concentration at the Non-monitoring Area in the Geum River, Korea. 2020 International Conference on Information and Communication Technology Convergence (ICTC).

Jeong, K. S., Kim, D. K., & Joo, G. J. (2006). River phytoplankton prediction model by Extreme Gradient Boosting: Model performance and selection of input variables to predict time-series phytoplankton proliferations in a regulated river system. Ecological Informatics, 1(3), 235-245.

Jia, W., Cheng, J., & Hu, H. (2020). A cluster-stacking-based approach to forecasting seasonal Chl-a concentration in coastal waters. IEEE Access, 8, 99934-99947.

Jia, W., Cheng, J., & Hu, H. (2020). A Cluster-Stacking-Based Approach to Forecasting Seasonal Chl-a Concentration in Coastal Waters. IEEE Access, 8, 99934-99947.

Kag, A., & Saligrama, V. (2021, July). Training recurrent neural networks via forward propagation through time. In International Conference on Machine Learning (pp. 5189-5200). PMLR.

Kalff, J., & Knoechel, R. (1978). Phytoplankton and their dynamics in oligotrophic and eutrophic lakes. Annual Review of Ecology and Systematics, 9(1), 475-495.

Karsoliya, S. (2012). Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. International Journal of Engineering Trends and Technology, 3(6), 714-717.

Khan, M. N., & Mohammad, F. (2014). Eutrophication: challenges and solutions. Eutrophication: Causes, Consequences and Control: Volume 2, 1-15.

Kim, H. R., Soh, H. Y., Kwak, M. T., & Han, S. H. (2022). Machine learning and multiple imputation approach to predict Chl-a concentration in the coastal zone of Korea. Water, 14(12), 1862.

Kim, T., Shin, J., Lee, D., Kim, Y., Na, E., Park, J. H., ... & Cha, Y. (2022). Simultaneous feature engineering and interpretation: Forecasting harmful algal blooms using a deep learning approach. Water Research, 215, 118289.

Kitsiou, D., & Karydis, M. (2011). Coastal marine eutrophication assessment: A review on data analysis. Environment International, 37(4), 778-801.

Kotsiantis, S. B. (2013). Decision trees: a recent overview. Artificial Intelligence Review, 39, 261-283.

Lahr, J., & Kooistra, L. (2010). Environmental risk mapping of pollutants: State of the art and communication aspects. Science of the Total Environment, 408(18), 3899-3907.

Lee, D.-J., Park, S.-Y., Jung, N.-C., Lee, H.-K., Park, J.-I., & Chun, M.-G. (2006). Chl-a Forecasting using PLS Based c-Fuzzy Model Tree. Journal of The Korean Institute of Intelligent Systems, 16, 777-784.

Li, D., Chen, X., Becchi, M., & Zong, Z. (2016, October). Evaluating the energy efficiency of deep convolutional neural networks on CPUs and GPUs. In 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom) (pp. 477-484). IEEE.

Li, X., Sha, J., & Wang, Z. L. (2018). Application of feature selection and regression models for Chl-a prediction in a shallow lake. Environmental Science and Pollution Research, 25, 19488-19498.

Li, X., Sha, J., & Wang, Z.-L. (2018). Application of feature selection and regression models for Chl-a prediction in a shallow lake. Environmental Science and Pollution Research, 25, 19488-19498.

Li, X., Sha, J., & Wang, Z.-L. (2018). Application of feature selection and regression models for Chl-a prediction in a shallow lake. Environmental Science and Pollution Research.

Liu, J., Zhang, Y., & Qian, X. (2009, June). Modeling Chl-a in Taihu Lake with machine learning models. In 2009 3rd International Conference on Bioinformatics and Biomedical Engineering (pp. 1-6). IEEE.

Liu, Y., & Wu, H. (2017). Water bloom warning model based on random forest. 2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), 45-48.

Liu, Y., & Wu, H. (2017, November). Water bloom warning model based on random forest. In 2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS) (pp. 45-48). IEEE.

Ly, Q. V., Nguyen, X. C., Lê, N. C., Truong, T. D., Hoang, T. H. T., Park, T. J., ... & Hur, J. (2021). Application of machine learning for eutrophication analysis and algal bloom prediction in an urban river: A 10-year study of the Han River, South Korea. Science of the Total Environment, 797, 149040.

Maimaitijiang, M. (2020). Multimodal remote sensing data fusion and machine learning for crop monitoring and food security (Doctoral dissertation, Saint Louis University).

Mamun, M., Kim, J. J., Alam, M. A., & An, K. G. (2019). Prediction of algal Chl-a and water clarity in monsoon-region reservoir using machine learning approaches. Water, 12(1), 30.

Merghadi, A., Yunus, A. P., Dou, J., Whiteley, J., ThaiPham, B., Bui, D. T., ... & Abderrahmane, B. (2020). Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. Earth-Science Reviews, 207, 103225.

Moustafa, E. B., & Elsheikh, A. H. (2023). Predicting Characteristics of Dissimilar Laser Welded Polymeric Joints Using a Multi-Layer Perceptrons Model Coupled with Archimedes Optimizer. Polymers, 15.

Mubarok, R. A., Maylawati, D., Alam, C., Gerhana, Y. A., Zulfikar, W. B., & Saputra, M. I. N. (2023). Prediction of interest in advanced studies using the K-nearest neighbor. 2023 9th International Conference on Wireless and Telematics (ICWT).

Mustapha, I. B., & Saeed, F. (2016). Bioactive molecule prediction using extreme gradient boosting. Molecules, 21.

Nieto, P., Gonzalo, E. G., Fernández, J. R. A., & Muñiz, C. D. (2016). A hybrid PSO optimized SVM-based model for predicting a successful growth cycle of the Spirulina platensis from raceway experiments data. J. Comput. Appl. Math., 291, 293-303.

Nigsch, F., Bender, A., van Buuren, B., Tissen, J., Nigsch, E., & Mitchell, J. B. O. (2006). Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization. Journal of Chemical Information and Modeling, 46(6), 2412-2422.

Oliver, S., Corburn, J., & Ribeiro, H. (2019). Challenges regarding water quality of eutrophic reservoirs in urban landscapes: a mapping literature review. International Journal of Environmental Research and Public Health, 16(1), 40.

Ostendorf, B. (2011). Overview: Spatial information and indicators for sustainable management of natural resources. Ecological Indicators, 11(1), 97-102.

Otchere, D. A., Ganat, T. O. A., Gholami, R., & Ridha, S. (2021). Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. Journal of Petroleum Science and Engineering, 200, 108182.

Oyedele, A., Ajayi, A., Oyedele, L. O., Delgado, J. M. D., Akanbi, L., Akinade, O., ... & Bilal, M. (2021). Deep learning and boosted trees for injuries prediction in power infrastructure projects. Applied Soft Computing, 110, 107587.

Park, J., Kim, K. T., & Lee, W. H. (2020). Recent advances in information and communications technology (ICT) and sensor technology for monitoring water quality. Water, 12(2), 510.

Park, J., Lee, W. H., Kim, K. T., Park, C. Y., Lee, S., & Heo, T. Y. (2022). Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. Science of the Total Environment, 832, 155070.

Park, Y., Cho, K. H., Park, J., Cha, S. M., & Kim, J. H. (2015). Development of early-warning protocol for predicting Chl-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Science of the Total Environment, 502, 31-41.

Park, Y., Cho, K., Park, J., Cha, S., & Kim, J. (2015). Development of early-warning protocol for predicting Chl-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. The Science of the Total Environment, 502, 31-41.

Parry, R. M., Jones, W. D., Stokes, T. H., Phan, J., Moffitt, R. A., Fang, H., Shi, L., Oberthuer, A., Fischer, M., Tong, W., & Wang, M. D. (2010). k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. The Pharmacogenomics Journal, 10(4), 292-309.

Paerl, H. W., & Huisman, J. (2020). Climate change: A catalyst for global expansion of harmful cyanobacterial blooms. *Environmental Microbiology Reports*, 12(1), 3-4. https://doi.org/10.1111/1758-2229.12843

Phaisangittisagul, E. (2016, January). An analysis of the regularization between L2 and dropout in single hidden layer neural network. In 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS) (pp. 174-179). IEEE.

Picos, A., & Peralta-Hernández, J. (2018). Genetic algorithm and Extreme Gradient Boosting model for prediction of discoloration dye from an electro-oxidation process in a press-type reactor. Water Science and Technology: A Journal of the International Association on Water Pollution Research.

Rajaee, T., & Boroumand, A. (2015). Forecasting of Chl-a concentrations in South San Francisco Bay using five different models. Applied Ocean Research, 53, 208-217.

Ramadas, M., & Samantaray, A. K. (2018). Applications of remote sensing and GIS in water quality monitoring and remediation: A state-of-the-art review. In Water Remediation (pp. 225-246).

Ray, S. (2019, February). A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 35-39). IEEE.

Rybka, K., Janaszek-Mańkowska, M., Siedlarz, P., & Mańkowski, D. (2019). Machine learning in determination of water saturation deficit in wheat leaves on basis of Chl a fluorescence parameters. Photosynthetica, 57(1), 226-230.

Salditt, M., Humberg, S., & Nestler, S. (2023). Gradient tree boosting for hierarchical data. Multivariate Behavioral Research, 58, 911-937.

Santana-Falcón, Y., Brasseur, P., Brankart, J. M., & Garnier, F. (2020). Assimilation of chlorophyll data into a stochastic ensemble simulation for the North Atlantic Ocean. Ocean Science.

Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D., & Džeroski, S. (2010). Predicting gene function using hierarchical multi-label decision tree ensembles. BMC Bioinformatics, 11, 2.

Shen, M., Luo, J., Cao, Z., Xue, K., Qi, T., Ma, J., ... & Duan, H. (2022). Random forest: An optimal Chl-a algorithm for optically complex inland water suffering atmospheric correction uncertainties. Journal of Hydrology, 615, 128685.

Shi, S., Wang, Q., Xu, P., & Chu, X. (2016, November). Benchmarking state-of-the-art deep learning software tools. In 2016 7th International Conference on Cloud Computing and Big Data (CCBD) (pp. 99-104). IEEE.

Shin, J., Yoon, S., & Cha, Y. (2017). Prediction of cyanobacteria blooms in the lower Han River (South Korea) using ensemble learning algorithms. Desalination and Water Treatment, 84, 31-39.

Shin, Y., Kim, T., Hong, S., Lee, S., Lee, E., Hong, S., ... & Heo, T. Y. (2020). Prediction of Chl-a concentrations in the Nakdong River using machine learning methods. Water, 12(6), 1822.

Shin, Y., Kim, T., Hong, S., Lee, S., Lee, E., Hong, S.-W., Lee, C., Kim, T., Park, M., Park, J., & Heo, T.-Y. (2020). Prediction of Chl-a Concentrations in the Nakdong River Using Machine Learning Methods. Water.

Smith, B., Pahlevan, N., Schalles, J., Ruberg, S., Errera, R., Ma, R., ... & Kangro, K. (2021). A Chl-a algorithm for Landsat-8 based on mixture density networks. Frontiers in Remote Sensing, 1, 623678.

Smith, V. H., & Schindler, D. W. (2009). Eutrophication science: where do we go from here? Trends in Ecology & Evolution, 24(4), 201-207.

Stanimirova, I., Üstün, B., Cajka, T., Riddelova, K., Hajslova, J., Buydens, L. M. C., & Walczak, B. (2010). Tracing the geographical origin of honeys based on volatile compounds profiles assessment using pattern recognition techniques. Food Chemistry, 118(1), 171-176.

Stumpf, R. P., Davis, T. W., Wynne, T. T., Graham, J. L., Loftin, K. A., Johengen, T. H., ... & Burtner, A. (2016). Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria. Harmful Algae, 54, 160-173.

Su, J., Wang, X., Zhao, S., Chen, B., Li, C.-H., & Yang, Z. (2015). A Structurally Simplified Hybrid Model of Genetic Algorithm and Support Vector Machine for Prediction of Chlorophyll a in Reservoirs. Water, 7, 1610-1627.

Sundararajan, K., Garg, L., Srinivasan, K., Bashir, A. K., Kaliappan, J., Ganapathy, G. P., ... & Meena, T. (2021). A contemporary review on drought modeling using machine learning approaches. CMES-Computer Modeling in Engineering and Sciences, 128(2), 447-487.

Tahmasebi, P., Kamrava, S., Bai, T., & Sahimi, M. (2020). Machine learning in geo- and environmental sciences: From small to large scale. Advances in Water Resources, 142, 103619.

Talib, A. (2006). Comparative ecological study of two Dutch lakes using computational modelling (Doctoral dissertation).

Tian, W., Liao, Z., & Wang, X. (2019). Transfer learning for neural network model in Chl-a dynamics prediction. Environmental Science and Pollution Research, 26, 29857-29871.

Tian, W., Liao, Z., & Wang, X. (2019). Transfer learning for neural network model in Chl-a dynamics prediction. Environmental Science and Pollution Research.

Tian, W., Liao, Z., & Zhang, J. (2017). An optimization of Extreme Gradient Boosting model for predicting chlorophyll dynamics. Ecological Modelling, 364, 42-52.

Tjärnberg, A., Mahmood, O., Jackson, C. A., Saldi, G. A., Cho, K., Christiaen, L. A., & Bonneau, R. A. (2021). Optimal tuning of weighted kNN-and diffusion-based methods for denoising single cell genomics data. PLoS Computational Biology, 17(1), e1008569.

Tomppo, E., Gagliano, C., Natale, F. D., Katila, M., & McRoberts, R. E. (2009). Predicting categorical forest variables using an improved k-Nearest Neighbour estimator and Landsat imagery. Remote Sensing of Environment, 113(1), 500-517.

Tsihrintzis, V. A., Hamid, R., & Fuentes, H. R. (1996). Use of geographic information systems (GIS) in water resources: A review. Water Resources Management, 10, 251-277.

Tu, J. V. (1996). Advantages and disadvantages of using Extreme Gradient Boostings versus logistic regression for predicting medical outcomes. Journal of Clinical Epidemiology, 49(11), 1225-1231.

Tung, T. M., & Yaseen, Z. M. (2020). A survey on river water quality modelling using artificial intelligence models: 2000–2020. Journal of Hydrology, 585, 124670.

Turner, R. K., Morse‐Jones, S., & Fisher, B. (2010). Ecosystem valuation: A sequential decision support system and quality assessment issues. Annals of the New York Academy of Sciences, 1185(1), 79-101.

Urbanovich, E., Afonnikov, D. A., & Nikolaev, S. V. (2021). Determination of the quantitative content of chlorophylls in leaves by reflection spectra using the random forest algorithm. Vavilov Journal of Genetics and Breeding, 25, 64-70.Jia, W., Cheng, J., & Hu, H. (2020). A Cluster-Stacking-Based Approach to Forecasting Seasonal Chl-a Concentration in Coastal Waters. IEEE Access, 8, 99934-99947.

Wei, Y., Huang, H., Chen, B., Zheng, B., & Wang, Y. (2019). Application of extreme learning machine for predicting Chl-a concentration in artificial upwelling processes. Mathematical Problems in Engineering, 2019.

Winter, J. G., & Duthie, H. C. (2000). Epilithic diatoms as indicators of stream total N and total P concentration. Journal of the North American Benthological Society, 19(1), 32-49.

Wu, C., Fu, X., Li, H., Hu, H., Li, X., & Zhang, L. (2023). A Method Based on Improved Ant Colony Algorithm Feature Selection Combined With GA-SVR Model for Predicting Chl-a Concentration in Ulansuhai Lake. IEEE Access, 11, 93180-93192.

Wu, D., Li, R., Zhang, F., & Liu, J. (2019). A review on drone-based harmful algae blooms monitoring. Environmental Monitoring and Assessment, 191, 1-11.

Wujek, B., Hall, P., & Günes, F. (2016). Best practices for machine learning applications. SAS Institute Inc.

Yajima, H., & Derot, J. (2018). Application of the Random Forest model for Chl-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. Journal of Hydroinformatics, 20, 206-220.

Yang, L., Chen, Y., & Zhang, L. (2022). Research on an improved prediction model based on decision tree algorithm. 2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA), 206-209.

Yang, L., Driscol, J., Sarigai, S., Wu, Q., Lippitt, C. D., & Morgan, M. (2022). Towards synoptic water monitoring systems: A review of AI methods for automating water body detection and water quality monitoring using remote sensing. Sensors, 22(6), 2416.

Yao, H., Huang, Y., Wei, Y., Zhong, W., & Wen, K. (2021). Retrieval of Chl-a concentrations in the coastal waters of the Beibu Gulf in Guangxi using a gradient-boosting decision tree model. Applied Sciences, 11(17), 7855.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., ... & Zhang, L. (2020). Deep learning in environmental remote sensing: Achievements and challenges. Remote Sensing of Environment, 241, 111716.

Zamyadi, A., Dorner, S., Sauvé, S., Ellis, D., & Boland, S. (2019). Cyanobacterial blooms and drinking water: Impacts, detection, and treatment. *Science of The Total Environment*, 682, 600-617. https://doi.org/10.1016/j.scitotenv.2019.05.244

Zennaro, E., Servadei, L., Devarajegowda, K., & Ecker, W. (2018, August). A machine learning approach for area prediction of hardware designs from abstract specifications. In 2018 21st Euromicro Conference on Digital System Design (DSD) (pp. 413-420). IEEE.

Zhang, T., Huang, M., & Wang, Z. (2020). Estimation of Chl-a Concentration of lakes based on SVM algorithm and Landsat 8 OLI images. Environmental Science and Pollution Research, 27, 14977-14990.

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies, 58, 308-324.

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., ... & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. Eco-Environment & Health.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., & Hinkelmann, R. (2021). Ensemble machine learning paradigms in hydrology: A review. Journal of Hydrology, 598, 126266.