Thesis for the Degree of Master of Engineering

Estimation of Chlorophyll-a Concentration in the Nakdong River Using Sentinel-2 MSI and Algal Bloom Influence Factors Data Fusion: Comparison and Evaluation of Machine Learning Models Performance

by

So Ryeon Park

Division of Earth Environmental System Science

(Major of Spatial Information Engineering)

The Graduate School

Pukyong National University

February, 2024

Estimation of Chlorophyll-a Concentration in the Nakdong River Using Sentinel-2 MSI and Algal Bloom Influence Factors Data Fusion: Comparison and Evaluation of Machine Learning Models Performance
(Sentinel-2 MSI 와 녹조 영향인자를 융합 활용한 낙동강 chlorophyll-a 농도 추정: 머신러닝 모델 성능 비교 및 평가)

Advisor: Prof. Jin-Soo Kim

by
So Ryeon Park

A thesis submitted in partial fulfillment of the requirements
for the degree of
Master of Engineering
in Division of Earth Environmental System Science
(Major of Spatial Information Engineering), The Graduate School,
Pukyong National University

February, 2024

Estimation of Chlorophyll-a Concentration in the Nakdong River Using Sentinel-2 MSI and Algal Bloom Influence Factors Data Fusion: Comparison and Evaluation of Machine Learning Models Performance

A dissertation

by

So Ryeon Park

Approved by:

_____

(Chairman) Prof. Kyung-soo Han

_____

(Member) Prof. Yang-won Lee

_____

(Member) Prof. Jin-soo Kim

February 16, 2024

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ANN	Artificial Neural Network
AutoML	Automated Machine Learning
BOD	Biochemical Oxygen Demand
CatBoost	Categorical Boosting
DO	Dissolved Oxygen
GPR	Gaussian Process Regression
LightGBM	Light Gradient Boosting Machine
LR	multiple Linear Regression
MLP	Multi-Layer Perceptron
MSI	Multi Spectral Instrument
OLI	Operational Land Imager
OOB	Out-Of-Bag
SS	Suspended Solids
SVR	Support Vector Regression
TOC	Total Organic Carbon
XAI	eXplainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting

Estimation of Chlorophyll-a Concentration in the Nakdong River Using Sentinel-2 MSI and Algal Bloom Influence Factors Data Fusion : Comparison and Evaluation of Machine Learning Models Performance

So Ryeon Park

Division of Earth Environmental System Science (Major of Spatial Information Engineering), The Graduate School,
Pukyong National University

ABSTRACT

Due to global warming, the frequency and extent of algal blooms are increasingly prevalent worldwide. In Korea, the Nakdong River faces severe algal bloom issues every year. Algal blooms cause various damages such as ecological, economic, and aesthetic damages, periodic monitoring is essential for preemptive management and rapid response. Chlorophyll-a (chl-a) concentration is utilized as an indicator of algal bloom occurrences, and the use of satellite enables the detection of algal blooms over extensive areas. Numerous studies recently have utilized machine learning techniques to achieve more accurate estimations of chl-a concentrations. Various factors affect algal blooms occurrence, and

identifying the cause of their occurrence remains challenging. Therefore, it is essential to apply a study using diverse input data to comprehensively consider these factors. This study fused Sentinel-2 satellite data with water quality, meteorological, and hydrological factors data to estimate chlorophyll-a concentrations for eight weirs along the Nakdong River over the last five years. AutoML selected six models (CatBoost, Extra Trees, Gradient Boosting, LightGBM, Random Forest, and XGBoost), and the SHAP method identified a total of 27 fused input variables. CatBoost (R2 = 0.862, RMSE = 5.560 mg/m3, MAE = 4.120 mg/m3) demonstrated superior performance, and all six models achieved significant results with R2 above 0.8. SHAP was conducted to analyze the importance of features, and Suspended Solids (SS) emerged as the most important factor in all six models. The ranking of variable importance varied by model, water quality variables such as Dissolved Oxygen (DO), pH, Total Organic Carbon (TOC), and Biochemical Oxygen Demand (BOD), and satellite variables using the band combinations of red-edge and red band were identified as common top-ranking variables. The feasibility of chl-a estimation was assessed by exhibiting the spatial patterns of the estimated chl-a values using CatBoost. This study confirmed the applicability of fusion data for estimating chl-a concentrations, and it is expected to be utilized for nationally and globally chl-a monitoring in the future.

1. Introduction

1.1. Background

An algal bloom is a phenomenon that changes the color of water bodies (rivers, lakes, etc.) to green, primarily due to the extensive proliferation of cyanobacteria. This is distinct from red tide, characterized by a reddish hue resulting from the extensive proliferation of flagellates and diatoms (Lim et al., 2020). Algae growth is responsive to various environmental factors, including temperature, precipitation, and water residence time. In water bodies with slow or stagnant flow rates, the accumulation of nutrients, such as nitrogen and phosphorus, which are vital for algae growth, leads to eutrophication. This nutrient build-up, in turn, triggers the proliferation of algae. In rivers, algal blooms are highly influenced by hydrological factors, as opposed to in static water bodies like lakes and reservoirs. Numerous studies have shown that hydrological conditions, such as water flow, significantly influence the migration, diffusion, and accumulation of algae (Xia et al., 2020).

Algal blooms have become increasingly prevalent worldwide. In recent years, their frequency and extent have risen globally, driven by the interactive effects of multiple stressors. These stressors include climate

change resulting from global warming and various human-induced factors like wastewater discharge and urbanization (Rodríguez-López et al., 2023; Zhou et al., 2021). According to Dai et al. (2023), the spatial extent of algal blooms expanded by 13.2%, and the frequency increased by 59.2% from 2003 to 2020. The deterioration of water quality caused by algal blooms results in adverse environmental and economic impacts, including reduced water transparency, increased water treatment costs, and restrictions on water-related recreational activities (Pretty et al., 2003). Additionally, the cyanobacteria responsible for algal blooms produce odorous substances, causing discomfort and affecting the taste of tap water. Certain cyanobacteria species also produce toxins, posing health risks to both humans and animals (Jeon et al., 2015). With ongoing global warming and the persistence of human-induced factors, the deleterious effects associated with algal blooms are expected to intensify.

In Korea, algal blooms persist as an annual issue, causing ongoing environmental damage. To tackle this issue, the Ministry of Environment actively monitors algal bloom outbreaks using the algal bloom alert system. This system relies on weekly measurements of the concentration of chlorophyll-a (chl-a), a photosynthetic pigment indicative of algal blooms and eutrophication states, as well as pH, and cyanobacteria cell count. Based on these parameters, the system issues algae warnings

classified into levels of caution, warning, and bloom. Notably, the Nakdong River, one of the four major rivers in Korea, faces a particularly severe problem. In 2022, this river experienced 700 days of algae alerts out of the total number of algae alert days, encompassing caution, warning, and bloom levels (ME, 2022).

The escalating frequency and severity of algal blooms, both globally and in Korea, underscore the need for monitoring, research, and proactive management. To effectively anticipate and minimize the impact of algal blooms, continuous monitoring is essential. This requires the development of an accurate model for estimating algal bloom occurrence, taking into account various influencing factors, such as meteorological and hydrological conditions. Therefore, comprehensive monitoring that considers a diverse range of influencing factors is crucial for effective management.

1.2. Literature review

Over the past several decades, various methodologies have been applied in research to monitor algal blooms. Numerous studies have focused on estimating the concentration of chl-a, enabling the assessment of the spatial extent and severity of algal blooms (Kim et al., 2022).

Remote sensing methods for algal bloom detection have been widely utilized in various studies due to their capacity for efficiently monitoring extensive areas regularly, a notable advantage over the classical method of field sampling (Shi et al., 2022b; Zhang et al., 2019). Numerous studies have sought to estimate chl-a concentration by applying band ratio combination algorithms and spectral indices, taking into account the spectral features of chl-a (Park et al., 2018; Rodríguez-López et al., 2023; Zhou et al., 2021). Chl-a exhibits low reflectance in the blue and red bands, contrasting with high reflectance in the green and red edge bands. Notably, the red edge band is characterized by exceptionally high reflectance (Jensen, 2006). Compared to sensors like the Medium Resolution Imaging Spectrometer (MERIS) and Moderate Resolution Imaging Spectroradiometer (MODIS) (~0.3-1 km), the Sentinel-2 Multispectral Instrument (MSI) (10-20 m) and Landsat-8 Operational Land Imager (OLI) (30 m) provide relatively high spatial resolution. Moreover, the Sentinel-

2 MSI enables periodic monitoring of small water bodies with high temporal resolution (revisit time of approximately five days) and provides spectral data from the red edge region, crucial for chl-a concentration estimation. Sentinel-2 MSI imagery is more suitable for monitoring algal blooms in relatively narrow inland water bodies.

Satellite-based methods for estimating chl-a concentration are increasingly incorporating machine learning techniques to address issues such as atmospheric calibration errors and the complex spectral characteristics that vary with water quality. Machine learning methods offer the advantage of estimating chl-a concentrations in water bodies with diverse spectroscopic characteristics and are not sensitive to atmospheric calibration errors (Li et al., 2021). Furthermore, machine learning facilitates efficient analysis and processing of data related to various factors, enabling rapid model construction and computation. Consequently, machine learning has gained widespread application in studies focusing on water quality prediction (Kim et al., 2021). Hafeez et al. (2019) employed Landsat-5 Thematic Mapper (TM), Landsat-7 Enhanced Thematic Mapper Plus (ETM+), and Landsat-8 OLI imagery, utilizing four machine learning models—Support Vector Regression (SVR), Artificial Neural Network (ANN), Cubist Regression Trees (CB), and Random Forest—to estimate chl-a concentrations in coastal waters of

Hong Kong. In another study, Rodríguez-López et al. (2020) utilized Landsat-8 OLI data along with multiple Linear Regression (LR) models to estimate chl-a concentrations in Lake Villarrica, Chile. Shi et al. (2022b) applied LR, ANN, multiple Bayesian Regression (BR) and Extreme Gradient Boosting (XGBoost) models to estimate chl-a concentrations in small water bodies belong to Beijing, leveraging satellite imagery from Sentinel-2 and Gaofen-6. Another study employed Random Forest and XGBoost models with Sentinel-2 imagery to estimate chl-a concentrations in Chagan Lake, China (Shi et al., 2022a). Additionally, Kim et al. (2022) utilized Sentinel-2 MSI imagery and five models—Random Forest, SVR, Multilayer Perceptron (MLP), Gaussian Process Regression (GPR), and Light Gradient Boosting Machine (LightGBM)—to predict chl-a concentrations in 78 different water bodies belonging to four rivers in Korea.

Machine learning-based studies integrate water quality, meteorology, and hydrology data, along with satellite imagery, to investigate the multitude of factors influencing algal bloom occurrences. In one study, five models—RF, GPR, XGBoost, SVR, and Categorical Boosting (CatBoost)—were employed to estimate chl-a concentration in the artificial Tri An Reservoir in Vietnam (Nguyen et al., 2022). The input data were divided into two cases: water quality factors and Sentinel-2 MSI

imagery, and the results were compared. Chen et al. (2020) utilized water quality factors data and a total of 10 models, including Random Forest, LR, SVR, and Decision Tree (DT), to estimate chl-a concentrations in rivers and lakes in China. In another study, Kim and Park (2023) used water quality factors data and applied CatBoost, XGBoost, and LightGBM to estimate chl-a concentrations in Daecheong Lake of the Geum River, Korea. Another study applied Random Forest, SVR, and ANN to estimate chl-a concentration in the Han River, Korea, using water quality and meteorological factors data (Kim and Ahn, 2022). Additionally, a study in the Nakdong River selected final input variables, combining collected water quality, meteorological, and hydrological factor data, through forward selection methods, and applied five models—Random Forest, SVR, XGBoost, Recurrent Neural Network (RNN), and Long-Short-Term Memory (LSTM)—to estimate chl-a concentration (Shin et al., 2020).

Collectively, machine learning-based studies aim to develop models that accurately estimate field-measured chl-a concentrations. These studies can be categorized into those utilizing satellite data with chl-a spectral properties as input and those using water quality, weather, and hydrology data. However, there remains a deficiency in research that comprehensively considers various factors influencing algal bloom development. Algal blooms occur due to a combination of complex factors,

the need for studies that account for a diverse set of input variables to achieve more accurate estimations of chlorophyll-*a* concentrations.

The range of machine learning applications is expanding, and the process of selecting, training, and optimizing a suitable model based on specific objectives and data characteristics requires considerable time investment and background knowledge. Automated machine learning (AutoML) is a technology that automates the complex process of model construction, spanning from data preprocessing to model training and evaluation. This innovation enables the simultaneous testing of various models with just a few lines of code and facilitates accessible development even for non-experts. Numerous studies are currently assessing the applicability of AutoML based on these advantages (Waring et al., 2020). However, Musigmann et al. (2022) confirmed that traditional machine learning methods exhibit higher and more stable performance levels and interpretability than AutoML. In addition, during the data pre-processing stage, numerous aspects necessitate human intervention. Therefore, it is crucial to judiciously employ AutoML for specific research purposes. In this study, we applied AutoML techniques for model selection to streamline the model-building process and reduce time consumption.

Machine learning finds extensive application across diverse fields, including image processing, classification, and predictive monitoring.

With the growing interest in machine learning, there is an increasing focus not only on performance improvement but also on the interpretability of models. Many models developed to address complex problems demonstrate high performance, but they may lack an intuitive interpretation of their structure and processes. Consequently, Explainable Artificial Intelligence (XAI) technology has been developed to help users understand the system and final results of models. Interpreting models through XAI enables users to gain trust and ensures effective model construction and management (Arrieta et al., 2020). One XAI method, Shapley Additive Explanations (SHAP), calculates and provides the contribution of each input variable and aids in the understanding of prediction results through visualization. Various studies utilize the SHAP method for identifying important variables and analyzing the impact of input variables on the model (Kim et al., 2022; Shi et al., 2022a).

1.3. Objectives

This study aimed to develop a model suitable for estimating chl-a concentrations in the Nakdong River, Korea, by integrating satellite data with water quality, meteorological, and hydrological factor data. To achieve this goal, the following steps were performed: (1) AutoML was utilized to select a learning model and conduct initial model-specific training; (2) the optimal combination of input variables for the chl-a estimation model through SHAP value analysis was determined; (3) secondary model-specific training was conducted, with the results being compared and analyzed; (4) factor importance analysis was conducted for each model using the SHAP method, followed by further analysis to interpret the model results; (5) spatial distribution map of the estimated chl-a concentration was generated. The overall research flow chart is depicted in Fig. 1.

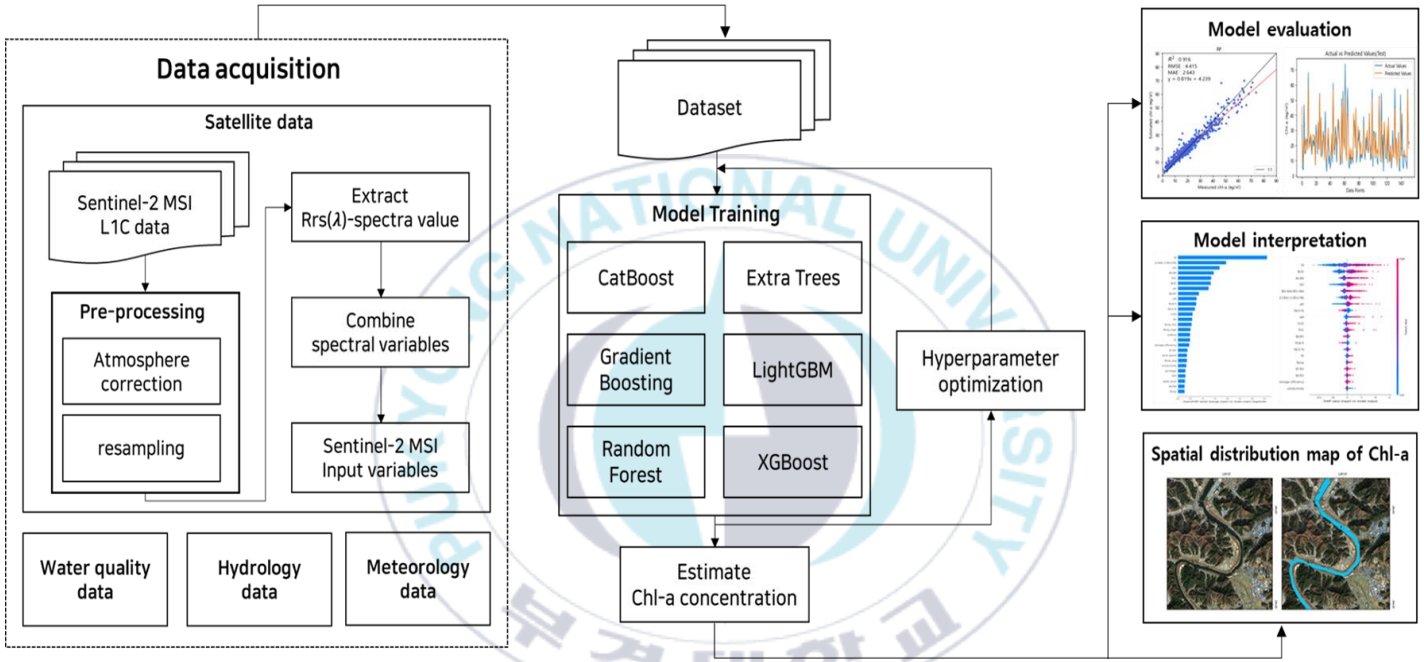


Fig. 1. Flow chart of this study.

2. Study Area and Data

2.1. Study Area

The study area encompassed on eight weirs situated along the Nakdong River watershed, located in order from upstream to downstream: Sang-ju, Nak-dan, Gu-mi, Chil-gok, Gang-jeong Go-ryeong, Dal-seong, Hap-cheon Chang-nyeong, and Chang-nyeong Ham-an (Fig. 2). In response to the need for water resource management from floods and droughts and the enhancement of water quality, a total of 16 multi-functional weirs were installed on the four major rivers, and eight weirs were installed on the Nakdong River. Following the installation of these weirs, the Nakdong River acquired characteristics of a closed water body, where river flow is controlled, and pollutants accumulate (Lee et al., 2014). Monitoring water quality changes by comparing concentrations of factors such as Dissolved Oxygen (DO), Suspended Solids (SS), and Total Organic Carbon (TOC) before and after the weir installation in the Nakdong River, identified a trend of deteriorating water quality, indicating that the weir installation made the Nakdong River watershed susceptible to algal blooms (Cho et al., 2018). The Nakdong River is grappling with severe eutrophication due to point source pollution, including livestock

wastewater and domestic sewage from the upstream section, as well as pollutants discharged from major cities and industrial areas such as Gu-mi and Dae-gu in the middle and downstream sections (Lee and Kim 2021). The nutrient-rich conditions in the eutrophicated Nakdong River contribute to the growth of cyanobacteria, resulting in significant damage from algal blooms each summer when water temperatures increase. To carry out continuous algal bloom monitoring, the study period was set for the last five years from January 2018 to December 2022.

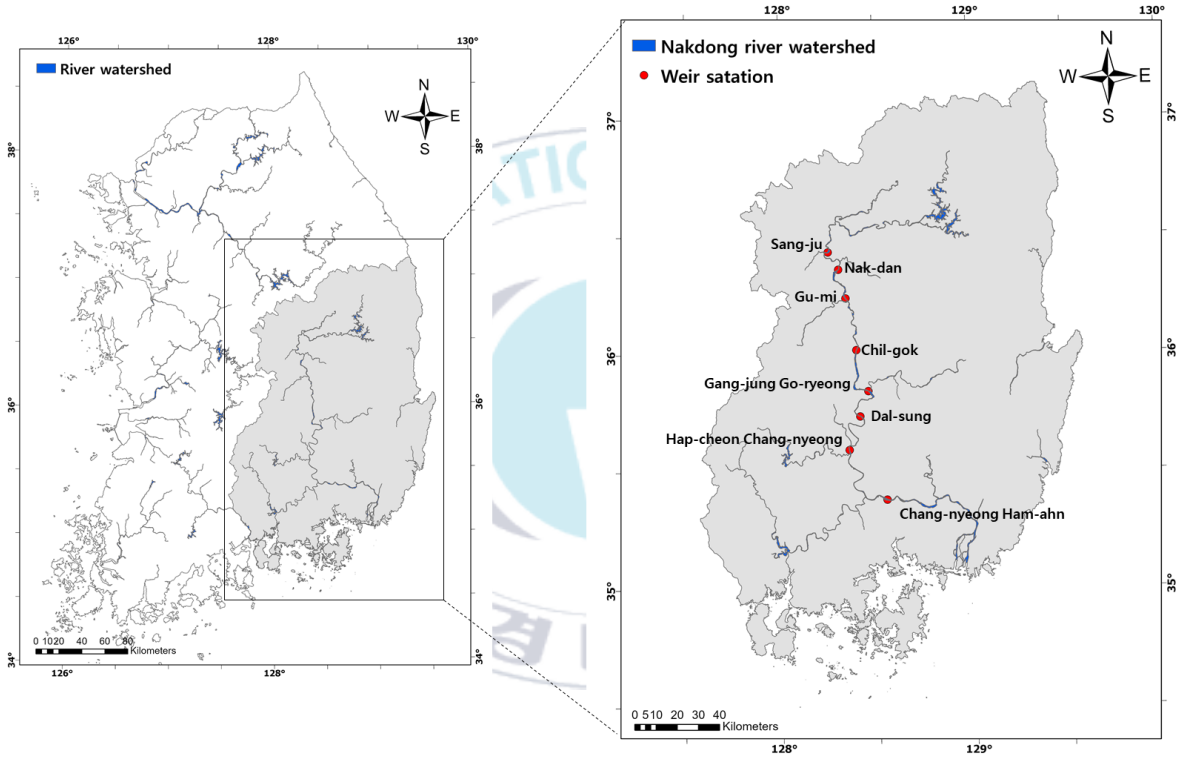


Fig. 2. Study Area: eight weirs along the Nakdong River.

2.2. Data

2.2.1. Satellite data

This study utilized data from the Sentinel-2 MSI sensor provided by the European Space Agency (ESA). Sentinel-2A and Sentinel-2B were launched in June 2015 and March 2017, respectively, and each satellite has a revisit period of 10 days, which can be reduced to 5 days when the two satellites are used together. The Sentinel-2 MSI sensor comprises a total of 13 spectral bands spanning the visible, near-infrared, and shortwave infrared regions, providing data with spatial resolutions of 10, 20, and 60 m (Table 1).

From January 2018 to December 2022, 294 cloud-free Sentinel-2 Level-1C (L1C) images were collected with a date difference of ±24 hours from the field measurement dates of chl-a . The downloaded images were converted to Level-2A (L2A) images utilizing an atmospheric correction processor Sentinel 2 Correction (Sen2Cor) (Version2.10). Recent studies have validated the Sen2Cor processor for its reasonable accuracy in estimating chl-a concentrations in highly turbid water (Kim et al., 2022; Nguyen et al., 2021). The Sen2Cor processor performs atmospheric, terrain, and cirrus correction to produce a bottom-of-atmosphere L2A product with scene classification and aerosol optical thickness. All bands

were resampled to 10 m to calculate bands between 10, 20 and 60 m resolution.

In this study, a total of 13 satellite input variables were used, consisting of six single spectral bands and seven combinations of spectral bands. In previous studies, the band ratio algorithm, which applies the characteristic that the reflectance value of the red edge band is very large, shows a high contribution to the model prediction, so the input variables were selected with this consideration (Table 2).

Table 1. Sentinel-2 MSI band information.

Band index	Description	Central wavelength (nm)	Resolution (m)
Band 1	Costal aerosol	443	60
Band 2	Blue	490	10
Band 3	Green	560	10
Band 4	Red	665	10
Band 5	Red edge	704	20
Band 6	Red edge	740	20
Band 7	Red edge	783	20
Band 8	NIR	842	10
Band 8A	Red edge	865	20
Band 9	Water vapor	945	60
Band 10	Cirrus	1375	60
Band 11	SWIR	1610	20
Band 12	SWIR	2190	20

Table 2. Satellite input variables used to machine learning

Data	Variable names	Variable details	Reference
Spectral bands	B1, B2, B3, B4, B5, B6	Rrs (443), Rrs (492), Rrs (560), Rrs (665), Rrs (704), Rrs (740)	
Two-band ratio	B1 / B3	Rrs (443) / Rrs (560),	(Chavula et al., 2009)
	B2 / B3	Rrs (492) / Rrs (560)	(Moses et al., 2009)
	B3 / B4	Rrs (560) / Rrs (665)	(Ha et al., 2017)
	B5 / B4	Rrs (704) / Rrs (665)	(Gurlin et al., 2011)
	B6 / B5	Rrs (740) / Rrs (704)	(Li et al., 2021)
	(B5-B4) / (B5+B4)	[Rrs (704) – Rrs (665)] / [Rrs (704) + Rrs (665)]	(Gitelson et al., 2008)
Three-band ratio	((1/B4) – (1/B5)) * B6	[Rrs (665)-1 – Rrs (704) -1] × Rrs (740)	(Gitelson et al., 2011)

2.2.2. Water quality data

The water quality data were provided by the Ministry of Environment Water Environment Information System (http://water.nier.go.kr) and are measured once a week upstream of the weir point. The water quality data affected by different water environments are diverse, and were collected based on various water quality factors used in previous studies (Jung et al., 2021; Lee et al., 2020; Nguyen et al., 2021). Therefore, the water quality measurement data utilized in the study included the output variable chl-a and a total of 16 input variables Temp, pH, DO, cell, BOD, COD, TN, TP, TOC, conductivity, DTN, NH_3-N, NO_3-N, DTP, PO_4-P, and SS as detailed in the Table 3.

2.2.3. Meteorology data

The meteorological factor data were obtained from the nearest measured point data within a 10-kilometer radius of the weir point, provided by the Korea Meteorological Administration weather data open portal (https://data.kma.go.kr). Meteorological factor data are provided on a minute-by-minute, hourly, and daily basis, and in this study, we used data measured on the same day as the chl-a data. According to Zhou et al. (2021), wind speed and air temperature have been identified as crucial meteorological factors influencing algal blooms. Rising air temperatures

attributed to climate warming considered a likely cause for the increased frequency and extent of algal blooms (Jung et al., 2021; Kosten et al., 2012). Therefore, the meteorological measurement data utilized in the study temp_avg, temp_high, temp_low, and wind_speed as shown in the Table 3.

2.2.4. Hydrology data

The hydrological data are measured from water level observation points according of the weir point and were provided by Mywater water information portal (https://www.water.or.kr/). Hydrological data are on a measured every 10 minutes and hourly, daily basis, and this study, we used data measure on the same day as the chl-a data. Since 2017, the Ministry of Environment has been improving water flow through weirs opening monitoring to address issues arising from increased algal blooms after weirs installation. It is necessary to consider hydrological data, including weir operational data, in estimating chl-a concentrations. Therefore, the hydrological measured data utilized in the study water level, pondage, storage efficiency, rainfall, inflow, and outflow as shown in the Table 3.

Table 3. Water quality, meteorology, hydrology input variables used to machine learning.

Category	Variable names	Variable details
Water quality	Temp	Water temperature (°C)
	pH	Potential of hydrogen
	DO	Dissolved oxygen (mg/L)
	Cell	Cyanobacteria cells (cells/mL)
	BOD	Biochemical oxygen demand (mg/L)
	COD	Chemical oxygen demand (mg/L)
	TN	Total nitrogen (mg/L)
	TP	Total phosphorus (mg/L)
	TOC	Total organic carbon (mg/L)
	Conductivity	Electric conductivity (μS/cm)
	DTN	Dissolved total nitrogen (mg/L)
	NH_3-N	Ammonia nitrogen (mg/L)
	NO_3-N	Nitrate nitrogen (mg/L)
	DTP	Dissolved total phosphorus (mg/L)
	PO4-P	Phosphate phosphorus (mg/L)
	SS	Suspended solids (mg/L)
Meteorology	Temp_avg	1-day average of air temperature (°C)
	Temp_high	Maximum air temperature (°C)
	Temp_low	Minimum air temperature (°C)
	Wind_speed	1-day average of wind speed (m/s)
Hydrology	Water level	Water level of the weir (EL.m)
	Pondage	Water storage capacity (100,000,000 m^3)
	Storage efficiency	Water storage rate (%)
	Rainfall	Rainfall of weir region (mm)
	Inflow	Inflow amount of water (m^3/s)
	Outflow	Outflow amount of water (m^3/s)

3. Methodology

3.1. AutoML

There are various tools for AutoML, such as Auto_keras, Auto-sklearn, Pycaret, and H_2O AutoML (Ferreira et al., 2022). In this study, the chl-a concentration estimation model was selected using Pycaret, an open source Python library with built-in packages such as classification, clustering, and regression. It provides results by comparing the performance of various models such as SVR, MLP, Random Forest, and XGBoost. Ensemble results for the top N models, based on predefined metrics, can also be obtained.

AutoML was employed to randomize the train and test datasets 8:2 ratio for all input variable data, and then applied the built-in regression model to compare the results. The top-performing six models (Catboost, Extra trees, Gradient boosting, LightGBM, Random Forest, and XGBoost) were then selected and built as a chl-a estimation model for eight weirs of the Nakdong River.

3.2. Machine Learning Method

The six machine learning models identified through AutoML can be categorized into two prominent ensemble model learning methodologies: bagging algorithms and boosting algorithms. Ensemble methods, which amalgamate multiple learning models to construct a final model with enhanced predictive capabilities. Random Forest and Extra Trees belong to bagging algorithm. Conversely, Gradient Boosting, XGBoost, LightGBM, and CatBoost belong to boosting algorithm.

Bootstrap aggregating, or Bagging, is a technique for data generation and modeling in which the bootstrap method is applied to produce a conclusive predictive model. Bootstrap, a statistical technique, involves the random generation of data, allowing for duplicates from the provided dataset, thereby constructing a dataset of equivalent size through resampling extraction. The outcomes of multiple decision trees are amalgamated to derive a single prediction result.

In contrast to bagging, boosting is a sequential learning method that enhances model performance by assigning weights to next learner based on the learning outcomes of the previous one. The objective is to progressively combine the learning outcomes of weak learners, ultimately create a strong learner. By assigning weights to incorrect predictions, the

boosting technique improves prediction accuracy, thereby reducing errors through repeated training. However, it may be susceptible to the influence of outliers.

3.2.1. Bagging algorithm

Random Forest, a representative bagging-based model, has found widespread application in numerous studies due to its uncomplicated structure and high efficiency (Breiman, 2001). This approach utilizes bootstrap samples derived from the training data to construct multiple decision trees in a randomized fashion, consolidating their outcomes to craft an optimal model. The model performance evaluation is achievable through the computation of the out-of-bag (OOB) error, attained by training with the out-of-bag subtraction of unrecovered samples through bootstrap and subsequent validation with the OOB (Nguyen et al., 2021). It is widely utilized in water quality prediction studies due to its effective handling of strongly nonlinear input variables and robustness against outliers (Wang et al., 2021).

Extremely randomized trees or Extra Trees, constitutes a model characterized by increased randomness and enhanced performance compared to Random Forests (Geurts et al., 2006). Mitigating the risk of overfitting is achieved by utilizing the entire dataset as training data

instead of employing the bootstrap method for decision tree generation (John et al., 2016). In contrast to Random Forests, which assess multiple features to determine the optimal node partitioning method, Extra Trees conducts partitioning by randomly selecting features and subsequently selecting the optimal node partitioning method, leading to faster computation speed. Leveraging these advantages, it has been utilized in numerous regression analyses.

3.2.2. Boosting algorithm

Gradient boosting is one of the most reputed boosting-based models and serves as the foundation for various ensemble model designs (Friedman, 2001). At each stage, the weak learner predicts the errors of the previous weak learner and minimizes errors through iteratively learning. The errors are quantified using a loss function, and the objective is to minimize the loss function through gradient descent method. The inclusion of gradient in the model name is due to the fact that the predicted errors during learning align with the gradient of the loss function. It uses a decision tree as the fundamental learner and is applicable to both regression and classification analyses.

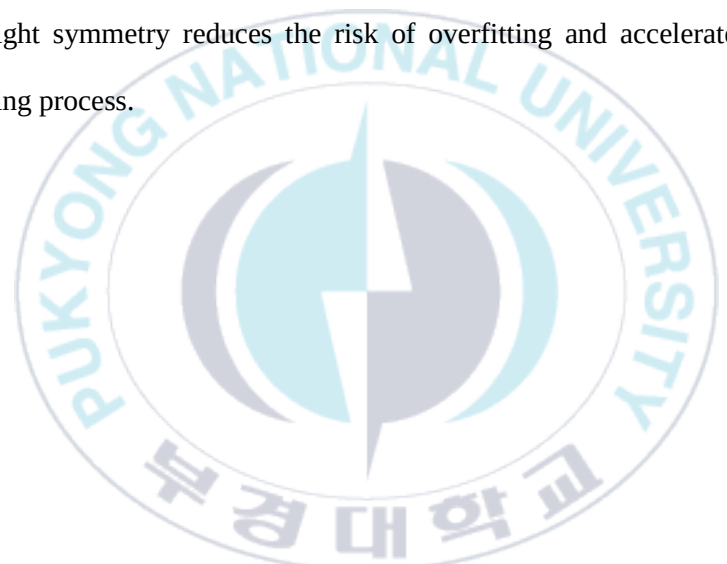
Extreme gradient boosting, or XGBoost, is a model based on gradient boosting and is a machine learning paradigm celebrated for its exceptional

performance across diverse fields (Chen and Guestrin, 2016). It has improved computational efficiency by reconfiguring the model architecture of conventional gradient boosting to parallel learning. Functioning as a classification and regression tree (CART)-based ensemble model, XGBoost demonstrates outstanding predictive prowess in both classification and regression tasks. By employing regularization, the model complexity is controlled, and L1 and L2 regularization is applied to prevent overfitting. Additionally, XGBoost performs second-order derivative computation to minimize errors and possesses intrinsic capabilities for handling missing values.

Light gradient boosting machine or LightGBM is one of the gradient boosting-based models developed by Microsoft (Ke et al., 2017). In addition to enhancing the learning speed, two algorithms—gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB)—were applied to reduce the volume of data and input variables utilized in constructing the model. Unlike the traditional level-wise tree splitting in the general GBM series, which balanced splits to expand horizontally, LightGBM adopts a leaf-wise approach. This asymmetrical deepening of the tree not only saves time but also conserves memory resources.

Categorical boosting or CatBoost is the most recent gradient boosting-based model (Dorogush et al., 208). It excels in handling categorical

variables and introduces an ordered boosting technique to prevent target leakage—a prevalent issue in existing gradient boosting. CatBoost addresses the target leakage problem by training the model in diverse ways through the creation of random permutations. In the construction of a new tree, CatBoost utilizes an oblivious decision tree (ODT), employing the same division criterion for trees at the same level. A balanced tree with left-right symmetry reduces the risk of overfitting and accelerates the learning process.



3.3. Model Accuracy Assessment

In this study, the performance of the models was assessed using the evaluation metrics r-squared score (R^2), root mean square errors (RMSE), and mean absolute errors (MAE). R^2 has a range of values between 0 and 1, indicates how well the model describes the data, with values closer to 1 signifying a stronger correlation between input and output variables. RMSE and MAE represent the difference between the actual and predicted values, so the smaller the value, the better the performance of the model. The formula of the evaluation indicator is as shown in equation (1-3), where $\hat{y}_i$ is the predicted value, y_i is the observed value, and $\bar{y}$ is the average value of the observations.

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

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \quad (3)$$

3.4. SHAP

SHAP, an eXplainable Artificial Intelligence (XAI) technique, is a method that applies the game theory concept of shapley value to interpret the output results of predictive models. Shapley value, based on cooperative game theory, signifies the value distributed to each participant based on their marginal contribution to the total gains achieved through cooperative play in a game. It illustrates how much and in what way each variable has contributed to representing the overall results. SHAP enables both global and local interpretations of input variables, offering a more accurate analysis than traditional methods of feature importance. Additionally, SHAP allows representation not only of feature importance but also the influence on individual predicted values, and provides a visual representation of Shapley values. Through SHAP analysis, selecting variables that play an important role in model predictions is one of the key factors in building a more accurate model.

In this study, Tree SHAP, specifically designed for tree-based models, was employed to analyze the model outcomes. Two SHAP analyses were conducted. First, we applied SHAP value analysis to the training model results to identify important variables and select the final input variable combination. Then, with the application of the selected input variables to

the model and the SHAP analysis technique was utilized to scrutinize the feature contributions to the estimation of chl-a concentration.

3.5. Model train and test

Utilizing the SHAP values from the results of the initial trained model, the final input variables were determined, including 13 water quality factors, 4 meteorological factors, 5 hydrological factors, and 6 satellite factors, as presented in the (Table 4). The train data (n=604) and test data (n=152) were randomly partitioned in an 8:2 ratio, and the same dataset was applied to the six models for performance comparison and analysis. Model optimization was conducted using the grid-search cross-validation (Grid-search CV) method. Grid-search CV is a technique that identifies the optimal hyperparameter combination by comparing prediction performance across various combinations of hyperparameter values specified by the user as a list.

Table 4. Selected input variables used to machine learning.

Category	Variable names
Water quality	Temp, pH, DO, Cell, BOD, COD, TP, TOC, Conductivity, NH_3-N, DTP, PO_4-P, SS
Meteorology	Temp_avg, Temp_high, Temp_low, Wind_speed
Hydrology	Water level, Pondage, Storage efficiency, Outflow
Satellite	B2, B1 / B3, B3 / B4, B5 / B4 B6 / B5, ((1/B4) – (1/B5)) * B6

4. Results and Discussion

4.1. Model performance

In this study, a total of 27 input variables, encompassing water quality, meteorological, hydrological, and satellite factors selected through the SHAP technique, were employed in six models: CatBoost, Extra Trees, Gradient Boosting, LightGBM, Random Forest, and XGBoost. These models were used to estimate chl-a concentrations at eight weirs along the Nakdong River. Table 5 compares the results of applying all variables and the variables selected through SHAP value analysis to the six models. Model performance comparison was performed on the test set (n = 152), independent of the training set (n = 604), using R^2, RMSE, and MAE as evaluation metrics. Upon analyzing the variation in the evaluation metrics, we observed that the selected variables enhanced the estimation performance of the six models, with XGBoost demonstrating the most marked improvement. This observation highlights the feasibility of using SHAP as an input variable selection method, as well as for model interpretation and input variable importance analysis (Yoon et al., 2021).

Table 5. Comparison of model performance with R^2, RMSE and MAE.

Method	Model	R^2	RMSE	MAE
All variables	**CatBoost**	0.841	5.971	4.360
	Extra Trees	0.797	6.752	5.099
	Gradient Boosting	0.825	6.262	4.555
	LightGBM	0.844	5.912	4.484
	Random Forest	0.790	6.871	5.108
	XGBoost	0.798	6.735	4.894
Selected variables	**CatBoost**	0.862	5.560	4.120
	Extra Trees	0.812	6.500	4.745
	Gradient Boosting	0.832	6.137	4.431
	LightGBM	0.857	5.662	4.153
	Random Forest	0.805	6.616	4.933
	XGBoost	0.835	6.088	4.494

In the comparison of model results using the selected variables, CatBoost outperformed the other algorithms across all evaluation metrics (R^2 = 0.862, RMSE = 5.560 mg/m^3, MAE = 4.120 mg/m^3) (Fig. 3a). LightGBM (R^2 = 0.857, RMSE = 5.662 mg/m^3, MAE = 4.153 mg/m^3) performed comparably to CatBoost (Fig. 3d). XGBoost (R^2 = 0.835, RMSE = 6.088 mg/m^3, MAE = 4.494 mg/m^3) and Gradient Boosting (R^2 = 0.832, RMSE = 6.137 mg/m^3, MAE = 4.431 mg/m^3) performed similarly to one another (Fig. 3c,f). In contrast, the bagging algorithms, Extra Trees (R^2 = 0.812, RMSE = 6.500 mg/m^3, MAE = 4.745 mg/m^3) and Random Forest (R^2 = 0.805, RMSE = 6.616 mg/m^3, MAE = 4.933 mg/m^3), performed relatively poorly compared to the boosting algorithms (Fig. 3b,e). All models displayed a tendency to underestimate high chl-a concentrations (> 40 mg/m^3). This underestimation would be particularly relevant during summer months when high chl-a concentrations are more frequent, coinciding with a higher probability of encountering high cloud cover that can impede satellite imagery availability. Consequently, collecting satellite data during these periods poses a challenge**.** To address this issue and enhance model performance, additional data collection efforts in other regions and with alternative satellite data are warranted to construct a more comprehensive dataset for high chl-a concentration scenarios.

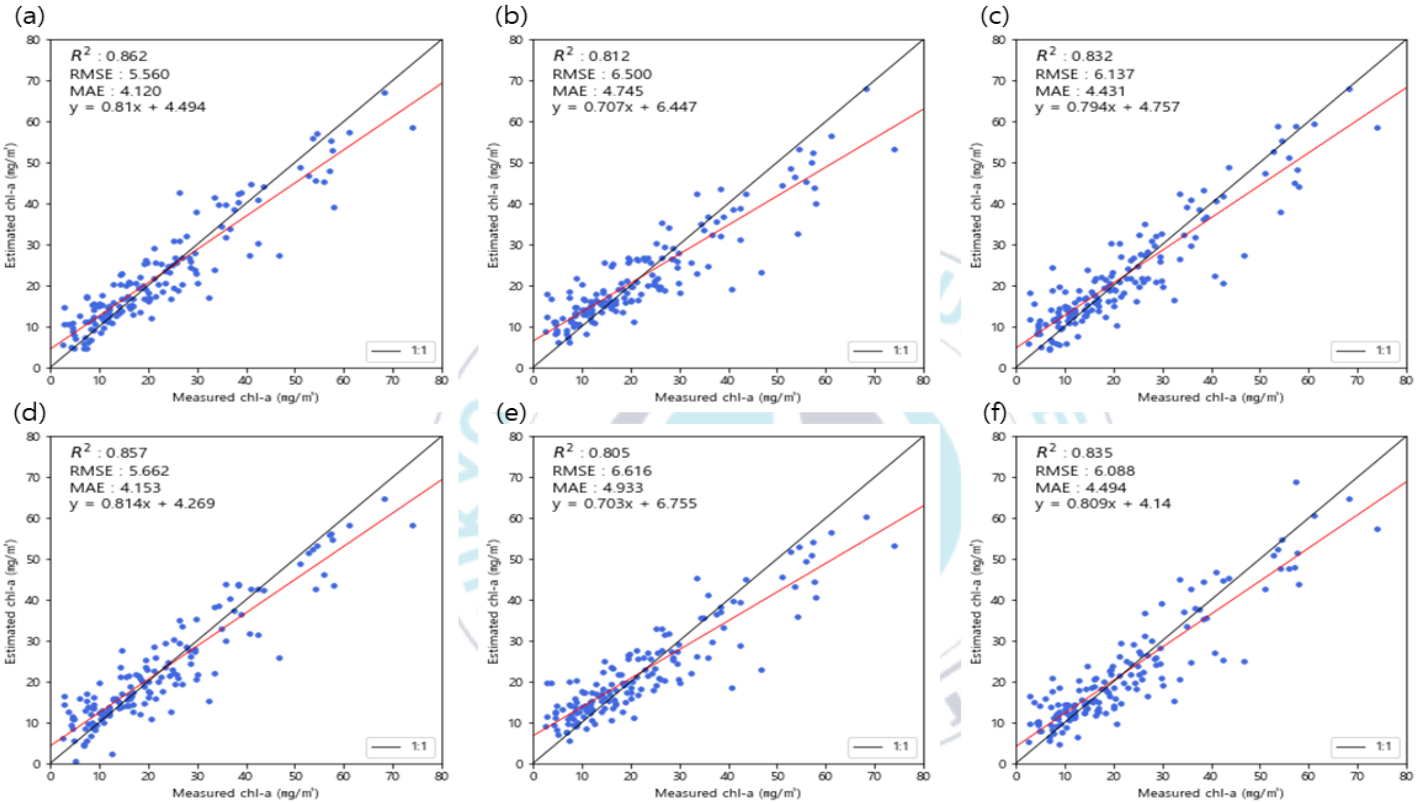


Fig. 3. The relationship between measured and estimated chl-a by machine learning algorithms (a) CatBoost (b) Extra Trees (c) Gradient Boosting (d) LightGBM (e) Random Forest (f) XGBoost. The black solid line represents the 1:1 line.

CatBoost and LightGBM both exhibited relatively high R^2 values above 0.85, along with similar RMSE and MAE values. Further comparative analysis is presented in Fig. 4. CatBoost exhibited superior prediction results for relatively low chl-a concentrations, while LightGBM demonstrated better prediction accuracy for relatively high chl-a concentrations. Notably, LightGBM tended to underestimate low chl-a concentrations. Based on this comprehensive comparison, CatBoost was selected as the final model for this study.

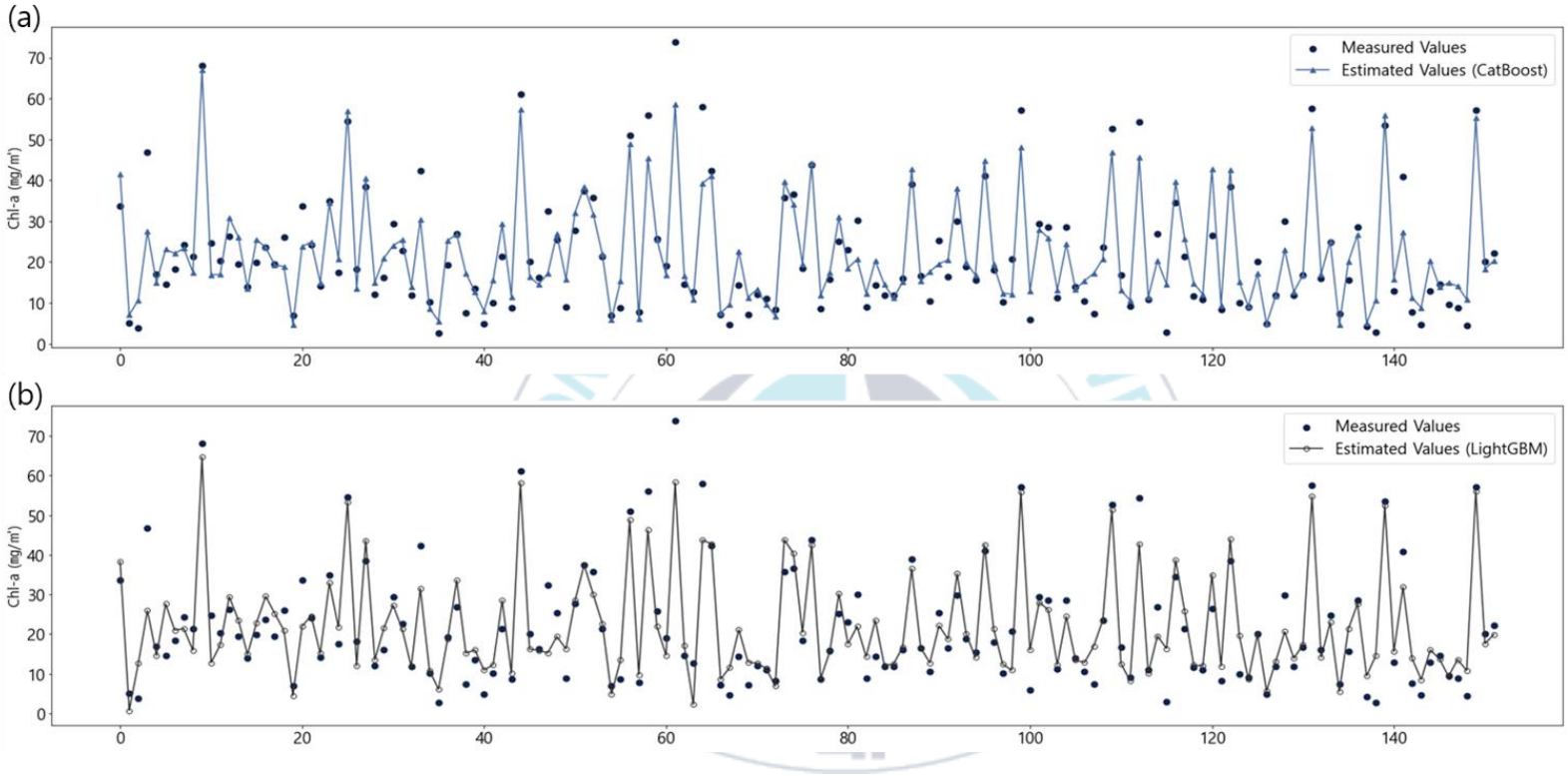


Fig. 4. Comparison between measured and estimated chl-a concentration values using (a) CatBoost and (b) LightGBM ; The black solid circle represents the measured value

Similar to our study, previous research has utilized a variety of models to estimate chl-a concentrations, employing different input data configurations. A study by Shi et al. (2022a) employed Sentinel-2 MSI data, incorporating six spectral bands and six band combinations. Notably, XGBoost (R^2 = 0.80, RMSE = 2.42 mg/m^3) demonstrated superior performance, outperforming Random Forest (R^2 = 0.79, RMSE = 2.51 mg/m^3). In another investigation, data from six spectral bands and four band combinations of Sentinel-2 MSI were employed. Among the five models assessed, LightGBM emerged as the top performer, achieving an R^2 of 0.75, RMSE of 15.15 mg/m^3, and MAE of 9.49 mg/m^3 (Kim et al., 2022). Nguyen et al. (2021), where the input data were divided into water quality data and Sentinel-2 data, GPR exhibited the best performance with water quality data, attaining an R^2 value of 0.85, while CatBoost excelled when using satellite data, boasting an R^2 value of 0.84. A study by Kim and Ahn (2022) utilized water quality and meteorology data to build Random Forest, SVR, and ANN models, with Random Forest (R^2 = 0.747, RMSE = 8.617 mg/m^3, MAE = 4.109 mg/m^3) exhibiting the best performance. Furthermore, in a study by Kim and Prak (2023), which utilized water quality data for building CatBoost, LightGBM, and XGBoost models, CatBoost demonstrated the highest performance,

despite the model performance evaluation metrics differing from those employed in our study.

In our study, the fusion of satellite data with water quality, meteorology, and hydrology data was undertaken to estimate chl-a concentrations at eight weirs along the Nakdong River. The application of six models yielded notable results, with all models achieving significant R^2 values exceeding 0.8. This suggests the successful development of a model applicable to monitoring chl-a concentrations in the Nakdong River. Notably, CatBoost emerged as the most reliable model, indicating the superior performance of boosting-based algorithms. Consistent with prior research, our findings align with the demonstrated efficiency of boosting-based algorithms in water quality prediction. Unlike earlier studies that relied on either satellite data or independently measured data (water quality, meteorology, and hydrology) to predict chl-a concentrations, our study exhibited comparatively better predictions using a combined dataset. This emphasizes the enhanced accuracy achieved through the fusion of satellite data with water quality, meteorological, and hydrological factor data. Furthermore, expanding data collection efforts to include the other major river basins (Han River, Geum River, and Yeongsan River), other than the Nakdong River, is expected to contribute to the development of a model applicable on a nationwide scale.

4.2. Model interpretation with SHAP

To assess the contribution of the selected 27 variables in model training for chl-a concentration estimation, the SHAP method was applied to conduct a feature importance analysis. Fig. 5 and 6 illustrate bar plots representing the impact of each variable on chl-a concentration prediction by calculating the average absolute value of the Shapley values. A longer bar indicates a higher contribution to chl-a concentration prediction, while a shorter bar suggests a lower impact. In all six models, SS emerged as the most influential variable. Although the ranking of variable importance varied by model, water quality variables such as DO, pH, TOC, and biochemical oxygen demand (BOD), and satellite variables B5/B4, ((1/B4)-(1/B5))*B6, were identified as common top-ranking variables. Across all models, the mean absolute SHAP value of SS exceeded 3, with the gradient boosting model registering the highest mean absolute SHAP value for SS at 4.034. In the case of CatBoost, the top-performing model in this study, the ranking of variable importance, from highest to lowest, was as follows: SS, DO, B5/B4, pH, TOC, ((1/B4)-(1/B5))*B6, and BOD.

The contribution of factors of low importance tended to vary by model, but it was observed that pondage and wind_speed contributed less to the estimation of chl-a concentration in most models. When averaging the

mean absolute SHAP value for each factor in each model to obtain a ranking, B3/B4 (mean value across six models = 0.244) was found to be the lowest among the 27 factors, followed by pondage (0.245), wind_speed (0.248), and Temp (0.270). Bagging-based algorithms, Extra Trees and Random Forest had very low SHAP values for low-ranking factors compared to boosting-based algorithms. The bottom five factors for each model are detailed in Table 6. Notably, the importance of meteorological and hydrological factors appeared relatively low compared to water quality factors. The relatively lower importance of meteorological factors is due to the varying distances between the point where the meteorology data were collected and the weir point for each data point.

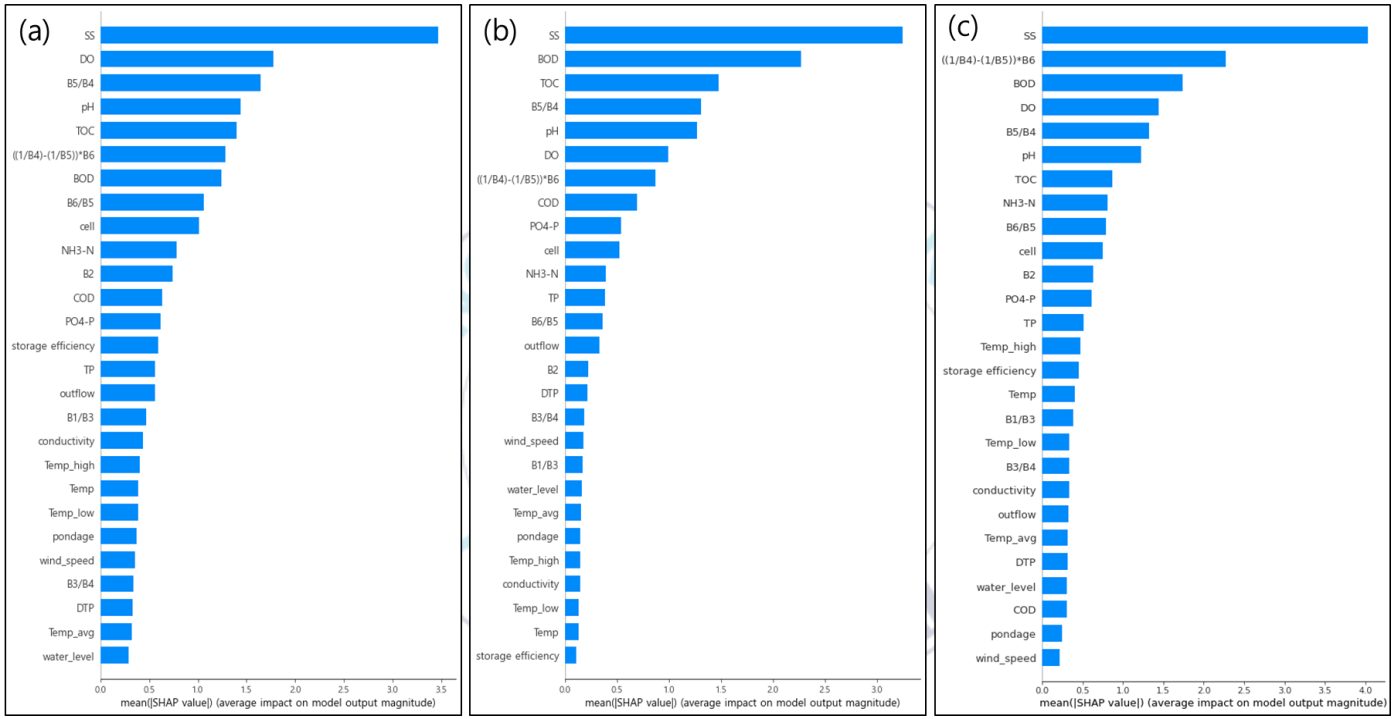


Fig. 5. SHAP summary bar plot for chl-a, estimating (a) CatBoost (b) Extra Trees (c) Gradient Boost

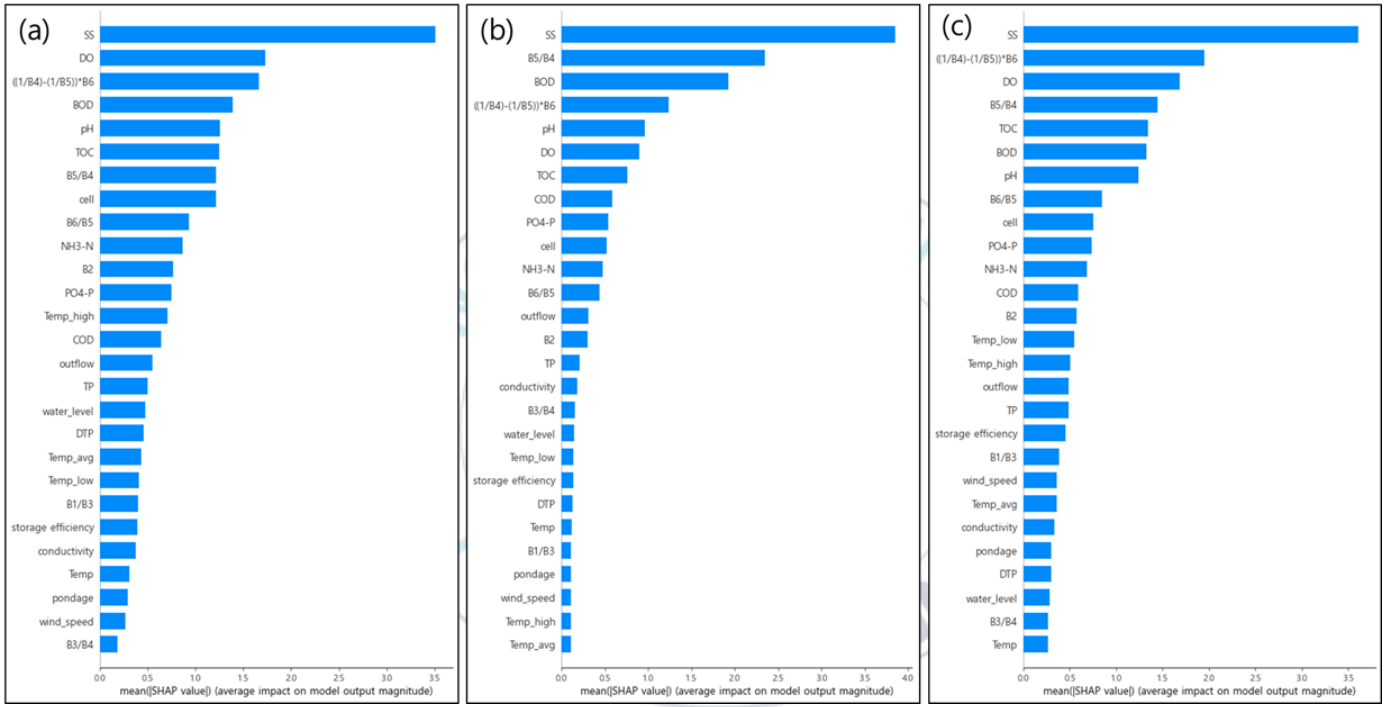


Fig. 6. SHAP summary bar plot for chl-a, estimating (a) LightGBM (b) Random Forest (c) XGBoost

Table 6. The bottom five ranked variables in terms of importance for each model.

Rank	CatBoost	Extra Trees	Gradient Boosting	LightGBM	Random Forest	XGBoost
1	wind_speed	Temp_high	DTP	conductivity	B1/B3	pondage
2	B3/B4	conductivity	water_level	Temp	pondage	DTP
3	DTP	Temp_low	COD	pondage	wind_speed	water_level
4	Temp_avg	Temp	pondage	wind_speed	Temp_high	B3/B4
5	water_level	storage efficiency	wind_speed	B3/B4	Temp_avg	Temp

Fig. 7 and 8 depict dot plots illustrating the impact of individual variables on model learning, showcasing both positive and negative influences. These plots factors of greater importance, following the same order as in the bar plot. Red dots represent features with large values, while blue dots represent features with small values. A SHAP value less than 0.0 indicates a decrease in the predicted value, whereas a value greater than 0.0 signifies an increase in the predicted value. Taking SS as an example, red dots are consistently distributed where the SHAP value is greater than 0.0 across all models, suggesting a positive correlation between SS values and the estimated chl-a concentration. Factors such as DO, BOD, pH, TOC, B5/B4, and ((1/B4)-(1/B5))*B6, identified as top factors in the six models, exhibited positive correlations, although the degree of influence varied by model. Conversely, B6/B5 and PO_4-P displayed a trend of high estimated chl-a concentration when the values were small in most models, indicating a negative correlation. In Fig. 7(b), factors at the bottom, such as conductivity, Temp_low, and Temp, present an ambiguous distribution of dots around the SHAP value of 0.0, posing a challenge for interpreting the results of the model for factors with such a distribution. Unlike conventional factor importance analysis, SHAP analysis offers the advantage of evaluating the impact of individual data on model predictions.

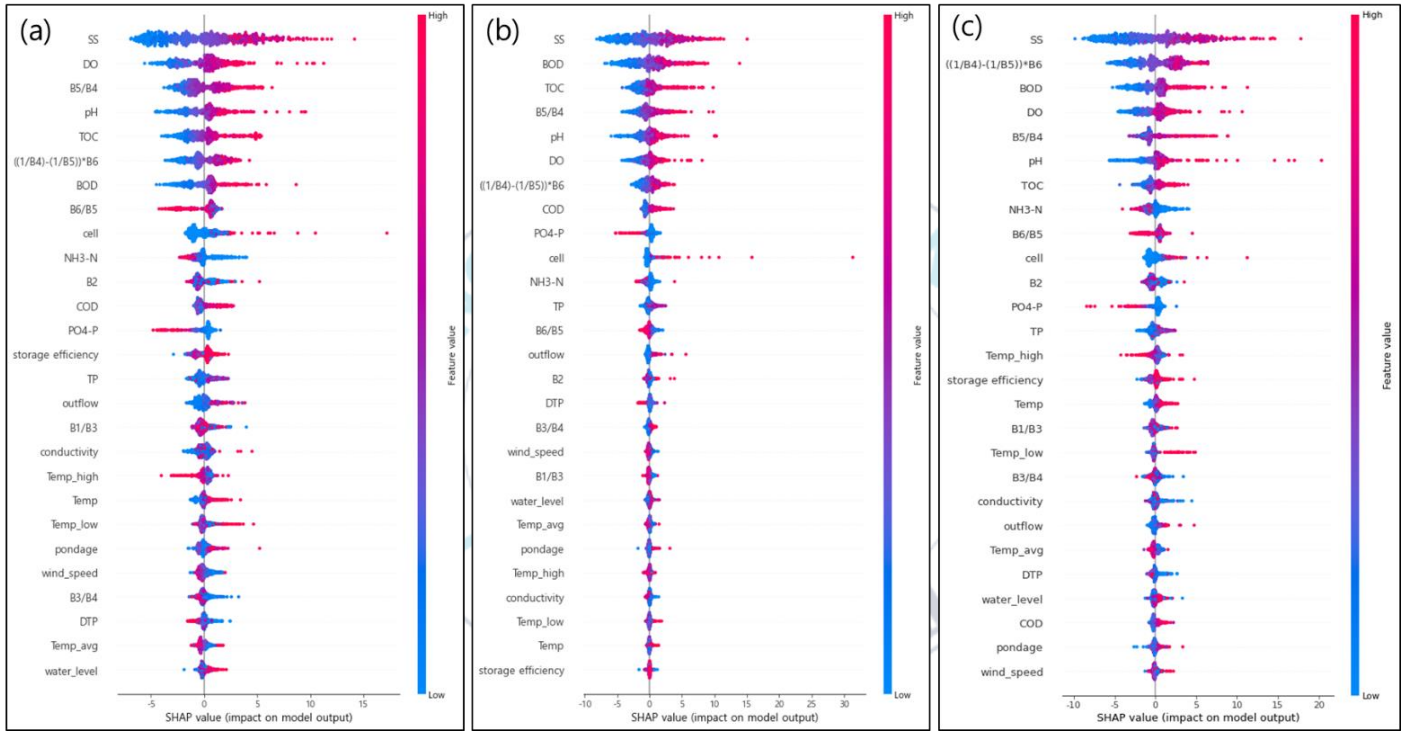


Fig. 7. SHAP summary dot plot for chl-a, estimating (a) CatBoost (b) Extra Trees (c) Gradient Boost

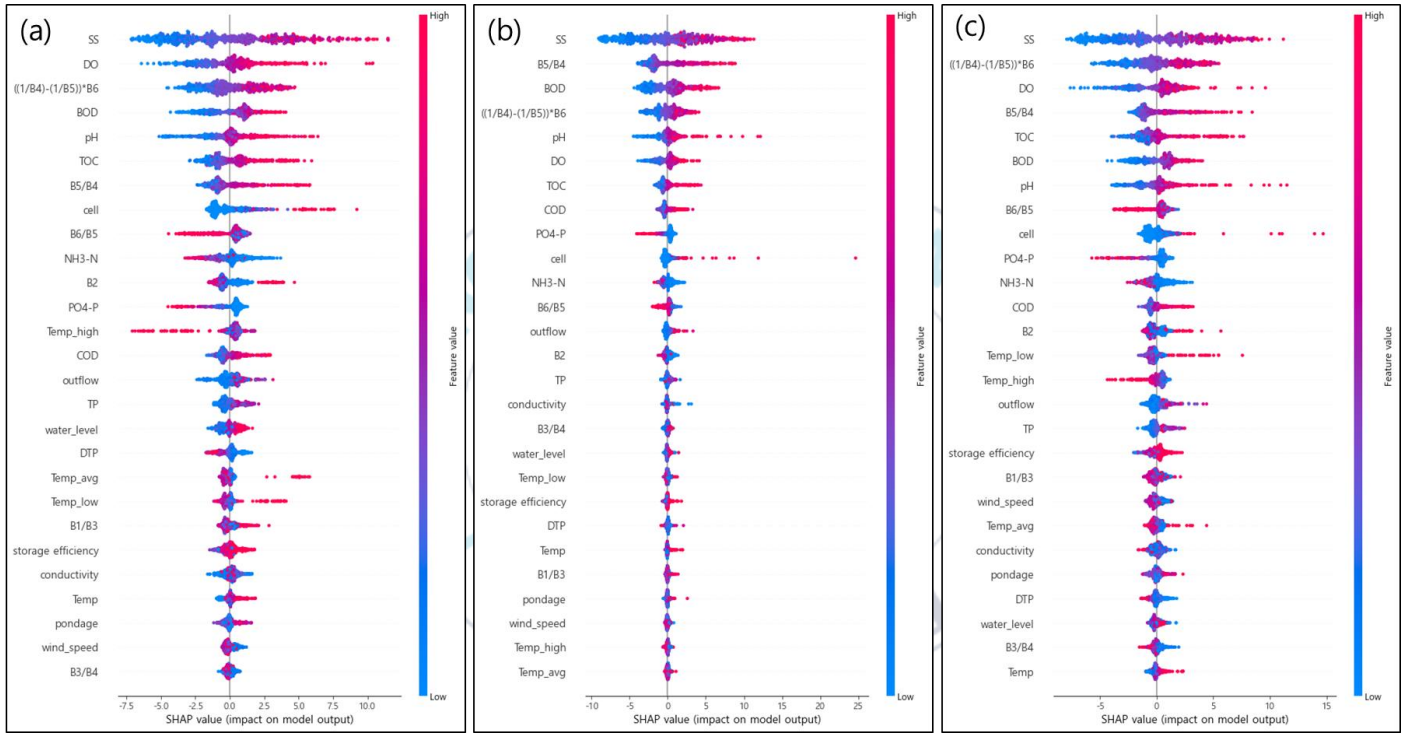


Fig. 8. SHAP summary dot plot for chl-a, estimating (a) LightGBM (b) Random Forest (c) XGBoost

The SHAP method aids in analyzing the influence of specific variables and provides localized interpretations for these variables. Fig. 9, a dependence plot, illustrates how the top three factors—SS, DO, and B5/B4—impact the prediction of chl-a concentration in CatBoost. As with the dot plot, red dots indicate large feature values, while blue dots represent small values. The SHAP value increased monotonically as the variable value increased, indicating a positive correlation between all three factors and chl-a. This plot not only reveals the variable interactions but also allows for correlation analysis by specifying particular variables. Fig. 10 presents an interaction plot analyzing the correlation between the top three factors of CatBoost and the variable with the most significant interaction. It appears that SS is most dependent on B5/B4, DO on TP, and B5/B4 on DO. Small TP values tend to coincide with large DO values, indicating a negative correlation (Fig. 10(b)). However, Fig. 10 (a and c) shows an unclear value distribution, making it challenging to analyze the correlation between the two variables in this case.

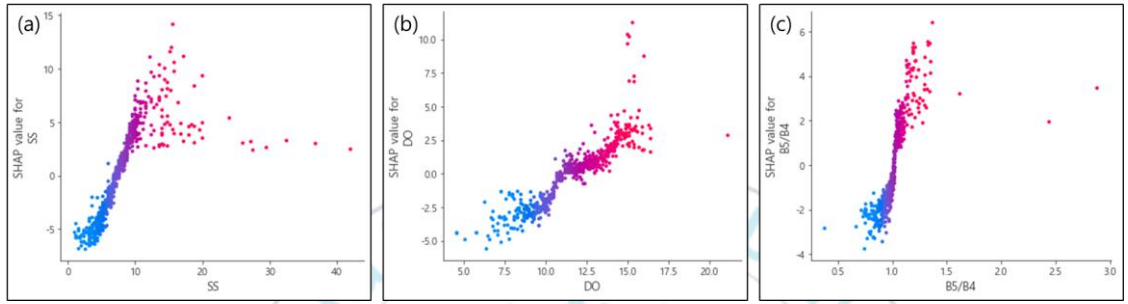


Fig. 9. SHAP dependence plots showcasing the impact of the three most crucial input variables : (a) SS (b) DO (c) B5/B4

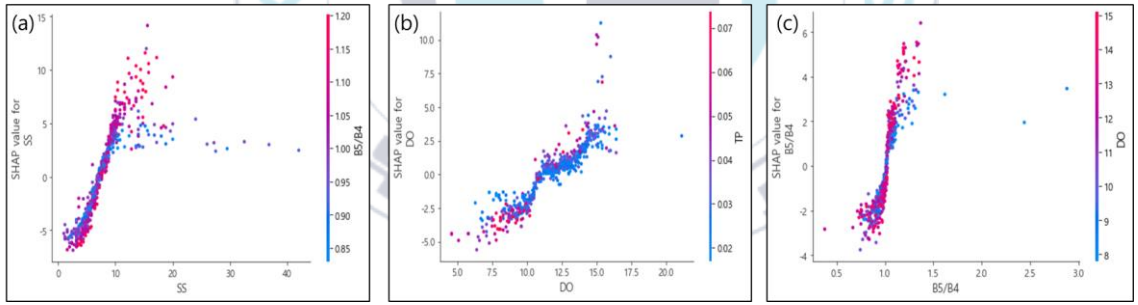


Fig. 10. Plots of SHAP interaction effects (a) SS (b) DO (c) B5/B4

Various studies have explored the comparison of features among different input variables, yielding diverse results in the analysis of variable importance. In a study assessing the LightGBM model's important factors for estimating chl-a concentration through SHAP analysis, Kim et al. (2022) identified B5/B4 (mean absolute SHAP value = 15.02), B6/B5 (3.64), and B2/B3 (3.01) as significant contributors to chl-a concentration prediction. The study utilized single spectral bands and band combinations as input data, highlighting the greater importance of band combinations than single spectral. Another study comparing the performance of band ratio algorithms found that the chl-a concentration estimation model, utilizing B5/B4 and (B5-B4)/(B5+B4), demonstrated high performance (Park et al., 2018). Nguyen et al. (2021) applied water quality factors and satellite data to machine learning, respectively. The study, water quality factors such as TN and TP and satellite data such as Sentinel-2 B3 and B6 data, emerged as crucial variables for estimating chl-a concentrations. According to Kim and Ahn (2020), the paramount factor influencing algal blooms was the water quality variable TOC, with an importance score of 0.27, followed by TN (0.19), pH (0.13), and water temperature (0.8). Meteorological factors related to temperature exhibited relatively low importance compared to water quality factors. Xia et al. (2020) found that hydrology variables, such as water level and flow rate, had a more

substantial impact on algae occurrence than water quality variables. A study by Jung et al. (2021), an analysis of factors influencing algal blooms at eight different weir sites along the Nakdong River was conducted using Random Forest. Although the ranking of important factors varied among the weirs, common important factors included outflow in the upper reaches, and DO and temperature in the middle and lower reaches. Furthermore, in studies conducted by Lee et al. (2020) and Lee and Kim et al. (2021), correlation analyses were conducted between water quality, hydrology factors, and chl-a concentrations at weir points located in the middle and lower reaches of the Nakdong River. Commonly identified factors in these analyses were BOD, SS, and DO.

The causes of algal blooms vary across different study areas due to diverse characteristics, including climatic and environmental factors. Furthermore, the significant factors influencing chl-a concentrations at the eight weirs along the Nakdong River differed based on the unique characteristics of each location. Consistent with prior research, DO, BOD, SS, and TOC were identified as factors of high importance in this study. Notably, pH, indicating the indirect impact of CO_2 supply by algae blooms, exhibited substantial contributions among water quality factors (Kim and Ahn, 2022). SS, recognized as the most crucial factor in the present study, serves as a key indicator for water quality monitoring alongside chl-a.

Recent studies employing satellite imagery have utilized SS concentrations for water quality assessment (Hafeez et al., 2019). The band ratio algorithm, specifically B5/B4 and ((1/B4)-(1/B5))*B6, leveraging the combination of red (with low reflectance) and the red edge band (with high reflectance), demonstrated significant contributions. The Sentinel-2 MSI, equipped with the red edge band, is deemed highly suitable for chl-a concentration estimation models (Li et al., 2021). Hydrology factors emerge as crucial variables, given the potential impact of sluice gate operations on algae occurrence with changing water levels (Xia et al., 2020). Temperature, a well-established factor linked to algal blooms in numerous studies, was of low importance in this study but is deemed indispensable for accurate chl-a concentration estimations.

4.3. Spatial distribution map of Chl-a

CatBoost model was used to estimate and map spatial distribution of chl-a concentrations in three weirs (Nak-dan, Gu-mi, and Dal-sung) in different seasons. Chl-a was estimated simultaneously in June and November 2019 for Nak-dan weir and Gu-mi weir in the upstream section and for Dal-sung weir in the middle section (Fig 11 and 12). The spatial distribution maps exhibit that chl-a concentrations are generally higher in the summer months. The distribution of chl-a showed spatial variability in different seasons. In the Nak-dan weir, relatively high chl-a was observed summer (modeled chl-a = 24.383 mg/m^3), and the low chl-a occurred in autumn (12.302 mg/m^3) (Fig. 11). In contrast, the Dal-sung weir showed similar trends in chl-a concentrations in summer (13.278 mg/m^3) and autumn (14.504 mg/m^3). The Nak-dan weir, the upstream of the Nakdong River, is temporarily eutrophicated during the summer months despite having fewer pollution sources (ME, 2022). The Dal-sung weir occurs frequently algal blooms due to the influence of industrial and domestic wastewater from the Kumho River (Lee et al., 2020). The number of data is not equivalent between weirs, suggesting that further analysis is essential. The small number of data makes it difficult to analysis specific spatial distributions.

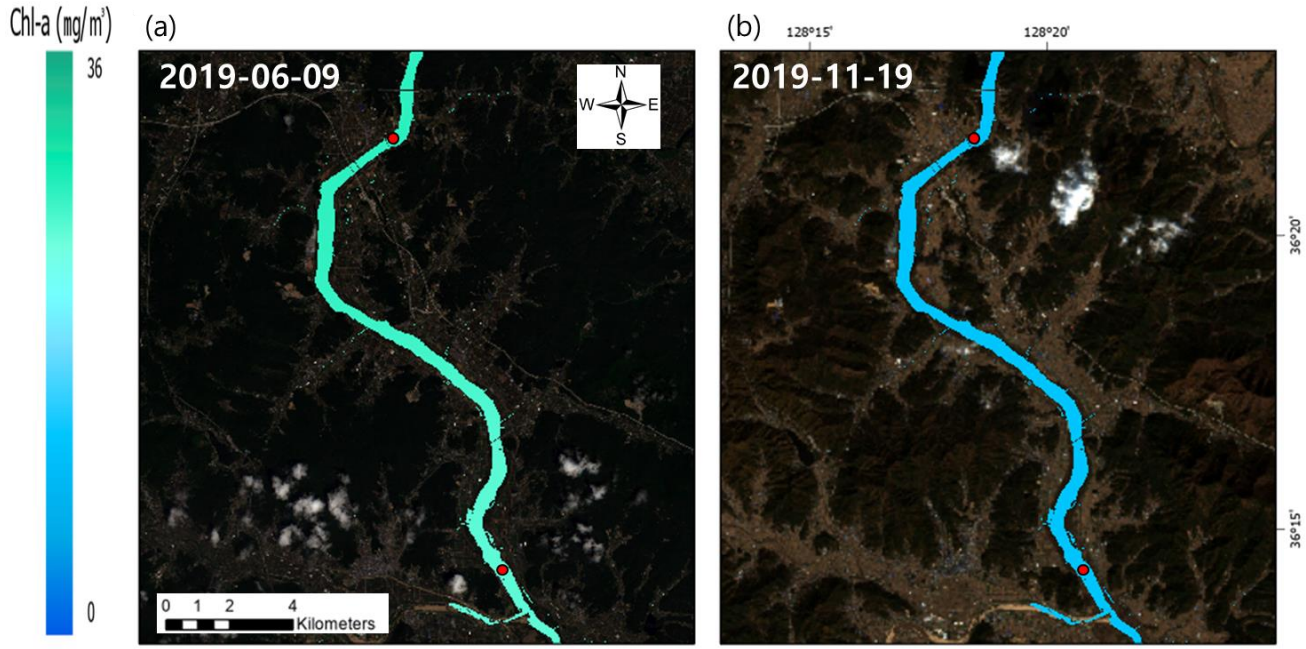


Fig. 11. Spatial distributions of modeled chl-a using CatBoost with Sentinel-2 true color composite images for Nak-dan weir and Gu-mi weir on (a) 09/06/2019 and (b) 19/11/2019.

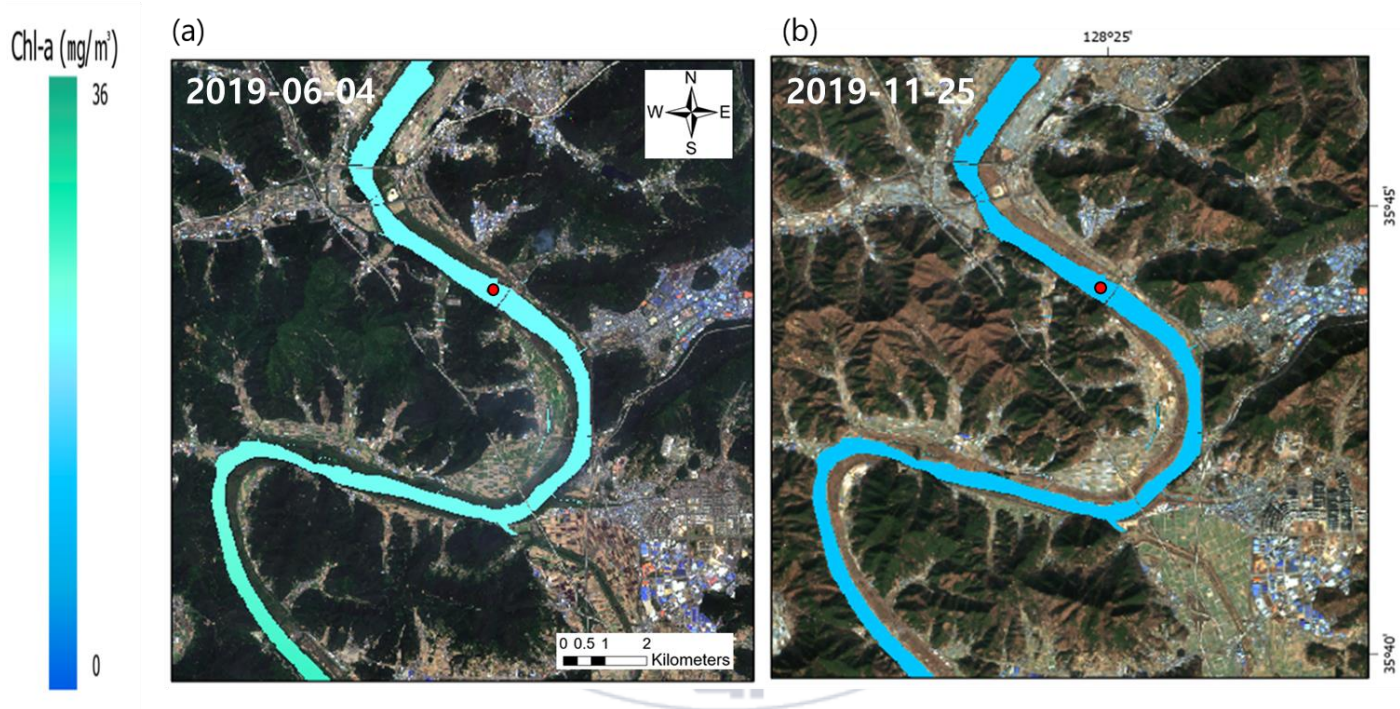


Fig. 12. Spatial distributions of modeled chl-a using CatBoost with Sentinel-2 true color composite images for Dal-sung weir on (a) 04/06/2019 and (b) 25/11/2019.

5. Conclusions

In this study, we fused Sentinel-2 MSI data and water quality, meteorological, and hydrological factors data to develop an accurate chl-a concentration estimation model for eight weirs along the Nakdong River. Diverse input data were employed to assess the fused dataset applicability. Twenty-seven input variables selected through the SHAP analysis, were applied to six models chosen via AutoML to compare and evaluate the results of chl-a concentration estimation. CatBoost outperformed the other algorithms (R^2 = 0.862, RMSE = 5.560 mg/m^3, MAE = 4.120 mg/m^3) and LightGBM (R^2 = 0.857, RMSE = 5.662 mg/m^3, MAE = 4.153 mg/m^3) showed comparable performance to CatBoost. XGBoost and Gradient Boosting exhibited similar performance, while the bagging-based ensemble models Random Forest and Extra Trees exhibited comparatively lower accuracy. SHAP method applied to analysis impact of input variables across six models. In all models, SS emerged as the most influential variable and exhibited positive correlations with chl-a concentrations. Although the ranking of variable importance varied by model, common top-ranking variables included water quality factors (DO, pH, TOC, and BOD) and satellite factors (B5/B4, ((1/B4)-(1/B5))*B6).

For satellite factors, highlighting the greater importance of band combinations than single spectral bands. Meteorological and hydrological factors exhibited relatively low importance compared to water quality factors. The spatial distribution map of chl-a generally provides realistic representations with seasonally distinct patterns. All six models obtained significant results with R^2 values above 0.8, suggesting that the fused data contributed to the accuracy of chl-a concentration estimations. Specifically, this approach should be extended to development the nationwide scale chl-a concentration estimation models through expand data collection efforts to include four major river basins. Further enhanced universality of the model can be realized through integration with nationally available datasets.

6. References

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Chavula, G., Brezonik, P., Thenkabail, P., Johnson, T., & Bauer, M. (2009). Estimating chlorophyll concentration in Lake Malawi from MODIS satellite imagery. Physics and Chemistry of the Earth, Parts A/B/C, 34(13-16), 755-760.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., ... & Ren, H. (2020). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water research, 171, 115454.

Cho, H. K., Lim, H. J., & Kim, S. M. (2018). Comparison of water quality

before and after four major river project for water monitoring stations located near 8 weirs in Nakdong River. Journal of Agriculture & Life Science, 52(6), 89-101.

Dai, Y., Yang, S., Zhao, D., Hu, C., Xu, W., Anderson, D. M., ... & Feng, L. (2023). Coastal phytoplankton blooms expand and intensify in the 21st century. Nature, 615(7951), 280-284.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.

Ferreira, L., Pilastri, A., Romano, F., & Cortez, P. (2022). Using supervised and one-class automated machine learning for predictive maintenance. Applied Soft Computing, 131, 109820.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63, 3-42.

Gitelson, A. A., Dall'Olmo, G., Moses, W., Rundquist, D. C., Barrow, T., Fisher, T. R., ... & Holz, J. (2008). A simple semi-analytical model for remote estimation of chlorophyll-a in turbid waters: Validation. Remote Sensing of Environment, 112(9), 3582-3593.

Gitelson, A. A., Gao, B. C., Li, R. R., Berdnikov, S., & Saprygin, V.

(2011). Estimation of chlorophyll-a concentration in productive turbid waters using a Hyperspectral Imager for the Coastal Ocean—the Azov Sea case study. Environmental Research Letters, 6(2), 024023.

Gurlin, D., Gitelson, A. A., & Moses, W. J. (2011). Remote estimation of chl-a concentration in turbid productive waters—Return to a simple two-band NIR-red model?. Remote Sensing of Environment, 115(12), 3479-3490.

Ha, N. T. T., Thao, N. T. P., Koike, K., & Nhuan, M. T. (2017). Selecting the best band ratio to estimate chlorophyll-a concentration in a tropical freshwater lake using sentinel 2A images from a case study of Lake Ba Be (Northern Vietnam). ISPRS International Journal of Geo-Information, 6(9), 290.

Hafeez, S., Wong, M. S., Ho, H. C., Nazeer, M., Nichol, J., Abbas, S., ... & Pun, L. (2019). Comparison of machine learning algorithms for retrieval of water quality indicators in case-II waters: A case study of Hong Kong. Remote sensing, 11(6), 617.

Jensen, J., 2006. Remote sensing of the environment: An earth resource perspective (2nd ed.). Pearson Education India

Jeon, B. S., Han, J., Kim, S. K., Ahn, J. H., Oh, H. C., & Park, H. D. (2015). An overview of problems cyanotoxins produced by cyanobacteria

and the solutions thereby. Journal of Korean Society of Environmental Engineers, 37(12), 657-667.

John, V., Liu, Z., Guo, C., Mita, S., & Kidono, K. (2016). Real-time lane estimation using deep features and extra trees regression. In Image and Video Technology: 7th Pacific-Rim Symposium, PSIVT 2015, Auckland, New Zealand, November 25-27, 2015, Revised Selected Papers 7 (pp. 721-733). Springer International Publishing.

Jung, W. S., Kim, S. E., & Kim, Y. D. (2021). Analysis of influential factors of cyanobacteria in the mainstream of Nakdong river using random forest. Journal of Wetlands Research, 23(1), 27-34.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Kim, J. & Park, J. (2023). Evaluation of Multi-classification Model Performance for Algal Bloom Prediction Using CatBoost. Journal of Korean Society on Water Environment, 39(1), 1-8.

Kim, K. M., & Ahn, J. H. (2022). Machine learning predictions of chlorophyll-a in the Han river basin, Korea. Journal of Environmental Management, 318, 115636.

Kim, S. H., Park, J. H., & Kim, B. (2021). Prediction of cyanobacteria harmful algal blooms in reservoir using machine learning and deep

learning. Journal of Korea Water Resources Association, 54(12), 1167-1181.

Kim, Y. W., Kim, T., Shin, J., Lee, D. S., Park, Y. S., Kim, Y., & Cha, Y. (2022). Validity evaluation of a machine-learning model for chlorophyll a retrieval using Sentinel-2 from inland and coastal waters. Ecological Indicators, 137, 108737.

Kosten, S., Huszar, V. L., Bécares, E., Costa, L. S., van Donk, E., Hansson, L. A., ... & Scheffer, M. (2012). Warmer climates boost cyanobacterial dominance in shallow lakes. Global Change Biology, 18(1), 118-126.

Lee, S. H., Kim, B. R., & Lee, H. W. (2014). A study on water quality after construction of the weirs in the middle area in Nakdong River. Journal of Korean Society of Environmental Engineers, 36(4), 258-264.

Lee, S. M., & Kim, I. K. (2021). A comparative study on the application of boosting algorithm for Chl-a Estimation in the downstream of Nakdong River. Journal of Korean Society of Environmental Engineers, 43(1), 66-78.

Lee, S. M., Park, K. D., & Kim, I. K. (2020). Comparison of machine learning algorithms for Chl-a prediction in the middle of Nakdong River (focusing on water quality and quantity factors). Journal of

Korean Society of Water and Wastewater, 34(4), 277-288.

Li, S., Song, K., Wang, S., Liu, G., Wen, Z., Shang, Y., ... & Mu, G. (2021). Quantification of chlorophyll-a in typical lakes across China using Sentinel-2 MSI imagery with machine learning algorithm. Science of the Total Environment, 778, 146271.

Lim, W., Go, W., Kim, K. Y., & Park, J. W. (2020). Variation in Harmful Algal Blooms in Korean coastal waters since 1970. Journal of the Korean Society of Marine Environment & Safety, 26(5), 523-530

Ministry of Environment (ME). (2022). Report on the algal bloom occurrence and countermeasures, Korea.

Moses, W. J., Gitelson, A. A., Berdnikov, S., & Povazhnyy, V. (2009). Estimation of chlorophyll-a concentration in case II waters using MODIS and MERIS data—successes and challenges. Environmental research letters, 4(4), 045005.

Musigmann, M., Akkurt, B. H., Krähling, H., Nacul, N. G., Remonda, L., Sartoretti, T., ... & Mannil, M. (2022). Testing the applicability and performance of Auto ML for potential applications in diagnostic neuroradiology. Scientific reports, 12(1), 13648.

Nguyen, H. Q., Ha, N. T., Nguyen-Ngoc, L., & Pham, T. L. (2021). Comparing the performance of machine learning algorithms for remote and in situ estimations of chlorophyll-a content: A case study

in the Tri An Reservoir, Vietnam. Water Environment Research, 93(12), 2941-2957.

O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., ... & McClain, C. (1998). Ocean color chlorophyll algorithms for SeaWiFS. Journal of Geophysical Research: Oceans, 103(C11), 24937-24953.

Park, S., Lee, S., Yun, Y., Shin, D., Park, S., & Lee, Y., 2018. Estimation of chlorophyll-a concentration for inland water using red-edge band of Sentinel-2 and RapidEye. The Geographical Journal of Korea, 52(3), 445–454.

Pretty, J. N., Mason, C. F., Nedwell, D. B., Hine, R. E., Leaf, S., & Dils, R. (2003). Environmental costs of freshwater eutrophication in England and Wales.

Rodríguez-López, L., Duran-Llacer, I., Gonzalez-Rodriguez, L., Abarca-del-Rio, R., Cárdenas, R., Parra, O., ... & Urrutia, R. (2020). Spectral analysis using LANDSAT images to monitor the chlorophyll-a concentration in Lake Laja in Chile. Ecological Informatics, 60, 101183.

Shi, X., Gu, L., Jiang, T., Zheng, X., Dong, W., & Tao, Z. (2022a). Retrieval of chlorophyll-a concentrations using Sentinel-2 MSI imagery in Lake Chagan based on assessments with machine learning

models. Remote Sensing, 14(19), 4924.

Shi, J., Shen, Q., Yao, Y., Li, J., Chen, F., Wang, R., ... & Zhou, Y. (2022b). Estimation of chlorophyll-a concentrations in small water bodies: comparison of fused Gaofen-6 and Sentinel-2 sensors. Remote Sensing, 14(1), 229.

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

Wang, R., Kim, J. H., & Li, M. H. (2021). Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. Science of the Total Environment, 761, 144057.

Waring, J., Lindvall, C., & Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial intelligence in medicine, 104, 101822.

Xia, R., Wang, G., Zhang, Y., Yang, P., Yang, Z., Ding, S., ... & Qian, C. (2020). River algal blooms are well predicted by antecedent environmental conditions. Water research, 185, 116221.

Yoon, Y. A., Lee, S. H., & Kim, Y. S. (2021). A Study on the Remaining Useful Life Prediction Performance Variation based on Identification and Selection by using SHAP. Journal of the Society of Korea

Industrial and Systems Engineering, 44(4), 1-11.

Zhang, F., Li, J., Shen, Q., Zhang, B., Tian, L., Ye, H., ... & Lu, Z. (2019). A soft-classification-based chlorophyll-a estimation method using MERIS data in the highly turbid and eutrophic Taihu Lake. International Journal of Applied Earth Observation and Geoinformation, 74, 138-149.

Zhou, Y., He, B., Fu, C., Giardino, C., Bresciani, M., Liu, H., ... & Liang, S. (2021). Assessments of trophic state in lakes and reservoirs of Wuhan using Sentinel-2 satellite data. European Journal of Remote Sensing, 54(1), 461-475.