Journal of Wetlands Research
Vol. 27, No.3, August 2025, pp.290-297

ISSN 1229-6031 (Print) / ISSN 2384-0056 (Online)
DOI https://doi.org/10.17663/JWR.2025.27.3.290

자동 머신러닝 AutoGluon 알고리즘을 활용한 하천 조류 발생 예측

김형민* · 박정수**†

*국립한밭대학교 환경공학과
**국립한밭대학교 건설환경공학과

Prediction of Algal Bloom in a River using AutoGluon, an Automated Machine Learning Algorithm

Hyungmin Kim* · Jungsu Park**†

*Department of Environmental Engineering, Hanbat National University, Korea
**Department of Civil and Environmental Engineering, Hanbat National University, Korea

(Received : 14 May 2025, Revised : 14 July 2025, Accepted : 18 August 2025)

요 약

하천의 과다한 조류 발생은 취수원 및 수생태 환경에 좋지 않은 영향을 줄 수 있으며 이에 대한 지속적인 관리가 중요하다. 본 연구에서는 낙동강 유입 지류 하천에서 조류 발생의 정량적 지표인 chlorophyll-a 농도를 예측하는 자동 머신러닝 모형 (AutoML: automated machine learning)을 구축하였다. AutoML은 머신러닝 모형의 구축을 위한 데이터 전처리, 모형 선정 및 최적화 과정의 편의성을 높여 상대적으로 용이한 모형 구축을 가능하게 하는 장점이 있으며, 본 연구에서는 AutoGluon을 이용하여 AutoML을 구축하고 기존에 널리 사용되는 머신러닝 모형인 random forest 및 XGBoost모형과 그 성능을 비교하였다. 모형 성능의 비교는 모형 성능의 평가에 활용되는 정량 지표인 Nash-Sutcliffe coefficient of efficiency (NSE), root mean squared error (RMSE) 및 RMSE-standard deviation ratio (RSR)를 활용하였다. 분석결과 AutoGloun, RF, XGB 세가지 모형의 RSR 값이 각각 0.564, 0.752, 0.811로 AutoGluon이 가장 우수한 성능을 보이는 것으로 확인되었다. 또한 AutoGluon 모형의 구축에 사용된 각 변수가 모형의 성능에 미치는 상대적 중요도인 feature importance를 확인하여 중요도가 낮은 변수부터 순차적으로 변수를 제거하면서 성능의 변화를 비교하였다. 분석결과 RSR이 0.542-0.579의 범위를 보여 입력변수가 제한적인 경우에도 일정 수준 이상의 안정적인 성능이 확보될 수 있음을 확인하였다.

핵심용어 : 녹조 발생, AutoGluon, 자동 머신러닝, 머신러닝, 수질 관리

Abstract

Excessive algal blooms in rivers can have negative impacts on water resources and aquatic ecosystems. Therefore, continuous management of algal bloom is essential. In this study, an automated machine learning (AutoML) model was developed to predict chlorophyll-a concentrations. AutoML has the advantage of simplifying the machine learning model development process by streamlining data preprocessing, model selection, and optimization, thus making model construction relatively easier. In this study, AutoGluon was used to implement AutoML, and its performance was compared with that of widely used machine learning models (i.e. Random Forest and XGBoost). Model performance was evaluated using three quantitative metrics: Nash – Sutcliffe Efficiency (NSE), Root Mean Squared Error (RMSE), and the Root Mean Squared Error – Observation Standard Deviation Ratio (RSR). The analysis showed that the RSR values for AutoGluon, RF, and XGB models were 0.564, 0.752, and 0.811, respectively, indicating that AutoGluon demonstrated the best performance. Additionally, the relative importance of the input features used in the development of the AutoGluon model was explored. Features were sequentially removed based on their importance ranking to assess the impact on model performance. The results showed that the RSR ranged from 0.542 to 0.579, verifying that the model maintained a stable performance even when input variables were limited.

Key words : Algal bloom, AutoGloun, Automated machine learning, Machine learning, Water quality management

†To whom correspondence should be addressed.
Department of Civil and Environmental Engineering, Hanbat National University
E-mail : parkjs@hanbat.ac.kr
• **Hyungmin Kim** Hanbat National University, Korea/Master Student(kimhm00531@gmail.com)
• **Jungsu Park** Hanbat National University, Korea/Associate professor(parkjs@hanbat.ac.kr)

1. 서 론

하천과 호수 등 담수 환경은 기후 변화와 인간 활동 등에 의해 다양한 영향을 받으며, 인간 활동에 의한 오염원 유입 증가에 따른 조류 발생의 심화는 하천 수질 안전성과 수생태 환경 등에 악영향을 줄 수 있어 지속적인 관리가 필요하다 (Gonzáleza et al., 2020). Chlorophyll-*a* (Chl-*a*) 농도는 조류 발생량을 정량적으로 평가할 수 있는 대표적인 지표로, 하천 및 댐 저수지 등에 발생되는 조류 관리를 위해 Chl-*a* 농도를 예측하기 위한 연구가 지속되고 있다.

조류 발생 및 수질 변화 등에 대한 예측은 현장 수질관리 및 관련 정책 수립 등을 위해 중요하며, 이를 위해 미국 환경청이 개발한 1차원 모형인 QUAL2E 및 3차원 수치 모형인 environmental fluid dynamic code (EFDC) 등 물질의 물리·화학·생물학적 기작 관계에 기반한 다양한 모형이 개발되어 현재까지도 활용되고 있다 (Kim and Son 2013; Shin et al., 2017).

머신러닝 모형은 기존의 기작 기반 모형과 달리, 예측 대상의 물리·화학·생물학적 인과관계를 명시적으로 수식화하거나 실험 등을 통해 필요한 계수를 산정하는 등의 과정이 필요 없이, 측정된 데이터로부터 입력 변수 간의 관계를 직접적으로 학습한다. 이러한 방식을 통해 구축된 머신러닝 모형은 녹조 발생과 같이 다양한 환경 변수가 복합적으로 작용하는 비선형 관계에 대해서도 우수한 예측 성능을 보이는 장점을 가지고 있어, 최근 국내외적으로 조류 발생 예측 등 현장 조류 관리를 위해 머신러닝 모형을 적용하려는 연구가 활발히 이루어지고 있다 (Lee et al., 2024; Park et al., 2024).

인공신경망 모형 등 비교적 초기에 개발된 모형부터, 다수의 단일 모형의 결과의 앙상블을 통해 모형의 성능을 개선하는 random forest (RF)와 XGBoost (XGB) 등의 앙상블 모형, 그리고 long short-term memory와 같은 딥러닝 알고리즘 등 다양한 머신러닝 알고리즘이 하천 Chl-*a* 농도 예측을 위한 모형 구축에 활용되고 있다 (Karimian et al., 2023; Park et al., 2024; Shin et al., 2020).

하지만 머신러닝 모형의 경우 데이터 전처리, 적합한 모형 선정 및 구축된 모형의 최적화 등 모형의 구축을 위한 여러 단계의 사전 작업이 필요하며 이를 위해 관련 분야의 전문적인 지식이 필요하다. 이러한 모형 구축 과정의 어려움은 모형의 실무 적용을 어렵게 하는 이유 중 하나이기도 하다. 자동 머신러닝 (AutoML: Automated Machine Learning)은 머신러닝 구축에 필요한 모형의 선정, 데이터 전처리 및 최적화 등의 과정을 자동화하여 분석자의 간섭을 최소화할 수 있어, 상대적으로 모형의 구축이 용이하면서도 안정적인 모형 성능을 확보할 수 있는 장점을 가지고 있다. 이러한 장점으로 AutoML은 복잡한 머신러닝 모형을 보다 쉽고 효율적으로 구축할 수 있는 도구로 관심을 받고 있으며, 다양한 AutoML의 개발 및 적용을 위한 관심이 계속되고 있다 (Karmaker et al., 2021; LeDell and Poirier, 2020; Salehin et al., 2024).

수질 분야에서도 오픈 소스 AutoML library인 Auto H2O를 활용하여 하천 Chl-*a*를 예측하는 모형을 구축하고 입력자료의 다양한 측정빈도가 모형의 성능에 미치는 영향을 비교하는 등 (Park et al., 2023), AutoML을 수질 및 조류 발생 예측 등에 활용하여 현장 수질관리 효율을 높이기 위한 기술개발 노력이 지속되고 있다 (Madni et al., 2023; Park 2024; Prasad et al., 2022).

데이터 기반 모형인 머신러닝 모형의 특성상 최신의 복잡한 모형이 우수한 항상 성능을 보이는 것이 아니므로, 모형의 구축에 사용된 데이터의 특성에 맞는 적정 모형의 선정이 필요하며 이러한 모형의 선정 및 평가에는 많은 인력과 시간이 소요된다. AutoGluon은 인공신경망 및 tree 기반 앙상블 모형 등 다양한 머신러닝 알고리즘을 내부에 포함하고 있으며, 이들 모형의 결과를 종합적으로 평가하여 최적의 성능을 내는 조합을 자동으로 탐색하는 앙상블 방식을 적용한다. 이를 통해 데이터에 적합한 최적 모형을 손쉽게 선정할 수 있으며, 안정적인 예측 성능을 확보할 수 있는 장점이 있다. 또한 시계열 예측, 구조화된 표 형식 (tabular) 데이터뿐만 아니라 이미지, 텍스트 등 다양한 형태의 데이터 분석이 가능하여, 현재까지도 널리 활용되는 대표적인 AutoML 프레임워크 중 하나이다 (Gao et al., 2024; Erickson et al., 2020, 2022).

본 연구에서는 적정 모형의 선정 및 데이터 전처리 등 모형 구축 과정의 자동화로 머신러닝 모형 구축의 효율성을 높인 AutoGluon을 사용하여 안동댐 하류 하천의 Chl-*a* 농도를 예측하는 머신러닝 모형을 구축하였으며, 기존에 널리 활용되는 머신러닝 알고리즘인 RF 및 XGB 모형을 활용하여 구축된 모형과 그 성능을 비교하여 AutoGluon의 적용성을 평가하였다. 모형의 구축을 위해 낙동강 안동댐 하류 하천에서 2014년부터 2023년까지 측정된 현장 수질 자료와 인근 기상 및 댐수문 자료를 활용하였으며, 모형의 입력 변수로 사용된 다양한 환경 인자가 AutoGluon 모형에 미치는 영향을 확인하기 위해, 모형 구축에 활용된 환경 변수의 상대적 중요도를 기반으로 선별적으로 입력 변수를 적용하여 모형의 성능에 미치는 영향을 확인하였다.

2. 재료 및 실험방법

2.1 입력 자료

안동댐은 낙동강 유역에 위치한 유역면적 1,584 ㎢, 저수용량 12억4800만 ㎥인 다목적댐으로 안동댐에서 방류되는 물은 안동댐 하류를 거쳐 낙동강으로 유입되어 대구, 부산, 창원 등 유역내 다양한 지역의 생활, 공업, 농업용수 및 하천유지용수로 활용되어 지속적인 수질관리가 필요한 지점이다 (Mywater, 2025, Noh et al., 2014).

본 연구에서는 낙동강유입 지류인 안동댐 하류 지역의 조류 발생을 예측하는 머신러닝 모형을 구축하였으며, 모형의 구축을 위해 하천 및 대상 유역의 수질, 기상 및 댐방류량 자료를 활용하였다 (Fig 1). 수질 자료는 환경부 국립환경과학원 물환경정보시스템 자동측정망 안동댐하류 지점 (site no. : S02023)에서 2014년 1월 1일 – 2023년 12월 31일까지

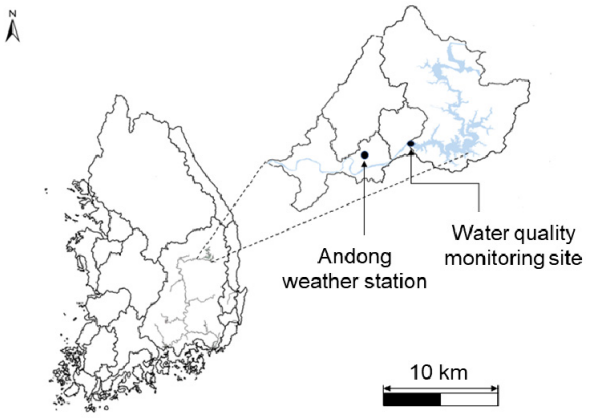


Fig. 1. Research site.

측정되어 공개된 일별 자료를 사용하였다 (Water Environment Information System [WEIS], 2025). 댐 방류량 자료는 국가수자원관리종합정보시스템 (Water Resources Management Information System [WAMIS], 2025)에 공개된 일별 안동댐 댐수문 자료를 사용하였다. 또한 대상 유역의 강수량, 일조량 등 기상 상황의 영향을 분석하기 위해 기상청 기상자료 개방포털에 공개된 안동기상대 (site no.: 136) 측정자료를 사용하였다 (Korea Meteorological Administration [KMA], 2025).

모형의 구축을 위해 조류 발생의 대표적인 정량지표 중 하나인 Chl-*a* (CHLA)를 예측의 대상이 되는 종속변수로 설정하고 조류 발생 등과 관련된 현장의 수질 및 기상 특성을 반영할 수 있는 환경 변수인 수온 (TEMP), 수소 이온농도 (pH), 전기전도도 (EC), 용존산소량 (DO), 총유기탄소 (TOC), 총질소 (TN), 총인 (TP), 댐 방류량 (Q), 평균기온 (AVER_TEMP), 일조량 (SUN), 일조시간 (SUN_H) 및 강수량 (RAIN)을 CHLA의 예측을 위한 독립변수로 활용하였다.

2.2 모형 구축

AutoML 모형의 구축을 위해 python open source library AutoGluon을 이용하였다 (Erickson et al., 2020). AutoGluon은 내부 알고리즘에 의해 선형회귀 모형, 신경망 (neural network) 모형 및 앙상블 모형 (LightGBM, CatBoost, RF 등) 다양한 모형을 구축하고 이러한 각 모형 결과의 앙상블을 통해 최종 결론을 도출하는 모형으로 데이터의 전처리, 다양한 개별 모형의 학습, 학습된 결과의 앙상블을 통한 최적의 모형 결과 도출까지 머신러닝 모형의 구축 전과정에서 외부의 개입을 최소화 하면서 모형 구축의 편의성을 높일 수 있도록 구성되어 있다 (Fig. 2).

AutoGluon은 모형의 최적화에 있어 모형의 기본 설정값을 우선적으로 적용할 것을 권장하고 있으며, 본 연구에서도 기본 설정값을 적용하여 모형의 학습을 수행하였다 (AutoGluon). 또한, 변수 선택에 따른 모형의 성능을 비교하기 위해 모든 입력 변수를 사용하여 구축한 모형 (M1)과 변수 중요도에 따라 선별적으로 변수를 선택하여 추가로 구성된 모형 (M2~M6)의 성능을 비교하였다. 이와 함께 AutoGluon의 적용성에 대한 평가를 위해, 널리 활용되는 대표적인 앙상블 머신러닝 모형인 RF와 XGB를 이용하여 모형을 구축하고 그 성능을 비교하였다.

RF는 weak learner로 불리는 의사결정나무를 활용하여 생성된 개별 모형 결과의 앙상블을 통해 최종 결과를 도출하는 알고리즘이며, XGB는 전 단계 weak learner의 결과를 다음 단계 개별 모형의 구축에 활용하여 점진적으로 모형의 성능을 향상시키는 gradient boosting decision tree 방식에 기반하여 구축되는 대표적인 앙상블 머신러닝 모형이다 (Chen and Guestrin, 2016; Friedman, 2001).

RF 및 XGB 모형의 최적화는 grid search 방식을 이용하였으며, time series cross-validation (n_split=6)을 적용하였다. 최적화를 위한 hyperparameter의 범위와 최적값은 Table 1에 제시하였다.

모형의 구축과 최적화 및 시각화 등을 위한 프로그램은 XGBoost 및 scikit-learn, NumPy, Pandas, matplotlib 등 Python 오픈소스 라이브러리를 활용하였다 (Chen and Guestrin, 2016; Harris, C. R. et al., 2020; McKinney, W., 2010; Pedregosa et al., 2011; XGBoost).

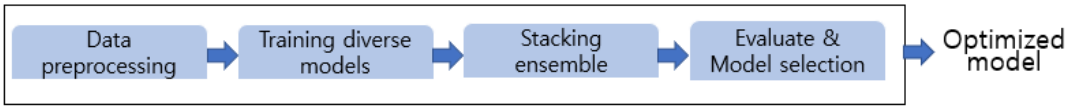


Fig. 2. A schematic diagram of AutoGluon.

Table 1. Hyperparameters used for model optimization

Model	Hyperparameter	Range	Optimal hyperparameter
RF	n_estimators	100, 200, 300, 400, 500	100
	min_sample_leaf	1, 2, 4	1
	min_sample_split	2, 5, 10	2
	max_depth	2, 3, 4, 5, 6	5
XGB	n_estimators	100, 200, 300, 400, 500	100
	learning_rate	0.01, 0.1, 2	0.01
	max_depth	2, 3, 4, 5, 6	6

모형 구축에 사용된 측정자료는 총 5% 정도의 결측치를 포함하고 있었으며, 모형의 구축을 위해 결측치가 발견된 지점에서 가까운 데이터 K개를 사용하여 결측값을 보간하는 방식인 K-Nearest Neighbors (KNN) 알고리즘을 사용하여 결측치 보간을 진행하였다. KNN을 이용한 결측값의 보정은 Python 오픈소스 라이브러리인 scikit-learn을 이용하였다 (Pedregosa et al., 2011).

우리나라는 사계절이 뚜렷하여 계절에 따른 기후 변화가 크며, 이러한 기후 특성으로 인해 조류 발생도 1년을 주기로 증감을 반복하는 경향을 보인다. 이러한 특성을 고려하여 모형의 학습 (training)과 평가 (testing) 자료를 연도를 기준으로 구분하였다. 전체 측정자료 중 2014년 1월 1일 – 2019년 12월 31일까지의 데이터를 모형의 학습 (training)에, 2020년 1월 1일 – 2023년 12월 31일까지의 데이터를 학습된 모형 성능의 평가 (testing)에 사용하여, 모형 구축에 사용된 training 및 testing 데이터의 구성 비율을 6:4로 설정하였다.

2.3 입력 변수 선정에 따른 모형 성능 분석

AutoGluon은 모델을 구성하는 데 사용된 개별 변수의 상대적 중요도를 정량적으로 확인할 수 있도록 feature_importance 함수를 제공한다. 이 함수는 대상 변수의 값들을 데이터 행 단위로 무작위로 섞어서 변형하고, 이렇게 변형된 새로운 데이터를 모형에 적용하여 적용 전후 모형의 성능변화를 비교하여 대상 변수의 중요도를 정량적으로 산출한다 (AutoGluon, 2025). 본 연구에서는 이러한 변수 중요도 산출 결과를 기반으로 변수의 중요도가 낮은 항목부터 순차적으로 제거하여 모형을 구축하고 변수의 선별적 적용이 모형의 성능에 미치는 영향을 비교하였다.

2.4 성능 평가

AutoGLuon과 RF, XGB 모형의 성능을 비교하기 위해서 모형성능의 정량적 평가를 위한 지표인 root mean squared error (RMSE), RMSE-standard deviation ratio (RSR) 및 Nash-Sutcliffe coefficient of efficiency (NSE)를 활용하였다 (Eqs. 1 – 3).

RSR 값은 0 – ∞의 범위를 가지며 값이 낮을수록 예측 성능이 뛰어나다고 평가하며 일반적으로 RSR〈0.7인 경우 모형의 결과가 실측값을 잘 예측한 것으로 판단한다 (Moriasi et al., 2007). RMSE는 실측 데이터와 모형의 예측 데이터 사이의 편차를 평균적으로 계산한 지표로 0 – ∞의 범위를 가지며 RMSE 값이 낮을수록 성능이 우수한 것을 나타낸다. NSE는 -∞ – 1의 범위를 가지며 값이 1에 가까울수록 모형이 실측값을 잘 예측하는 것으로 판단한다 (Bennett et al., 2013; Moriasi et al., 2007).

$$\mathrm{RSR} = \frac{\sqrt{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}}{\sqrt{\sum_{t=1}^{n}(y_t - \bar{y}_t)^2}} \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}} \tag{2}$$

$$NSE = 1 - \frac{\sum_{t=1}^{n}(\hat{y}_t - y)^2}{\sum_{t=1}^{n}(y_t - \bar{y}_t)^2} \tag{3}$$

where y_t is observed value at time t,
$\bar{y}_t$ is average of observed value,
$\hat{y}_t$ is model prediction at time t,
n is number of observation

Table 2. Statistical characteristics of input variables

Variable		Average	min	max	Standard deviation
Independent variables	TEMP (℃)	10.852	1.800	24.900	4.783
	pH	7.267	6.300	9.600	0.365
	EC (μS/cm)	176.951	101.000	230.000	24.180
	DO (mg/L)	8.765	1.100	18.600	2.726
	TOC (mg/L)	2.408	1.100	6.800	0.600
	TN (mg/L)	1.715	0.420	5.266	0.545
	TP (mg/L)	0.011	0.003	0.309	0.016
	Q (m^3/s)	23.215	0.000	176.272	19.409
	AVER_TEMP (℃)	12.851	-12.200	32.000	9.993
	RAIN (mm)	2.709	0.000	99.700	9.238
	SUN_H (hr)	15.048	0.000	31.280	3.750
	SUN (MJ/m^2)	7.048	0.000	13.300	7.231
Dependent variable	CHL_A (mg/m^3)	6.388	0.000	70.400	7.605

3. 결과 및 고찰

3.1 입력자료 특성

모형 구축을 위해 사용한 자료의 평균값, 최소값, 최대값, 표준편차값을 아래 Table 2에 나타냈다. 종속변수인 CHL_A의 평균은 6.388이며, 최소값과 최대값은 각각 0.000, 70.400 이었다. Figure 3은 모형 구축에 사용된 종속변수인 CHL_A의 training 및 testing에 사용된 자료의 분포를 보여준다. 연도별 차이는 있으나, training 기간 중 최대 CHL_A 발생 농도와 testing 기간 중 최대 CHL_A 발생 농도는 유사한 분포를 보이는 경향을 확인할 수 있었다.

3.2 모형 성능

AutoGluon을 활용하여 구축된 CHL_A 예측 모형 (M1)의 결과를 RF 및 XGB 모형의 성능과 비교한 결과, AutoGluon, RF 및 XGB의 RSR 값이 각각 0.564, 0.752, 0.811로 분석되어 AutoGluon이 가장 우수한 성능을 보였으며, XGB, RF 순으로 성능이 높은 것으로 확인되었다. 다른 지표인 RMSE, NSE 값 역시 AutoGluon이 4.329, 0.682로 가장 우수한 성능을 보였고, XGB, RF 순으로 성능이 좋은 것으로 분석되었다.

AutoGluon을 적용한 경우, XGB 및 RF에 비해 상대적으로 큰 차이로 우수한 성능을 보이는 것으로 분석되었다. Auto Gluon은 신경망, RF 등 다양한 단위 모델을 포함하고 있으며, 이들 개별 모델의 예측 결과에 가중치를 적용해 조합함으로써 최적의 성능을 도출하는 앙상블 방식을 통해 최종 예측을 수행한다. 따라서, 모형의 최적화 등 구축 과정이 상대적으로 단순함에도 불구하고 비교 대상인 단일 머신러닝 모형을 적용한 경우보다 우수한 성능을 보였다 (Fig. 4).

Figure 5는 실측값과 예측값의 1:1 관계를 보여주며, 그림의 빨간 점선에 가깝게 데이터가 분포할수록 모형이 실측값에 가까운 결과를 예측함을 나타낸다. 그래프를 통해 전체 구간에 걸쳐 AutoGluon이 XGB, RF보다 실측값을 잘 예측하는 경향을 보임을 시각적으로 확인할 수 있다.

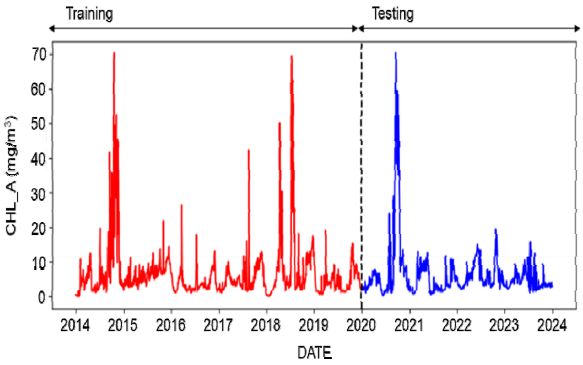
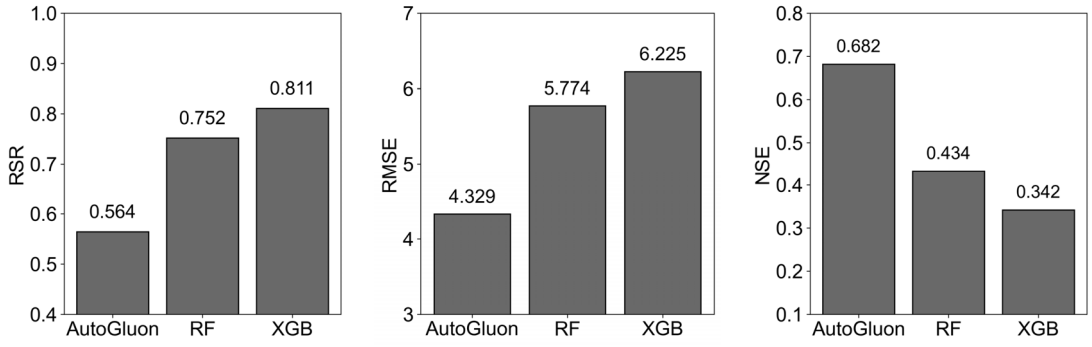
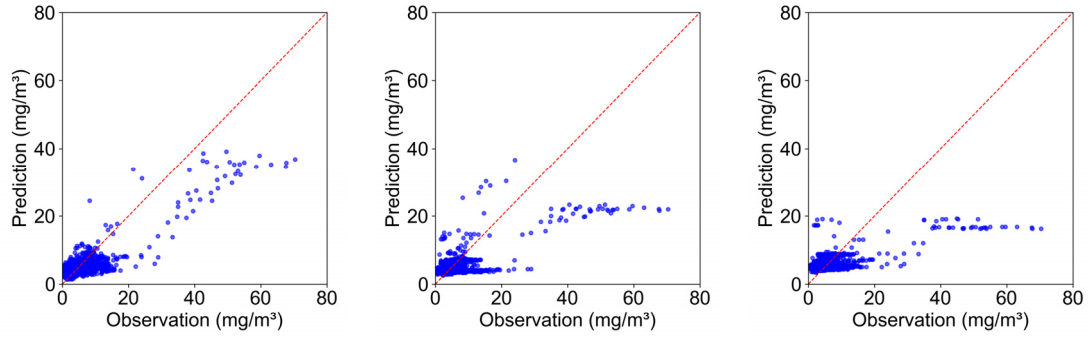


Fig 3. Temporal distribution of CHL_A used for model training and testing.



(a) RSR (b) RMSE (C) NSE

Fig 4. Evaluation of predictive performance of three models (AutoGluon, RF, and XGB)



(a) AutoGluon (b) RF (c) XGB

Fig 5. Relationship between observed and model-predicted CHLA (mg/m3).

3.3 변수 중요도 분석

본 연구에서는 AutoGluon의 내부 알고리즘을 이용하여 모형구축에 사용된 다양한 독립변수가 모형의 결과에 미치는 상대적 중요도를 산출하고 이를 영향이 큰 순으로 위에서 아래로 시각화하여 Fig. 6에 제시하였다.

변수 중요도 값이 클수록 CHL_A 예측에 독립 변수가 미치는 상대적 영향이 더 큰 것을 나타내며, 본 연구에서 사용된 독립변수에 대한 변수 중요도 분석 결과 pH가 모형의 결과에 미치는 영향이 가장 크고 이후 TEMP 〉 Q 〉 TOC 〉 DO 〉 EC 〉 TEMP_AVER 〉 TP 〉 TN 〉 SUN 〉 RAIN 〉 SUN_H 순으로 영향이 큰 것으로 분석되었다.

본 연구에서는 변수 중요도 분석 결과를 기반으로 중요도가 낮은 변수부터 순차적으로 변수를 제거하여 전체 변수를 모두 사용한 모형인 M1을 포함하여, Table 3에 제시된 바와 같이 M2−M6의 5가지 모형을 추가로 구성하여 총 6개의 모형을 구축하고 성능을 비교하였다.

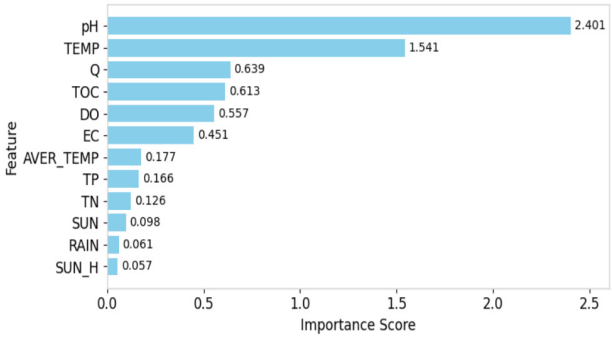


Fig 6. Feature importance of the AutoGluon model.

구축된 6개 모형의 성능을 분석한 결과, RSR 0.542−0.579, RMSE 4.163−4.448, NSE 0.664−0.706의 범위를 보였으며, 변수의 수가 늘어나거나 줄어듦에 따른 일정한 경향은 보이지 않았다 (Fig. 7). 하지만 전체적으로 RSR 〈 0.6 정도 수준의 성능을 보여, 독립변수의 수가 제한적인 경우에도 안정적인 성능을 보임을 확인할 수 있었다.

머신러닝 모형은 모형 구축에 활용되는 데이터의 특성을 학습하여 구현되는 데이터 기반 모형으로, 성능 확보를 위해서는 모형 특성에 맞는 양질의 데이터 확보가 필수적이다. 하지만 현장에서의 수질 자료 취득은 많은 시간과 비용이 요구되며, 특히 TN, TP와 같이 습식 분석이 필요한 항목의 경우 pH, DO, 수온 등 센서를 이용한 취득이 가능한 항목에 비해 일반적으로 더욱 많은 비용이 필요하다.

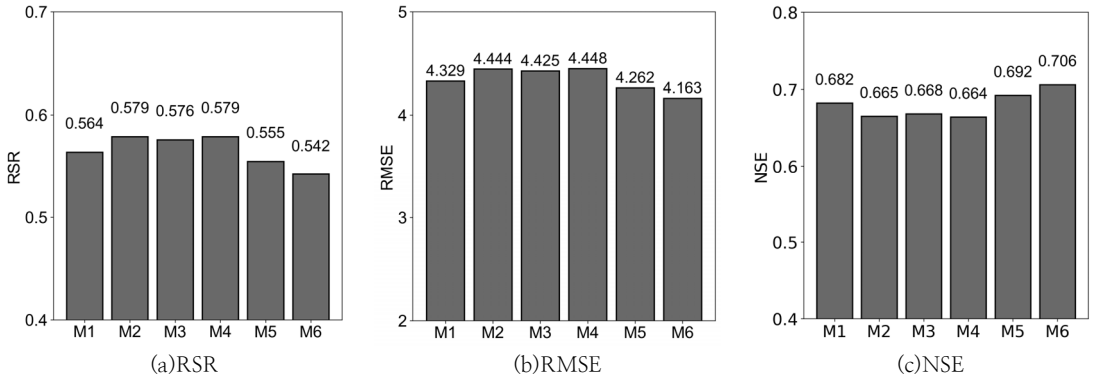
본 연구의 분석을 통해, 상대적으로 제한된 변수를 사용하는 경우에도 AutoML을 통해 일정 수준 이상의 안정적인 예측 성능을 확보할 수 있었다.

3.4 AutoML 모형 성능개선 및 적용 방안

최근 수년간 빠르게 발전하는 머신러닝 모형을 녹조 예측 등 물환경 관리에 적용하기 위한 연구가 다양한 분야에서 계속되고 있다. 하지만 머신러닝 모형의 구축을 위해서는 프로그램 구성 및 데이터 분석에 대한 일정 수준 이상의 전문적인 지식이 필요하다. 대부분의 머신러닝 모형은 Python, R 등을 이용한 직접적인 프로그래밍을 통해 구축되어야 하며, 이러한 특성은 머신러닝 모형을 현장 실무에 직접적으로 적용하는 것을 어렵게 하는 요인 중 하나이다. 또한 데이터 기반 모형의 특성상 머신러닝 모형의 성능은 모형의 구축에 사용된 자료의 특성에 크게 영향을 받게 된다. 따라서 머신러닝

Table 3. Models based on feature importance

Model	Independent Variable
M1	TEMP, pH, EC, DO, TOC, TN, TP ,Q, AVER_TEMP, RAIN, SUN, SUN_H
M2	TEMP, pH, EC, DO, TOC, TN, TP ,Q, AVER_TEMP, RAIN, SUN
M3	TEMP, pH, EC, DO, TOC, TN, TP ,Q, AVER_TEMP, SUN
M4	TEMP, pH, EC, DO, TOC, TN, TP ,Q, AVER_TEMP
M5	TEMP, pH, EC, DO, TOC, TP ,Q, AVER_TEMP
M6	TEMP, pH, EC, DO, TOC ,Q, AVER_TEMP



(a)RSR (b)RMSE (c)NSE

Fig 7. Performance comparison of models constructed based on feature importance.

모형의 성능 향상을 위해 데이터 특성에 맞는 적정한 모형의 선정 및 최적화가 필요하며, 이러한 작업의 수행을 위한 많은 비용 및 시간이 요구된다. AutoML은 이러한 머신러닝 모형의 적용 편의성을 높일 수 있도록 구축된 모형으로, 모형 구축을 위한 시간과 노력을 줄이며, 과정의 단순화를 통해 모형 구축 과정에서의 오류 등을 줄일 수 있는 장점을 가지고 있다.

본 연구는 AutoML을 이용하여 머신러닝 기반 녹조 발생 예측 모형의 활용 효율을 높일 수 있음을 보여주었으며, 입력 변수가 제한적인 경우에도 안정적인 머신러닝 모형 성능의 확보가 가능함을 확인하였다.

현장의 조류 발생은 수질 및 기상 등 다양한 환경 인자의 복합적인 영향에 따른 결과로, 실제 측정된 조류 발생 양상은 일반적으로 알려진 단일 인자와의 이론적 인과관계와는 다른 특성을 보일 수 있다. 머신러닝 모형은 이러한 다양한 환경 인자의 복잡한 상호작용이 최종적으로 반영된 결과인 실측 자료를 기반으로 학습되며, 본 연구에서 도출된 변수 중요도 결과 역시 이러한 현장 자료의 특성을 반영한 것이다. 따라서 본 연구 결과는 기존 이론적 인과관계와는 차이를 보일 수 있다. 향후 머신러닝 결과에 영향을 미치는 다양한 환경 인자에 대한 이해를 높이기 위한 지속적인 연구를 통해 현장 조류 관리 실무에 도움을 주는 의사결정 지원 도구로서 머신러닝 모형의 활용성을 높일 수 있을 것으로 판단된다.

4. 결 론

본 연구에서는 자동화된 머신러닝인 AutoGluon을 이용하여 하천 Chl-*a*를 예측하는 모형을 구축하였으며, 구축된 모형의 성능을 대표적인 머신러닝 알고리즘인 XGB 및 RF와 비교하였다. AutoGluon, RF 및 XGB 모형의 성능을 비교한 결과, RSR, NSE, RMSE 값 모두 AutoGluon으로 구축된 모형이 우수한 성능을 보이는 것으로 분석되어 녹조 발생 예측을 위한 AutoGluon의 적용 가능성을 제시하였다.

또한, AutoGluon 모형 구축에 활용되는 다양한 환경 인자를 대표하는 입력 변수의 구성이 모형 성능에 미치는 영향에 대한 정량적 분석을 위해 변수의 상대적 중요도를 산출하였으며, 중요도가 낮은 변수부터 순차적으로 제거하면서 모형 결과에 미치는 영향을 분석하였다. 분석 결과, 변수의 구성에 따라 차이는 있으나 변수를 제한적으로 적용한 모든 모형이 RSR $<$ 0.6 정도의 성능을 보여, 변수의 사용이 제한적인 경우에도 일정 수준 이상 안정적인 성능의 확보가 가능함을 확인하였다.

최근 물환경 관련 분야에서 고도화된 머신러닝 모형을 활용하기 위한 연구가 지속되고 있다. 머신러닝 모형을 적용하기 위해서는 데이터의 수집, 데이터 전처리, 적합한 모형 선정을 수행하기 위한 프로그래밍 및 모형 구축과 관련된 전문적인 지식이 필요하며, 이는 머신러닝 모형의 현장 적용을 제한하는 요인 중 하나가 되고 있다.

AutoML은 모형 구축의 용이성을 높이고, 모형 구축 과정에서 인위적인 개입을 최소화할 수 있도록 하여 모형 구축의 편의성과 현장 적용성을 높일 수 있는 장점을 가지고 있다. 향후 AutoML 모형의 물환경 분야 활용성을 높이기 위한 지속적인 연구를 통해 물환경 관리의 효율성을 높이는 데 기여할 수 있을 것으로 생각된다.

References

AutoGluon (2025). https://auto.gluon.ai/stable/tutorials/tabular/tabular-indepth.html (accessed May 14, 2025)

Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Northon, J. P., Perrin, C., Pierce, S. A., Robson, B. J., Seppelt, R., Voinov, A., Fath, B. D., and Andreassian, V. (2013). "Characterising performance of environmental models." Environmental Modelling and Software, 40, 1 – 20.

Chen, T., and Guestrin, C. (2016). "XGBoost: A scalable tree boosting system." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785 – 794.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. (2020). "Autogluon-tabular: Robust and accurate AutoML for structured data." arXiv preprint, arXiv:2003.06505.

Erickson, N., Shi, X., Sharpnack, J., and Smola, A. (2022). "Multimodal AutoML for image, text and tabular data." Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 4786 – 4787.

Friedman, J. H. (2001). "Greedy function approximation: A gradient boosting machine." Annals of Statistics, 29 (5), 1189 – 1232.

Gao, X., Lin, J., Qu, C., Wang, C., Wu, A., Zhu, J., and Xu, C. (2024). "Computer-aided diagnostic system with automated deep learning method based on the AutoGluon framework improved the diagnostic accuracy of early esophageal cancer." Journal of Gastrointestinal Oncology, 15(2), 535.

González, E. J., and Roldán, G. (2020). "Eutrophication and phytoplankton: Some generalities from lakes and reservoirs of the Americas." Microalgae: From Physiology to Application, Edited by Vítová, M., IntechOpen, Chapter 2. pp. 27-35.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). "Array programming with NumPy." Nature, 585, 357 – 362. https://doi.org/10.1038/s41586-020-2649-2

Karimian, H., Huang, J., Chen, Y., Wang, Z., and Huang, J. (2023). "A novel framework to predict chlorophyll-a concentrations in water bodies through multi-source big data and machine learning algorithms." Environmental

Science and Pollution Research, 30(32), 79402－79422.
Karmaker, S. K., Hassan, M. M., Smith, M. J., Xu, L., Zhai, C., and Veeramachaneni, K. (2021). "AutoML to date and beyond: Challenges and opportunities." ACM Computing Surveys (CSUR), 54(8), 1－36.
Kim, N. C., and Son, J. H. (2013). "Water quality modeling for Gokgyochun by QUAL2E and QUAL2K." Journal of the Korean Society for Environmental Analysis, 16(2), 84－91.
Korea Meteorological Administration (KMA) (2025). Weather Data Service. https://data.kma.go.kr/cmmn/main.do (accessed May 14, 2025)
LeDell, E., and Poirier, S. (2020). "H2O AutoML: Scalable automatic machine learning." Proceedings of the AutoML Workshop at ICML, Vol. 2020, 24.
Lee, T., Kim, S., Lee, J., Kim, K., Lee, H., and Kim, D. (2024). "A study on algal bloom forecast system based on hydro-meteorological factors in the mainstream of Nakdong river using machine learning." Journal of Wetlands Research, 26(3), 245-253.
Madni, H. A., Umer, M., Ishaq, A., Abuzinadah, N., Saidani, O., Alsubai, S., and Ashraf, I. (2023). "Water-quality prediction based on H2O AutoML and explainable AI techniques." Water, 15(3), 475.
McKinney, W. (2010). "Data structures for statistical computing in Python." Proceedings of the 9th Python in Science Conference, 56－61.
Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations." Transactions of the ASABE, 50(3), 885－900.
MyWater (2025). https://www.water.or.kr/kor/menu/sub.do?menuId=13 (accessed May 1, 2025)
Noh, J., Kim, J. C., and Park, J. (2014). "Turbidity control in downstream of the reservoir: The Nakdong River in Korea." Environmental Earth Sciences, 71, 1871－1880.
Park J. (2023). "Comparison of Automated Machine Learning Model Performance for Predicting Chlorophyll-a Concentration according to Measurement Frequency of Input Data." Journal of Korean Society of Environmental Engineers, 45(4) 201-209.
Park, J., Patel, K., and Lee, W. H. (2024)."Recent advances in algal bloom detection and prediction technology using machine learning." Science of The Total Environment, 173546.
Prasad, D. V. V., Venkataramana, L. Y., Kumar, P. S., Prasannamedha, G., Harshana, S., Srividya, S. J., and Indraganti, S. (2022). "Analysis and prediction of water quality using deep learning and auto deep learning techniques." Science of the Total Environment, 821, 153311.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., and Duchesnay, É. (2011). "Scikit-learn: Machine learning in Python." Journal of Machine Learning Research, 12, 2825－2830.
Salehin, I., Islam, M. S., Saha, P., Noman, S. M., Tuni, A., Hasan, M. M., and Baten, M. A. (2024). "AutoML: A systematic review on automated machine learning with neural architecture search." Journal of Information and Intelligence, 2(1), 52－81.
Shin, C. M., Min, J. H., Park, S. Y., Choi, J., Park, J. H., Song, Y. S., and Kim, K. (2017). "Operational water quality forecast for the Yeongsan River using EFDC model." Journal of Korean Society on Water Environment, 33(2), 219－229.
Shin, Y., Kim, T., Hong, S., Lee, S., Lee, E., Hong, S., and Heo, T. Y. (2020). "Prediction of chlorophyll-a concentrations in the Nakdong River using machine learning methods." Water, 12(6), 1822.
Water Environment Information System (WEIS) (2025). https://water.nier.go.kr/ (accessed May 14, 2025)
Water Resources Management Information System (WAMIS) (2025). http://www.wamis.go.kr/
XGBoost. https://pypi.org/project/xgboost/ (accessed May 14, 2025)