Journal of Wetlands Research
Vol. 26, No.3, August 2024, pp.197-203

ISSN 1229-6031 (Print) / ISSN 2384-0056 (Online)
DOI https://doi.org/10.17663/JWR.2024.26.3.197

데이터 구성에 따른 하천 조류 예측 딥러닝 모형 (TabPFN) 성능 비교

양현석 · 박정수+

*국립한밭대학교 건설환경공학과

Comparing the Performance of a Deep Learning Model (TabPFN) for Predicting River Algal Blooms with Varying Data Composition

Hyunseok Yang* · Jungsu Park*+

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

(Received : 26 April 2024, Revised : 4 June 2024, Accepted : 26 June 2024)

요 약

하천에서 조류의 과다 발생은 취수원 관리 및 정수 처리에 악영향을 줄 수 있어 지속적인 관리가 필요하다. 본 연구에서는 딥러닝 알고리즘 중 작은 규모의 테이블 데이터에서도 상대적으로 우수한 성능을 보이는 것으로 알려진 tabular prior data fitted networks (TabPFN)을 사용하여 조류 발생 지표 중 하나인 chlorophyll-*a* (chl-*a*) 농도를 예측하는 다중 분류 모형을 구축하였다. 모형의 구축을 위해 부여지점 수질자동측정망에서 2014년 1월 1일부터 2022년 12월 31일까지 측정된 일일측정자료를 사용하였으며 입력 자료의 크기가 모형의 성능에 미치는 영향을 확인하기 위해 입력 자료의 평균값을 이용하여 1일, 3일, 6일, 12일의 측정 주기를 가진 입력 자료를 구성하였다. 각 모형의 성능을 비교한 결과 측정 주기가 길어져 입력 자료의 규모가 작은 경우에도 모형이 안정적인 성능을 보이는 것을 확인하였다. 각 모형의 macro average는 precision이 0.77, 0.76, 0.83, 0.84였으며, recall은 0.63, 0.65, 0.66, 0.74 F1-score는 0.67, 0.69, 0.71, 0.78로 분석되었다. Weighted average는 precision이 0.76, 0.77, 0.81, 0.84이며 recall은 0.76, 0.78, 0.81, 0.85 F1-score는 0.74, 0.77, 0.80, 0.84로 분석되었다. 본 연구에서는 TabPFN을 이용하여 구축한 chl-*a* 예측 모형이 작은 규모의 입력 자료에서도 안정적인 성능을 보이는 것을 확인하여 모형구축에 필요한 입력 자료가 제한적인 현장에서의 적용 가능성을 확인하였다.

핵심용어 : 딥러닝, 머신러닝, 머신러닝 자동화, 분류 모형, 조류관리

Abstract

The algal blooms in rivers can negatively affect water source management and water treatment processes, necessitating continuous management. In this study, a multi-classification model was developed to predict the concentration of chlorophyll-*a* (chl-*a*), one of the key indicators of algal blooms, using Tabular Prior Fitted Networks (TabPFN), a novel deep learning algorithm known for its relatively superior performance on small tabular datasets.

The model was developed using daily observation data collected at Buyeo water quality monitoring station from January 1, 2014, to December 31, 2022. The collected data were averaged to construct input data sets with measurement frequencies of 1 day, 3 days, 6 days, 12 days. The performance comparison of the four models, constructed with input data on observation frequencies of 1 day, 3 days, 6 days, and 12 days, showed that the model exhibits stable performance even when the measurement frequency is longer and the number of observations is smaller. The macro average for each model were analyzed as follows: Precision was 0.77, 0.76, 0.83, 0.84; Recall was 0.63, 0.65, 0.66, 0.74; F1-score was 0.67, 0.69, 0.71, 0.78. For the weighted average, Precision was 0.76, 0.77, 0.81, 0.84; Recall was 0.76, 0.78, 0.81, 0.85; F1-score was 0.74, 0.77, 0.80, 0.84. This study demonstrates that the chl-*a* prediction model constructed using TabPFN exhibits stable performance even with small-scale input data, verifying the feasibility of its application in fields where the input data required for model construction is limited.

Key words : Algal bloom management, Automated machine learning, Classification model, Deep learning, Machine learning

+To whom correspondence should be addressed.
Department of Civil and Environmental Engineering, Hanbat National University
E-mail : parkjs@hanbat.ac.kr

• **Hyunseok Yang** Hanbat National University, Korea / Master Student (didgustjr563@naver.com)
• **Jungsu Park** Hanbat National University, Korea / Associate professor (parkjs@hanbat.ac.kr)

1. 서 론

농업 및 산업활동의 증가로 인하여 하천으로 유입되는 영양염류 등의 오염물질과 기후변화에 따른 수온 증가는 하천의 조류를 지속적으로 발생시키는 원인 중 하나이며, 하천의 조류는 취수원의 수질 악화와 정수 처리비용 증가와 같은 문제를 발생시키기에 지속적인 관리가 필요한 수질 인자이다 (Schindler, 2006; Wurtsbaugh et al., 2019). 일반적으로 chlorophyll-*a* (chl-*a*) 농도는 하천의 조류를 정량적으로 나타낼 수 있는 수질 인자로 사용되며, chl-*a* 농도를 통해 하천의 부영양화 및 조류 번성의 정도를 추정하기도 한다 (Chen, 2024). 안정적인 수질관리를 위하여 실시간 수질 자료를 통한 정확한 수질 예측이 중요하며, 이를 위해 다양한 분야에서 널리 사용되고 있는 머신러닝 모형을 수질 예측에 활용하기 위한 연구가 지속되고 있다 (Blix and Eltoft, 2018; Kwon wt al., 2018).

머신러닝은 모형구축에 활용되는 입력 자료의 특성을 학습하여 데이터 기반의 의사결정을 내리기에 변수 간 물리, 화학, 생물학적 특성을 기반으로 하는 계수의 산정을 필요로 하지 않으며 복잡성과 비선형성을 해석하기에 효과적이라는 장점을 가지고 있다 (Amorim et al., 2021; Li et al., 2024). 수질관리 분야에서도 XGBoost (XGB), random forest (RF)와 같은 앙상블 (ensemble) 모형과 convolution neural networks (CNN), long short term memory (LSTM)과 같은 딥러닝 (deep learning) 모형 등 다양한 머신러닝 모형들이 수질 예측 등에 적용되고 있다 (Barzegar et al., 2020; Kim et al., 2022; Kim and Ahn, 2022; Shin et al., 2020).

머신러닝을 이용한 모형의 구축은 입력 자료의 전처리, 입력 자료의 적합한 모형의 선정, 선정된 모형의 최적화 단계로 진행된다. 특히, 입력 자료의 적합한 모형의 선정 및 최적화 과정은 머신러닝에 대한 전문 지식과 많은 시간, 컴퓨터 기능이 요구되며, 최근 몇 년간 머신러닝 모형구축의 편의성 및 활용성 등을 높일 수 있도록 모형구축을 자동화하는 automated machine learning (autoML)등에 대한 연구가 활발히 이루어지고 있다 (Chen et al., 2021; Moon et al., 2019; Tuggener et al., 2019). 머신러닝 모형은 일반적으로 초매개변수 (hyperparameter)를 조정하는 최적화 과정을 통해서 모형의 내부 구조를 결정한 후 모형을 구축하고 도출된 결과를 비교하여 최적의 모형을 찾는 과정으로 이루어져 있으나 입력 자료의 구성과 사용된 모형에 따라 같은 과정을 반복 수행해야 하기에 실제 수질 예측을 위해서 머신러닝 모형을 사용하기에는 많은 시간과 노력을 요구한다. 또한 머신러닝과 같은 데이터 기반 모형은 모형의 구축에 사용된 입력 자료의 특성이 성능에 많은 영향을 미치게 되어 모형구축에 필요한 충분한 자료의 확보가 중요하나 이를 위해서는 많은 비용과 인력 및 시간이 소요되어 머신러닝 모형의 활용성을 제한하는 요인이 되기도 한다.

본 연구는 입력 자료의 전처리 및 모형의 최적화 등 모형 구축 과정의 편의성을 높였으며 상대적으로 자료 수가 많지 않은 소규모 자료에서도 좋은 성능을 보이는 것으로 알려진 최신 딥러닝 모형 중 하나인 tabular prior-data fitted networks (TabPFN)을 이용하여 하천 chl-*a* 농도를 예측하는 분류 모형을 구축하였다. 모형의 구축을 위해 일일 수질 측정자료를 이용하였으며 측정자료의 평균값을 산정하여 자료의 크기를 다르게 한 다양한 입력 자료를 구성하여 입력 자료의 규모에 따른 모형의 성능을 비교하였다.

2. 재료 및 실험방법

2.1 입력 자료

본 연구에서는 환경부 국립환경과학원에서 제공하는 물환경정보시스템의 수질자동측정망 중 부여지점 (S03005)에서 2014년 1월 1일부터 2022년 12월 31일까지 측정된 일별 수질 측정자료를 활용하였다 (NEIR, 2023)(Fig. 1). 금강 유역은 총 3,611.82km의 길이와 9,914.02km^2 의 면적을 가지고 있으며 대전광역시 등 8개의 시·도 및 천안시, 청주시, 전주시 등 50여개의 시·군·구로 구성되어 있다. 금강 유역의 수질자동측정망 중 부여지점은 금강 본류 하류수질을 측정하는 지점으로 갑천, 미호강 등 금강유역의 주요 지류가 합류한 후의 수질을 모니터링할 수 있는 지점이다 (ME, 2023).

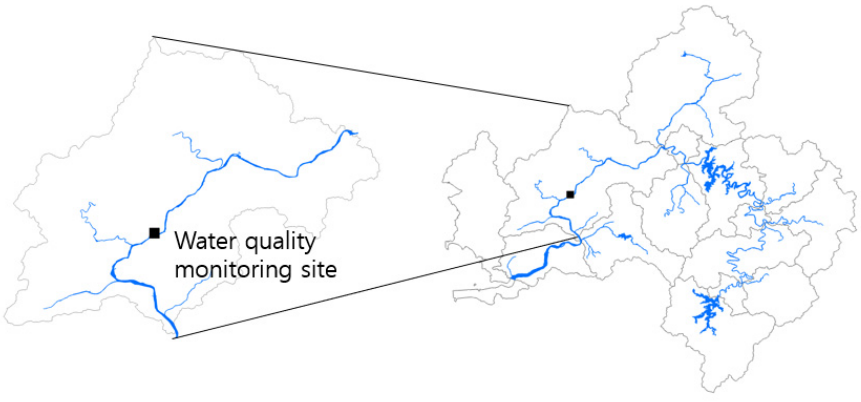


Fig. 1. Research site.

측정자료 중 수온(TEMP), pH, 전기전도도 (EC), 용존산소량 (DO), 총유기탄소 (TOC), 총질소 (TN), 총인 (TP)은 모형의 독립변수로 설정하고 조류의 발생정도를 나타내는 대표적인 수질인자인 chl-*a*를 종속변수로 설정하였다. 본 연구에서는 하천의 chl-*a*를 예측하는 다중 분류 모형을 구축하였으며, chl-*a*의 분류 기준은 world health organization (WHO)에서 제시한 기준에 따라 10 ㎍/L미만은 low (class1), 10 ㎍/L이상 50 ㎍/L미만은 moderate (class 2), 50 ㎍/L이상은 high (class 3)으로 분류하여 모형의 종속변수로 사용하였다 (Loftin et al., 2016).

2.2 모형구축

TabPFN은 트랜스포머(transformer) 알고리즘을 활용한 딥러닝 모형 중 하나로 상대적으로 소규모로 구성된 테이블 형식 데이터를 이용한 분류모형을 효과적으로 구축할 수 있는 알고리즘이다. 소규모 데이터에 대한 사전 학습을 통해 예측의 대상이 되는 데이터의 사후 예측 분포를 가장 잘 근사화하는 것을 목표로 하는 알고리즘으로 모형의 구축 시 소규모 데이터에 대한 사전 학습 결과를 활용하여 상대적으로 소규모의 입력자료에 대해서도 우수한 성능을 보이는 것으로 알려져 있다. 또한 autoML type의 library로 데이터 전처리 및 초매개변수 최적화 등 모형의 구축 및 최적화 과정의 편의성을 높이고 기존의 모형들에 비해 훨씬 빠른 학습속도를 보이는 장점을 가지고 있다 (Hollmann et al., 2022; Magadán, L ete al., 2023.). TabPFN 알고리즘을 이용한 모형 구축 과정에 대한 모식도를 Fig. 2에 시각화하였다.

본 연구에서는 TabPFN open source library (Hollmann et al., 2022)를 이용하고 환경부 물환경정보시스템에 공개된 부여지점의 총 3,287회의 일일 측정자료를 입력 자료로 활용하여 하천 chl-*a* 예측을 위한 분류 모형을 구축하였다 (model 1). 또한 일일 측정자료의 3일 평균값을 계산하여 1,097개의 데이터를 이용한 model 2, 6일 평균값을 계산하여 549개의 데이터를 이용한 model 3, 12일 평균값을 계산하여 275개의 데이터를 이용한 model 4 총 네 가지의 모형을 구축하여 입력 자료의 크기가 모형의 성능에 미치는 영향을 비교하였다. 각 model에 활용된 데이터 수의 변화를 Fig. 3에 시각화하였다.

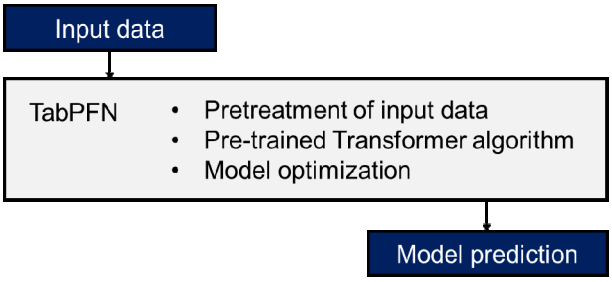
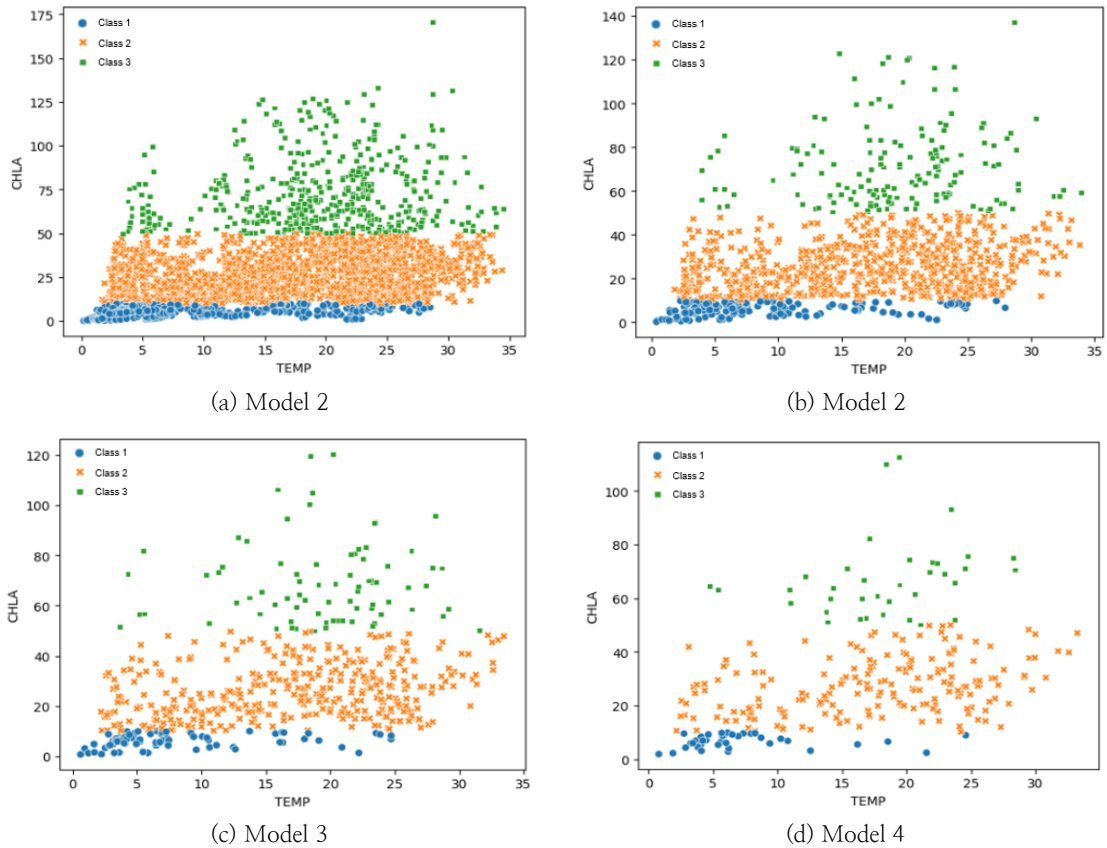


Fig. 2. Schematic of model development process in TabPFN.



(a) Model 2 (b) Model 2

(c) Model 3 (d) Model 4

Fig. 3. The changes in the amount of data for each model.

측정 항목은 각각 TEMP 11.4%, pH 11.5%, EC 12%, DO 12.5%, TOC 17.7%, TN 17.8%, TP 17.5%, chl-*a* 18.4%의 결측값을 포함하고 있으나 결측값이 대부분 수질의 변화가 크지 않은 기간에 분포하고 있어 결측값 주변의 k개의 측정자료로 결측값을 보정 하는 k-nearest neighbor (KNN)을 이용하여 결측값의 보정을 수행하였다. KNN은 python open source library인 scikit-learn을 이용하였으며, k 값은 3으로 적용하여 수행하였다 (Pedregosa et al., 2011).

전체 입력 자료 중 2014년 1월 1일부터 2020년 12월 31일까지의 자료는 학습자료 (training)로 사용하였고 2021년 1월 1일부터 2022년 12월 31일까지의 자료는 학습된 모형의 성능 평가 (testing)에 사용하였다. Training 자료와 testing 자료의 비율은 각각 78%와 22%로 구성하였으며 모형의 최적화 작업은 python open source library인 scikit-learn의 grid search를 이용하였다 (Pedregosa et al., 2011).

2.3 모형 성능 평가 방법

각 모형의 성능 평가 방법은 분류 모형의 성능 평가 방법 중 하나인 혼동행렬 (Confusion matrix)을 이용하여 모형 성능을 평가하였다.

Confusion matrix는 머신러닝 분류 모형의 실측값과 예측값 사이의 분포를 행렬의 형태로 나타낸 성능 평가 지표이며 실측값과 예측값이 일치한다면 True, 일치하지 않는다면 False로 구분한다. 예를 들어 실측값이 positive일 때 모형의 예측값도 positive라면 True Positive (TP), 실측값이 positive일 때 예측값이 negative라면 False Negative (FN), 실측값이 negative일 때 예측값이 positive라면 False Positive (FP), 실측값이 negative일 때 예측값이 negative라면 True Negative (TN)으로 분류한다 (Table 1).

Confusion matrix는 각 class에 해당하는 정밀도 (precision), 재현율 (recall), F1-score를 확인하고 산술 평균 (macro average), 가중 평균 (weighted average)을 계산하여 모형의 성능을 비교할 수 있다. Precision은 모형의 예측값 중 실측값과 일치하는 비율을 의미하고 recall은 실측값 중 모형의 예측값과 일치하는 비율을 의미한다. F1-score는 precision과 recall의 조화평균으로 두 지표 중 어느 하나만 너무 높거나 낮을경우 결과 해석 과정에서 발생할 수 있는 오류를 방지하기 위해 사용하는 지표이다. Macro average는 전체 class에서 해당 지표의 값을 모두 더한 후 class 수로 나누어 평균값을 구하는 것을 의미하고 weighted average는 class의 지표값에 전체 데이터 수에 대한 해당 class의 데이터 수의 비율만큼 가중하여 전체 평균을 계산한 것을 의미한다 (Eq. 1, 2, 3).

Table 1. Confusion matrix

Confusion Matrix		Predictive Values	
		Positive (P)	Negative (N)
Actual Values	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

$$\mathrm{Pr}ecision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$F1-score = 2 \times \frac{\mathrm{Pr}ecision \times Recall}{\mathrm{Pr}ecision + Recall} \tag{3}$$

3. 결과 및 고찰

3.1 모형 결과

본 연구는 입력 자료를 하천의 chl-*a*의 농도에 따라 3개의 class를 분류하고 TabPFN을 이용하여 다중 분류 모형을 구축하였다. Model 1의 training 자료 중 각 class에 해당하는 데이터 수는 class 1이 402개, class 2가 1,769개, class 3이 386개로 구성되었으며 TabPFN 구축된 모형의 성능을 Table 2에 제시하여 각 class 별 성능을 확인하였다.

Table 2. Performance of model 1 using observation data

	Precision	Recall	F1-score
Class 1	0.69	0.43	0.53
Class 2	0.76	0.92	0.83
Class 3	0.88	0.53	0.66
Macro average	0.77	0.63	0.67
Weighted average	0.76	0.76	0.74

Table 3. Performance of models with different observation range

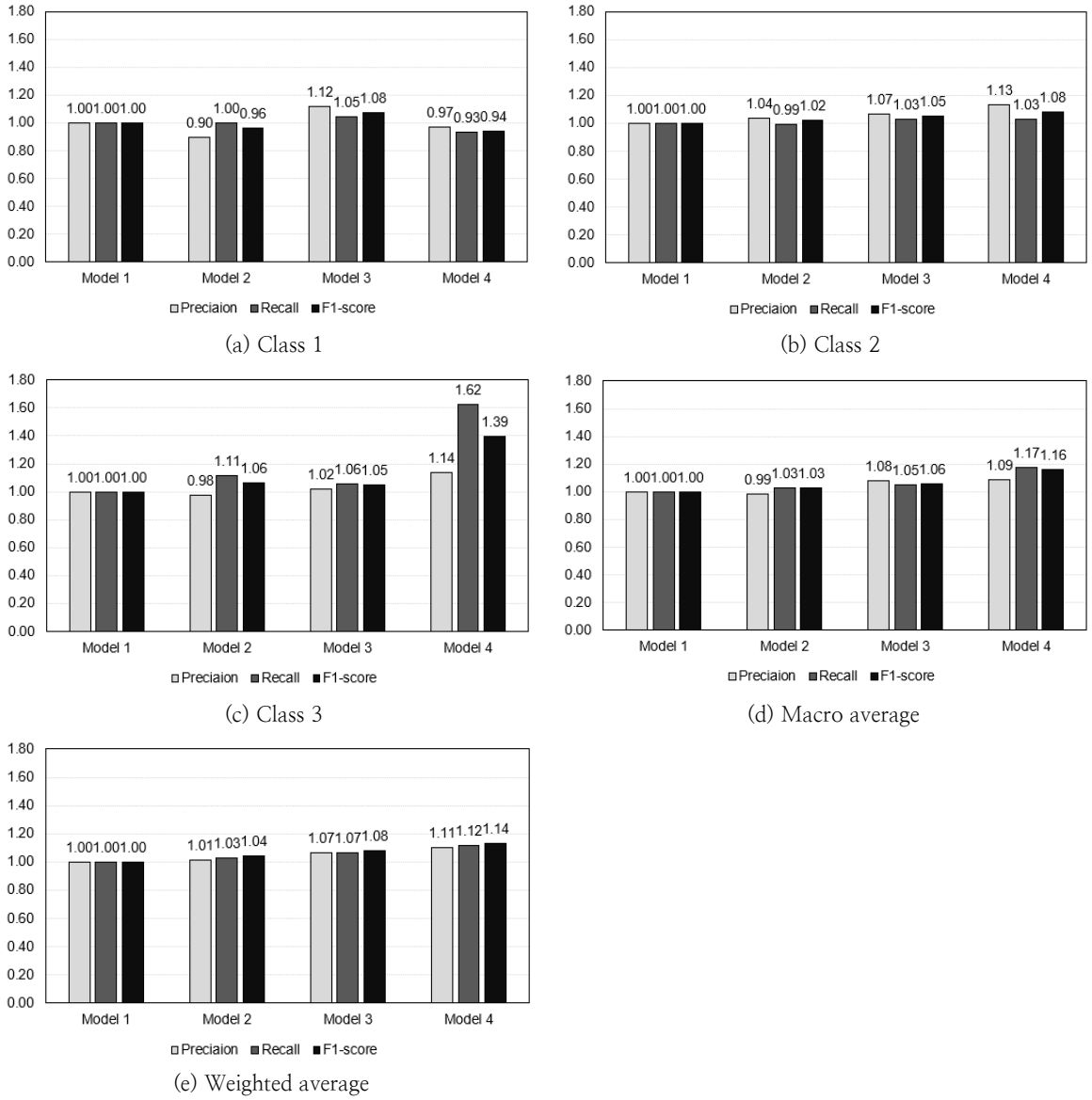
		Model 2	Model 3	Model 4
Class 1	Precision	0.62	0.77	0.67
	Recall	0.43	0.45	0.40
	F1-score	0.51	0.57	0.50
Class 2	Precision	0.79	0.81	0.86
	Recall	0.91	0.95	0.95
	F1-score	0.85	0.87	0.90
Class 3	Precision	0.86	0.90	1.00
	Recall	0.59	0.56	0.86
	F1-score	0.70	0.69	0.92
Macro average	Precision	0.76	0.83	0.84
	Recall	0.65	0.66	0.74
	F1-score	0.69	0.71	0.78
Weighted average	Precision	0.77	0.81	0.84
	Recall	0.78	0.81	0.85
	F1-score	0.77	0.80	0.84

Model 1의 precision, recall, F1-score를 각 class 별로 확인하면 class 1과 class 2는 각각 0.69, 0.43, 0.53과 0.76, 0.92, 0.83으로 분석되었고 class 3은 0.88, 0.53, 0.66으로 분석되었다. Macro average와 weighted average는 각각 0.77, 0.63, 0.67과 0.76, 0.76, 0.74로 분석되어 class 1과 class 3의 recall이 다른 지표들에 비하여 상대적으로 낮은 것을 확인하였다. 본 연구에서 모형의 구축에 사용된 자료중 class 2에 비해 class 1과 3에 해당되는 자료수가 적어 class 별 자료의 불균형이 있다. 이러한 자료의 불균형은 class별 모형 성능의 차이가 발생하는 원인이 될 수 있으며, 향후 입력 자료의 불균형 해소 등 모형 성능의 개선을 위한 지속적인 연구가 필요할 것으로 판단된다.

3.2 입력 자료 수에 따른 모형 성능 비교

입력 자료를 활용한 데이터 수에 따른 모형의 성능 변화를 확인하기 위하여 일일 측정자료를 3일, 6일, 12일로 나누어 각각의 평균값으로 입력 자료를 구성했을 경우 모형의 성능에 미치는 영향을 비교하였다. 입력 자료의 측정 주기에 따라 구성된 3개의 모형인 model 2, model 3, model 4의 성능 분석 결과를 Table 3에 제시하였다.

실측값을 입력 자료로 사용하여 구축된 model 1을 기준으로 각 model 별 모형의 class 별 성능을 비교하였다. Class 1은 model 1~model 4의 precision은 0.62~0.77의 범위를 보이고 데이터가 가장 작은 model 4의 precision이 0.67로 데이터 수에 따른 성능 차이는 크지 않았다. Recall과 F1-score도 각 0.40~0.45와 0.50~0.57의 범위를 보이고 있으며, model 4가 0.40과 0.50으로 가장 낮은 성능을 보이기는 했으나, 기존 모형에 비하여 큰 차이는 나지 않는 것을 확인하였다. Class 2는 class 1과 마찬가지로 모델별로 성능의 차이가 크지 않고 안정적인 성능을 보였다.



(a) Class 1

(b) Class 2

(c) Class 3

(d) Macro average

(e) Weighted average

Fig. 4. Performance of models with different observation range.

Class 3에서는 precision과 recall이 각각 0.86~1.00과 0.53~0.86으로 상대적으로 큰 차이의 범위를 보이며 데이터가 가장 적은 model 4에서 1.00과 0.86으로 가장 높은 지표값을 보이는 것을 확인하였다. F1-score도 0.66~0.92의 범위를 보이고 있으며 model 4에서 0.92의 가장 높은 성능을 보여 class 3에서 가장 큰 성능 차이를 보였다. 각 class의 성능을 산술 평균으로 계산하여 나타낸 macro average에서는 precision과 recall이 각각 0.76~0.84와 0.63~0.74의 범위를 보였으며 F1-score도 0.67~0.78의 범위를 보여 각 model 별 성능의 차이가 상대적으로 크지 않다는 것을 확인하였으며, class 3의 성능 차이로 인하여 macro average도 model 4가 가장 높은 지표값을 보이는 것을 확인하였다. Weighted average도 각 class의 데이터 수만큼 가중치를 포함하여 평균을 구하는 것이기에 macro average와 같이 각 model 별로 성능의 차이가 크지 않으며 class 3의 영향으로 인하여 model 4에서 좋은 성능을 보이는 것으로 판단된다. Model 1을 기준으로 각 model의 상대적 성능을 비교하여 성능변화율을 Fig. 4에 시각화하였다.

전체 모형의 성능을 비교한 결과 측정 주기를 길게 하여 모형의 입력 자료가 작은 규모로 구성된 model 4가 다른 model에 비해 유사하거나 우수한 성능을 보이는 것으로 확인되어 머신러닝 모형의 구축을 위해 필요한 입력 자료의 확보가 제한적인 경우에도 상대적으로 안정적 성능의 확보가 가능함을 보여주었다.

본 연구를 통해 TabPFN을 이용한 하천 chl-*a* 예측 모형을 구축하고 입력 자료로 활용된 자료의 수에 따른 모형의 성능을 비교하였다. 머신러닝 모형의 구축은 적용 대상 모형의 특성에 따라 적합한 입력 자료의 scaling을 수행하는 등 모형의 선정 및 모형의 초매개변수 최적화 등의 과정을 거쳐 수행된다. 이러한 과정은 관련 분야의 전문지식과 함께 많은 시간과 노력이 필요하다.

최근 머신러닝 모형의 구축에 필요한 전처리 및 모형 선정 등 모형구축 과정의 편의성을 높여주기 위해 autoML과 같이 다양한 기술이 개발되고 있으며, 상대적으로 머신러닝에 대한 전문지식이 많지 않은 경우에도 머신러닝 모형의 구축과 활용을 용이하게 할 수 있도록 하여 머신러닝 모형의 실무 활용성을 높이기 위한 연구가 계속되고 있다 (Chauhan et al., 2020; Feurer et al., 2015).

머신러닝 모형은 일반적으로 복잡한 내부 알고리즘을 가지고 있으며, 이러한 머신러닝 모형의 실질적인 적용을 위해서는 모형의 구축에 필요한 충분한 측정자료의 확보가 필요하다. 우리나라에서는 주기적인 하천 및 저수지의 현장 수질 측정 결과를 물환경정보시스템을 통해 공개하고 있으며, 수질자동측정망의 경우 1시간 간격 수질 측정 결과를 확인할 수 있으나 현재 운영되고 있는 대부분의 측정망은 주 혹은 월 단위의 측정 빈도로 측정 결과를 제공하고 있어 (ME, 2024) 딥러닝과 같이 상대적으로 복잡한 모형을 효율적으로 활용하기에 충분한 자료의 확보가 쉽지 않은 경우가 많다.

본 연구의 결과는 TabPFN을 이용하여 상대적으로 입력 자료가 적은 경우에도 안정적인 모형 성능을 확보할 수 있음을 보여주었으며, 입력 자료의 취득이 제한적인 현장 등에서 머신러닝 모형의 실무적용성을 높일수 있음을 확인하였다. 또한 입력 자료의 전처리 및 모형의 최적화에 필요한 과정을 최소화하는 autoML 모형을 이용하여 머신러닝 모형의 사용 편의성과 활용성을 높일 수 있음을 보여주었다. 머신러닝 모형과 같은 데이터 기반 모형의 성능을 높이기 위해서는 모형이 특성에 적합한 양질의 자료확보가 중요하며, 향후 자료취득의 효율성 향상과 자료취득 현황 등 현장의 현실적 특성을 고려한 머신러닝 모형의 실무적용성을 높이기 위한 지속적인 연구로 하천 수질관리 기술을 고도화하고 현장 관리 효율을 높이는데 기여할 수 있을 것으로 판단된다.

4. 결 론

머신러닝 모형의 구축을 위해서는 모형구축에 필요한 충분한 자료의 확보가 필요하며 머신러닝 모형중 상대적으로 좀더 복잡한 내부알고리즘을 가지는 딥러닝 모형의 경우 특히 모형의 성능확보를 위한 충분한 자료의 확보가 중요하다. 하지만 수질측정 자료의 경우 충분한 현장 자료의 확보가 어려운 경우가 많으며 이는 수질관리를 위한 딥러닝 모형의 현장 실무 적용을 제한하는 원인중 하나이다. 본 연구에서는 소규모 데이터에 대해서도 상대적으로 우수한 성능을 보이는 것으로 알려진 딥러닝 open source library인 TabPFN을 이용하여 조류 발생 정도를 예측하는 분류 모형을 구축하고 입력자료의 규모가 모형의 성능에 미치는 영향에 대한 분석을 수행하였다. 모형의 구축에는 부여지점 수질자동측정망의 현장 측정자료를 활용하였으며 chl-*a* 농도에 따라 조류 발생 정도를 3개의 class로 구분하여 모형의 구축에 활용하였다.

데이터 규모가 모형의 성능에 미치는 영향을 확인하기 위해 실측된 데이터인 일일 측정자료를 이용하여 1일, 3일, 6일, 12일의 평균값을 구해 다양한 측정 횟수를 가지는 데이터를 구성하였으며 구축된 데이터를 각각 model 1~4의 구축에 활용하여 데이터 크기가 모형의 성능에 미치는 영향을 확인하였다.

분석 결과 데이터 수가 가장 작은 model 4가 더 많은 자료를 이용하여 구축된 model 1~3에 비해 유사하거나 상대적으로 우수한 성능을 보여 입력 자료의 확보가 제한적인 경우에도 TabPFN을 이용하여 안정적인 성능을 보이는 딥러닝 모형의 구축이 가능함을 확인하였다.

본 연구를 통해 입력 자료 수가 제한적인 경우에도 현장 관리를 위한 딥러닝 모형의 적용성을 높일 있음을 확인하였다. 또한 딥러닝 모형의 구축을 위해 필요한 측정자료의 전처리 및 모형이 최적화 과정을 최소화하는 autoML library의 활용을 통해 딥러닝 모형의 실무적용성을 높일 수 있음을 보여주었다. 향후 관련 분야의 지속적인 연구를

통해 고도화된 딥러닝 모형의 현장 적용성을 높일 수 있을 것으로 판단된다.

Acknowledgment

이 성과는 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2022R1F1A1065518).

References

Amorim, F. D. L. L. D., Rick, J., Lohmann, G., and Wiltshire, K. H. (2021). Evaluation of machine learning predictions of a highly resolved time series of chlorophyll-*a* concentration. Applied Sciences, 11(16), 7208.

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

Blix, K., and Eltoft, T. (2018). Machine learning automatic model selection algorithm for oceanic chlorophyll-*a* content retrieval. Remote Sensing, 10(5), 775.

Chauhan, K., Jani, S., Thakkar, D., Dave, R., Bhatia, J., Tanwar, S., and Obaidat, M. S. (2020). Automated machine learning: The new wave of machine learning. IEEE. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 205-212).

Chen, C., Chen, Q., Yao, S., He, M., Zhang, J., Li, G., and Lin, Y. (2024). Combining physical-based model and machine learning to forecast chlorophyll-*a* concentration in freshwater lakes. Science of The Total Environment, 907, 168097.

Chen, Y. W., Song, Q., and Hu, X. (2021). Techniques for automated machine learning. ACM SIGKDD Explorations Newsletter, 22(2), 35-50.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., and Hutter, F. (2015). Efficient and robust automated machine learning. Advances in neural information processing systems, 28.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. (2022). Tabpfn: A transformer that solves small tabular classification problems in a second. arXiv preprint arXiv:2207.01848.

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

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

Kwon, Y. S., Baek, S. H., Lim, Y. K., Pyo, J., Ligaray, M., Park, Y., and Cho, K. H. (2018). Monitoring coastal chlorophyll-a concentrations in coastal areas using machine learning models. Water, 10(8), 1020.

Li, H., Li, X., Song, D., Nie, J., and Liang, S. (2024). Prediction on daily spatial distribution of chlorophyll-*a* in coastal seas using a synthetic method of remote sensing, machine learning and numerical modeling. Science of The Total Environment, 910, 168642.

Loftin, K. A., Graham, J. L., Hilborn, E. D., Lehmann, S. C., Meyer, M. T., Dietze, J. E., and Griffith, C. B. (2016). Cyanotoxins in inland lakes of the United States: Occurrence and potential recreational health risks in the EPA National Lakes Assessment 2007. Harmful algae, 56, 77-90.

Magadán, L., Roldán-Gómez, J., Granda, J. C., & Suárez, F. J. (2023). Early fault classification in rotating machinery with limited data using TabPFN. IEEE Sensors Journal.

Ministry of Environment (ME). (2023). The First Comprehensive Water Management Plan for the Geum River Basin, 2021-2030. Geum River Basin Management Commission pp 17-19

Ministry of Environment (ME). (2024). Water Quality Monitoring Program. Ministry of Environment pp 6-7

Moon, Y. H., Shin, I. H., Lee, Y. J., and Min, D. G. (2019). Recent research & development trends in automated machine learning. Electronics and Telecommunications Trends, 34(4), 32-42

National Institute of Environmental Research (NIER). (2023). Water Environment Information System, https://water.nier.go.kr/web, Accessed 4 December 2023

https:// water.nier.go.kr/web. Accessed 4 December 2023.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Schindler, D. W. (2006). Recent advances in the understanding and management of eutrophication. Limnology and oceanography, 51(1part2), 356-363.

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.

Tuggener, L., Amirian, M., Rombach, K., Lörwald, S., Varlet, A., Westermann, C., and Stadelmann, T. (2019). Automated machine learning in practice: state of the art and recent results. IEEE. In 2019 6th Swiss Conference on Data Science (SDS) (pp. 31-36).

Wurtsbaugh, W. A., Paerl, H. W., and Dodds, W. K. (2019). Nutrients, eutrophication and harmful algal blooms along the freshwater to marine continuum. Wiley Interdisciplinary Reviews: Water, 6(5), e1373.