ISSN(Print): 1225-7672 / ISSN(Online): 2287-822X
DOI https://doi.org/10.11001/jksww.2024.38.1.1

방류수질 예측을 위한 AI 모델 적용 및 평가

Application and evaluation for effluent water quality prediction using artificial intelligence model

김민철[1*] · 박영호[1] · 유광태[2] · 김종락[2]
Mincheol Kim[1*] · Youngho Park[1] · Kwangtae You[2] · Jongrack Kim[2]

[1]서울물재생시설공단 물재생연구소
[2](주)유앤유
[1]Water Regeneration Research Center, Seoul Water Recycling Corporation
[2]UnU Inc.

ABSTRACT

Occurrence of process environment changes, such as influent load variances and process condition changes, can reduce treatment efficiency, increasing effluent water quality. In order to prevent exceeding effluent standards, it is necessary to manage effluent water quality based on process operation data including influent and process condition before exceeding occur. Accordingly, the development of the effluent water quality prediction system and the application of technology to wastewater treatment processes are getting attention. Therefore, in this study, through the multi-channel measuring instruments in the bio-reactor and smart multi-item water quality sensors (location in bio-reactor influent/effluent) were installed in The Seonam water recycling center #2 treatment plant series 3, it was collected water quality data centering around COD, T-N. Using the collected data, the artificial intelligence-based effluent quality prediction model was developed, and relative errors were compared with effluent TMS measurement data. Through relative error comparison, the applicability of the artificial intelligence-based effluent water quality prediction model in wastewater treatment process was reviewed.

Key words: Effluent water quality prediction, Artificial intelligence model, Wastewater treatment process, Smart sensor

주제어: 방류 수질 예측, 인공지능 모델, 하수처리공정, 스마트센서

Received 1 September 2023, revised 11 December 2023, accepted 20 December 2023.

*Corresponding author: Mincheol Kim (E-mail: mckim@swr.or.kr; Fax. 82-2-3410-9769, Tel. 82-2-3410-9720)

1 김민철 (연구소장) / Mincheol Kim (Research Director)
서울특별시 강남구 개포로 625, 06333
625, Gaepo-ro, Gangnam-gu, Seoul 06333, Republic of Korea

1 박영호 (주임) / Youngho Park (Assistant Manager)
서울특별시 강남구 개포로 625, 06333
625, Gaepo-ro, Gangnam-gu, Seoul 06333, Republic of Korea

2 유광태 (대표이사) / Kwangtae You (CEO)
서울특별시 구로구 디지털로33길 27, 08380
27, Digital-ro 33-gil, Guro-gu, Seoul 08380, Republic of Korea

2 김종락 (연구소장) / Jongrack Kim (Research Director)
서울특별시 구로구 디지털로33길 27, 08380
27, Digital-ro 33-gil, Guro-gu, Seoul 08380, Republic of Korea

1. 서 론

국내 하수처리장에서는 궁극적으로 하수처리공정 TMS 모니터링 항목의 수질기준 이내의 안정적인 운영을 목표로 하고 있다. TMS 측정 시 유입부하 및 운영효율 변동, TMS장치 점검 등에 의해 수질 변동이 발생할 수 있으며, 또한 최근 기후변화 및 도시환경의 급격한 변화로 하수처리장에 유입되는 유량과 수질 변동이 발생하면서 공정 운영 중 예상치 못한 수질 사고가 발생하는 등 여러 어려움이 발생하고 있다. 이에 변동적인 유량 및 수질에 대한 대응 체계의 필요성이 대두되고 있으며, 방류 수질 예측 시스템의 개발을 통해 실제 하수처리장에 적용하는 등 다양한 기술개발 및 기술적용이 이뤄지고 있다 (Haghiabi et al., 2018). 특히 하수처리장의 경우 미생물의 대사 과정에 의해 처리되기 때문에 다양한 매개변수에 영향을 받으며, 또한 시간에 따라 실시간으로 변하는 비선형적인 특성을 나타낸다 (Hamed et al., 2004; Hong et al., 2004; Mjalli et al., 2007). 1990년대부터 복잡한 비선형 관계를 증명하기 위해 서포트 벡터 머신, 인공신경망, 유전 알고리즘, K-means 클러스터링 등 머신러닝 기법을 활용한 접근 방법이 하수처리 공정 및 오염물질 예측에 있어 다양하게 적용되어 왔다 (Ahmed et al., 2019). 머신러닝 기법은 다양한 알고리즘을 이용해 데이터를 분석하고, 이를 통해 학습하며, 학습한 내용을 기반으로 판단이나 예측을 수행하는 방법이다. 이는 의사 결정 기준에 대한 구체적인 지침을 소프트웨어에 직접 코딩해 넣는 것이 아닌, 대량의 데이터와 알고리즘을 통해 컴퓨터 그 자체를 학습시켜 수행 방법을 익히게하는 방법이다. 널리 사용중인 수학적 모델 중 하나인 활성슬러지모델(Activated Sludge Models, ASMs)과는 다르게 유입성상 분석값, 초기값 등 기초 필수 데이터 값에 대한 입력 없이 사용할 수 있다는 점에서 하수처리장에 적용하기 적합한 모델로 볼 수 있다 (You, 2020). 또한 활성슬러지 모델의 경우 하루 데이터에 대한 값에 대해 하루 평균값을 예측을 하는 반면, 머신러닝기법을 활용한 모델의 경우 1시간, 6시간 후 등 시간단위에 대한 예측이 가능하다는 점에서 하수처리장 적용 시 더욱 이점이 될 수 있다. Ribeiro et al (2013)은 포르투갈 북부에 위치한 폐수처리공장의 성능을 예측하기 위해 1년 동안 측정된 처리공정의 일별 평균값을 기반으로 앞선 설명한 머신러닝 기법 중 하나인 서포트 벡터 머신 알고리즘을 이용하여 BOD와 TSS예측 모델을 개발하고, 적용성을 평가한 사례가 있다 (Ribeiro et al., 2013). 이처럼 대용량 데이터 세트를 처리하기 위한 기술개발이 급속도로 발전하고 있으며, 수자원 관리, 홍수 예측, 강우 유출 모델링, 하수처리 모델링 등 환경공학 분야에서 적용사례 또한 증가하고 있다 (Mosavi et al., 2018; Adnan et al., 2021). 하지만 이런 대부분의 예측 모델들은 한 가지 종류의 모델만을 사용하여 예측을 수행하거나, 초기 모델 구성 시 결정한 입력 항목들을 고정적으로 사용하여 예측을 진행하기 때문에 한 가지 입력 항목에서만 이상 데이터가 발생하여도 전체 예측 성능에 영향을 주거나, 예측 결과와 상관성이 낮은 항목이 포함될 시 예측 성능을 저하시키는 문제점이 있다.

따라서 본 연구에서는 2개 이상의 서로 다른 예측 모델의 결과를 조합하여 최종 예측값을 도출하는 앙상블 모델을 적용하여 TMS 미래 수질을 예측할 수 있는 모델을 개발하였으며, 예측 성능을 향상시킬 수 있는 방안과 함께 개발된 예측 모델을 실제 대형하수처리장에 적용함으로써 대형 하수처리장에서의 방류수질 예측 가능성에 대해 검토하고자 했다.

2. 연구방법

2.1 연구대상 지역 및 데이터 수집

본 연구는 서남물재생센터를 대상으로 인공지능(AI) 기반 하수처리 수질 예측 모델을 적용하였으며 데이터 수집을 위한 데이터베이스를 구축하고, SCADA(Supervisory Control and Data Acquisition) 운영자료 및 TMS 수질 자료를 연계 수집하였다.

기존에 계측되던 처리장 운영자료(SCADA 연계) 외에 스마트센서를 설치하여 계측자료를 구축한 데이터베이스에 통합 관리되도록 구성하였으며, 스마트센서는 Fig. 1과 같이 2처리장 3계열의 최초침전지 후단(생물반응조 유입)과 최종침전지 후단에 설치하여 TOC, $TCOD_{CR}$, T-N, NH_4^+-N, T-P, TSS, 전기전도도(Electrical conductivity, EC), 수온 등의 항목을 1분 단위로 측정하여 데이터베이스에 저장하며(Table 1), 이를 예측 모델 학습에 사용하였다.

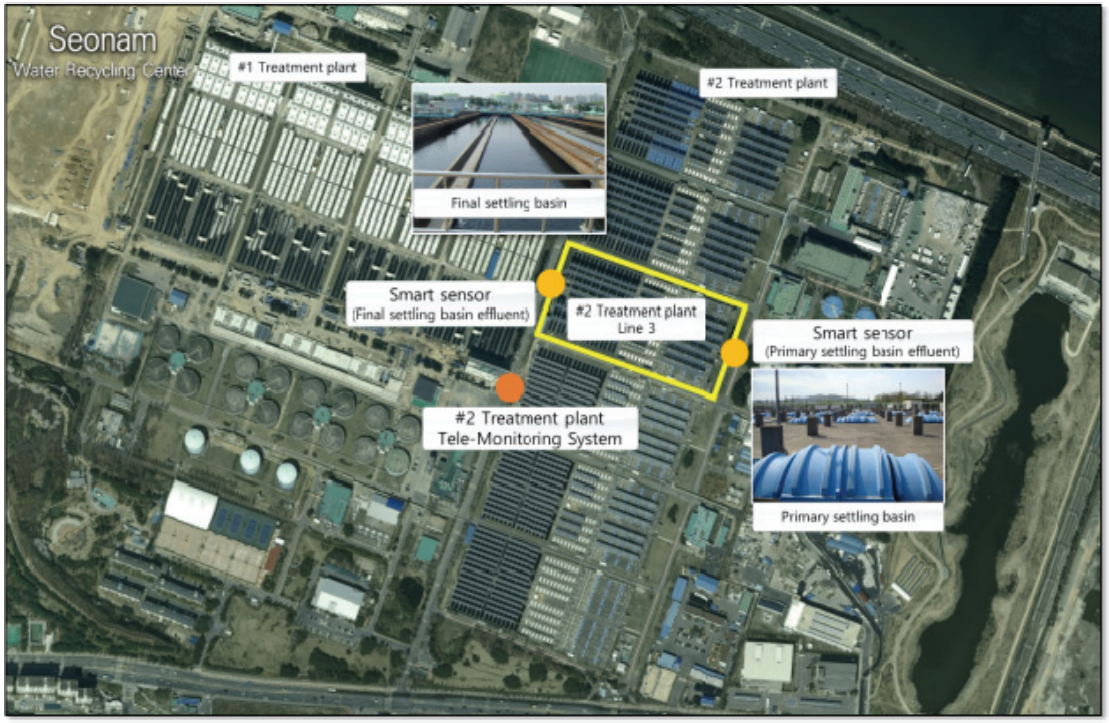


Fig. 1. Installation location of Smart Sensor.

Table 1. SCADA connection and smart sensor data collection items

No	Classification	Collection items
1	Water Quality (Connection of SCADA)	Influent quantity_Series 1-6
2		Drawing quantity of primary sludge_Series 1-6
3		Bioreactor MLSS_Series 1-6
4		Bioreactor DO_Series 1-6 A/B
5		Quantity of internal recycle_Series 1-6 (8 lines per series)
6		Quantity of return sludge_Series 1-6 A/B
7		Quantity of excess sludge_Series 1-6
8	#2 treatment plant TMS (Connection of SCADA)	COD measurement, State code
9		T-N measurement, State code
10		T-P measurement, State code
11	Smart sensor (#2 treatment plant series 3)	Bioreactor influent/effluent TOC
12		Bioreactor influent/effluent $TCOD_{CR}$
13		Bioreactor influent/effluent T-N
14		Bioreactor influent/effluent NH_4^+-N
15		Bioreactor influent/effluent T-P
16		Bioreactor influent/effluent TSS
17		Bioreactor influent/effluent EC
18		Bioreactor influent/effluent temperature

2.2 방류 수질 예측 모델 선정

본 연구에서 적용된 인공지능(AI) 기반 하수처리 수질 예측 모델은 Fig. 2와 같이 실시간 수집되는 스마트센서 측정데이터 및 SCADA 데이터, TMS 측정데이터로부터 시간 평균자료를 계산한 후 이를 이용한 자율지도 학습(Self Supervised Learning) 기반의 수질 예측 모델을 적용하였다. 해당 모델은 과거 운영 데이터를 사용하여 1시간 주기로 예측 모델을 자동으로 학습하고, 1시간, 3시간, 6시간, 12시간 후 미래 방류 수질을 예측한다.

생물학적 처리공정 특성상 하수처리시설은 유입부하 조건에 따라 방류 수질 예측 모델이 선형/비선형 특성을 가질 수 있다. 이에 따라 다양한 모델을 이용하여 최적 모델을 선정하는 앙상블 모델(Ensemble model)을 시스템에 구축하였으며, 하나의 모델을 사용하지 않고 2개 이상의 모델을 결합하여 다양한 조건을 반영한 최적의 예측값을 산정할 수 있도록 구성하였다. Table 2와 같이 PLS, SVM, RANSAC, MLP 모델이 적용되어 각 모델들을 동시에 검토하며, 과거 데이터를 사용하여 모델을 학습한 후 오차가 적은 모델을 예측 모델로 선정하므로, 예측성능이 우수한 최적 모델은 항상 일정하지 않고, 모델 학습에 사용된 운전데이터에 따라 변경될 수 있다.

수집된 수질 및 운전 인자들 중 예측하고자 하는 수질 항목에 영향을 미치는 항목들을 일부 선택하여 예측에 사용하였으며, Table 3과 같이 활용 가능한 입력 항목 후보들 가운데 실제 예측에 사용되는 항목들은 매 분석마다 학습을 통해 새롭게 선정하여 데이터 변동 및 시계열적 패턴에 따라 적절한 예측 모델 입력 항목을 업데이트하도록 구성하였다.

2.3 방류 수질 예측 모델 개발 절차

본 연구에 사용되는 방류 수질 예측 모델은 Usman et al. (2022)이 제시한 머신러닝(ML)과 순환신경망(RNN) 알고리즘을 활용한 하이브리드 모델링 방법을 기반으로 하였으며, 본 연구에 적용된 방류 수질 예측 모델은 최적 변수 선정(Feature Selection), 모델 생성 및 학습(Build and Train Model), 자료 예측(Data Prediction) 3단계의 과정으로 진행하였다.

최적 변수 선정 단계에서는 SCADA 연계 항목, 스마트센서 항목, TMS 수질 항목의 현재 및 과거 자료를 모두 사용하여 예측하고자 하는 방류 수질 항목 사이의 상관관계 분석을 수행 후 상관도가 높은 변수를 모델 입력 항목으로 최적 선정하였다.

두 번째 단계는 선택된 입력변수를 사용하여 앙상블 모델을 생성하는 것으로 PLS, SVN, RANSAC, MLP 모델을 다수 생성하였으며, 이후 설정된 모델 학습 기간의 자료를 조회하여 각 모델을 학습 수행 후 모델별 상대오차를 계산 후 오차가 가장 적은 모델을 예측 성능이 우수한 것으로 판단하여 예측 모델 그룹을 생성하였다.

세 번째 단계에서 현재 측정된 데이터를 사용하여 예측 모델 그룹에 포함된 모델별로 미래 수질을 계산하고 이를 통합하여 최적 예측값을 계산하였으며 본 연구에서는 일반적으로 가장 좋은 성능을 보이는 평균값을 적용하였다.

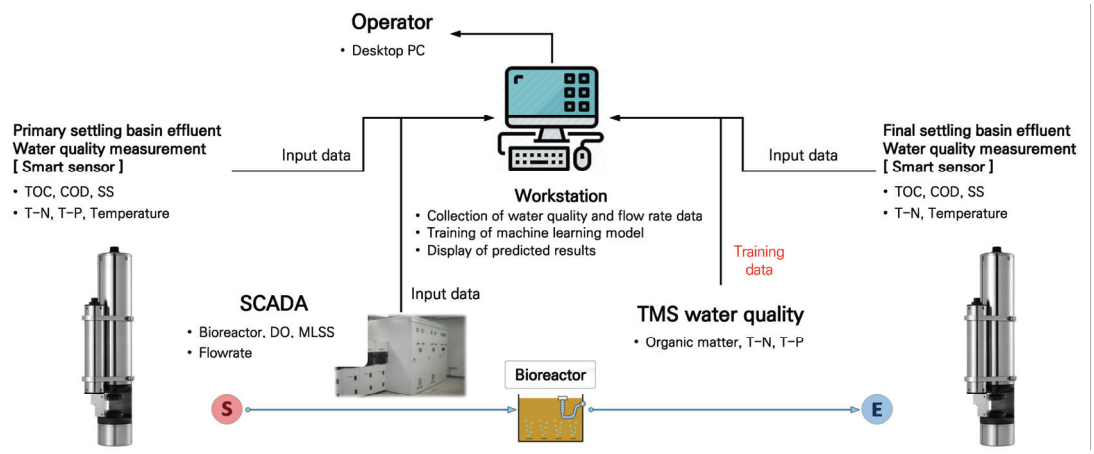


Fig. 2. Diagram of facility and equipment construction for effluent quality prediction.

Table 2. Water quality prediction model and ensemble model

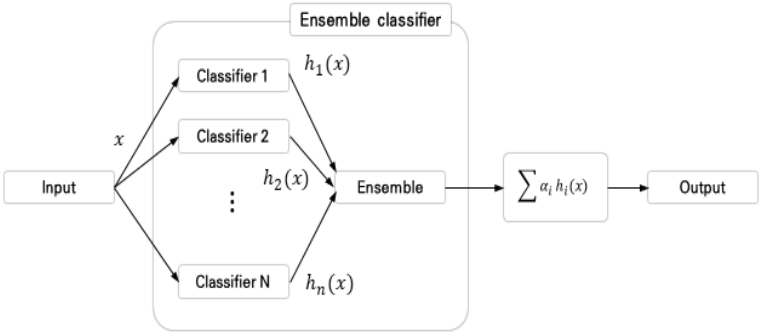
Models	Characteristics
PLS (Partial Least Squares Regression)	Through principal component analysis of the main ingredients, regression analysis is performed by projecting the data on the axis that best represent the variation between independent and dependent variables, while avoiding multicollinearity.
SVM (Support Vector Machine Regression)	As a method to reduce errors between groups by minimizing the margin between data groups, various kernel functions can be used to both linear and nonlinear data.
RANSAC (RANdom SAmple Consensus Regression)	The model which maximizes consensus between model data by training multiple randomly selected subsets of data individually, can generate a robust regression model on outliers.
MLP (Multi-Layer Perceptron)	An algorithm with an artificial neural network structure that includes one or more hidden layers between input data and output values, which exhibits different performance depending on the number of nodes and the type of activation function used in the hidden layers.
Ensemble model Concept map	

Table 3. Inputs and prediction items used in the effluent water quality prediction model

Classification	Unit process	Data items	Note
Input items	Influent	· Smart sensor COD, T-N, T-P, TSS, Temp, EC · Smart sensor raw data	Hourly Average
	Primary settling basin	· Drawing quantity of sludge	Hourly Average
	Bioreactor	· MLSS, DO	Hourly Average
	Final Settling basin	· Drawing quantity of sludge · Smart sensor COD, T-N, T-P, TSS, Temp, EC · Smart sensor raw data	Hourly Average
	TMS	· COD, T-N	Hourly Average
Prediction items	TMS	· COD and T-N data after 1, 3, 6, 12hr	Hourly Average

2.4 방류 수질 예측을 위한 입력 항목

방류 수질 예측에 필요한 운영 인자를 도출하기 위해 측정값의 변화가 상대적으로 크고, 예측 항목값에 많은 영향을 주는 것으로 판단되는 항목을 묶어 데이터 항목을 다음의 4개 그룹으로 분류하였으며 항목별 시간 평균자료를 활용하였다.

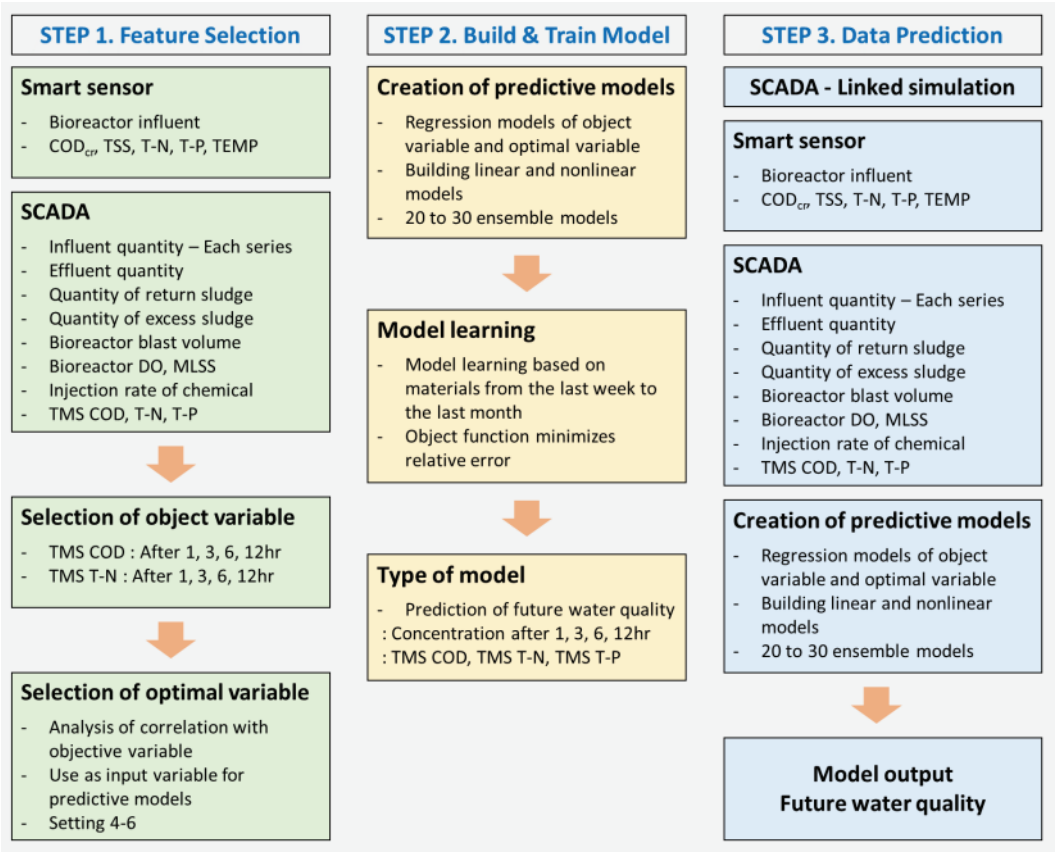


Fig. 3. The process of water quality prediction in this study.

- 그룹1 : 스마트센서 수질 측정항목(TOC, COD_{CR}, T-N, NH_4-N(유입), T-P, TSS, TEMP, EC)
- 그룹2 : 생물 반응조 DO, MLSS
- 그룹3 : 유입유량, 잉여슬러지 인발량
- 그룹4 : 내부반송유량, 슬러지반송유량

그룹2~4의 항목은 스마트센서가 설치된(서남물재생센터) 제2처리장 3계열의 자료만 사용한 경우와 전체 계열자료를 모두 사용한 경우로 다시 분류하였다.

- 그룹2-1, 3-1, 4-1 : 3계열 자료만 사용
- 그룹2-2, 3-2, 4-2 : 6개 모든 계열의 평균자료 사용

3. 연구결과

3.1 방류 수질 예측 모델 입력 항목 그룹 선정 결과

2처리장 방류 TMS 항목별 예측 모델을 입력 항목 조합을 다르게 설정하여 총 7가지 모델 케이스로 설정하였다 (Table 4). 스마트센서 자료만 사용하는 모델(GRP1), 스마트센서와 3계열 자료를 사용하는 3개의 모델 (GRP2/3/4-1), 스마트센서와 모든 계열 평균자료를 사용하는 3개의 모델 (GRP2/3/4-2)로 구분하였다.

각각의 입력 항목 그룹으로부터 예측 모델의 입력 항목을 선정하고 이를 사용하여 예측을 수행한 결과를 비교하여 예측성능이 우수한 입력 항목 그룹 및 예측 모델을 선정하고자 하였다. 2023년 1월 20일부터 2023년 2월 19일까지 약 31일간 매시간 예측을 수행하였으며, 모델별 학습에 사용되는 자료는 최근 14일간 데이터를 사용하도록 설정하였다. 예측성능은 실제 TMS 측정값과의 절대오차 및 상대오차 평균을 계산하여 Table 5와 같이 비교하였다.

스마트센서 데이터만 사용한 GRP1 케이스의 경우 단기 예측(1시간, 3시간)에 대해 다른 예측 모델에 비해 좋은 성능을 보였으며, 특히 T-N 예측 모델의 경우 해당 모델의 성능이 가장 우수하게 나타났다.

Table 4. Combination of prediction models by input item group

Input items		Prediction models case						
Group	Items	GRP 1	GRP 2-1	GRP 3-1	GRP 4-1	GRP 2-2	GRP 3-2	GRP 4-2
1	Smart sensor measurement list (Primary and final settling basin)	○	○	○	○	○	○	○
2-1	Bioreactor DO, MLSS (Series 3)		○	○	○			
3-1	Influent quantity, drawing quantity of sludge (Series 3)			○	○			
4-1	Quantity of internal recycle and return sludge (Series 3)				○			
2-2	Bioreactor DO, MLSS (Average of all series)					○	○	○
3-2	Influent quantity, drawing quantity of sludge (Average of all series)						○	○
4-2	Quantity of internal recycle and return sludge (Average of all series)							○

Table 5. Absolute and relative errors by model of predicted items

Predicted items		COD				T-N			
		After 1h	After 3h	After 6h	After 12h	After 1h	After 3h	After 6h	After 12h
Absolute error (mg/L)	GRP1	1.83	1.89	1.99	2.70	0.56	0.81	1.03	1.30
	GRP2-1	1.84	2.05	2.00	2.71	0.58	0.79	1.07	1.11
	GRP3-1	1.91	1.85	1.88	2.38	0.58	0.78	0.96	0.98
	GRP4-1	1.95	1.92	1.97	2.52	0.59	0.78	0.86	0.92
	GRP2-2	1.84	1.99	1.93	2.70	0.55	0.79	1.04	1.29
	GRP3-2	1.89	1.89	1.93	2.38	0.56	0.79	0.85	0.95
	GRP4-2	1.97	1.91	2.05	2.86	0.56	0.86	0.97	0.84
Relative error (%)	GRP1	15.54	16.06	16.83	22.19	3.83	6.28	8.43	11.06
	GRP2-1	15.58	17.05	16.95	22.23	3.97	6.08	8.79	9.11
	GRP3-1	16.00	15.82	16.13	20.50	3.94	6.02	7.79	7.91
	GRP4-1	16.22	16.24	16.86	20.85	4.09	6.04	6.89	7.35
	GRP2-2	15.52	16.61	16.37	22.14	3.70	6.08	8.59	11.01
	GRP3-2	15.86	16.03	16.41	20.38	3.85	6.15	6.84	7.70
	GRP4-2	16.54	16.19	17.26	23.27	3.77	6.78	7.90	6.71

※ Minimum relative error (red number), Within 5% of minimum relative error (yellow background)

COD 예측 모델의 경우, 3계열 항목 및 전계열 항목을 사용한 결과 간의 차이가 크지 않았다. 반면, T-N 예측 모델의 경우 3계열 항목보다 전 계열 항목을 사용한 경우의 오차가 더 적게 나타났다.

T-N, COD 두 가지 예측 항목 모두 모델 입력 항목으로 스마트센서, 생물반응조 DO/MLSS, 유입유량/슬러지 인발유량을 사용한 경우 대부분 상대오차가 적으며, 내부반송유량/슬러지반송유량은 예측성능에 큰 영향을 주지 않았다.

이와 같은 결과로부터, 예측성능이 우수한 스마트센

Table 6. Classification of predictive performance period

Classification	Predictive performance period
Period 1	2023.03.01.~2023.03.18. (For 18 days)
Period 2	2023.03.19.~2023.04.06. (For 19 days)
Period 3	2023.04.07.~2023.04.19. (For 13 days)
Period 4	2023.04.20.~2023.05.01. (For 12 days)
Period 5	2023.05.02.~2023.05.17. (For 16 days)

서 자료, 전 계열 생물반응조/유입유량/슬러지인발유량 자료 사용 모델 (GRP3-2)을 대상으로 학습기간을 변경하여 추가적인 예측 및 결과 분석을 수행하였다.

3.2 학습 데이터 기간별 수질 예측 결과

일반적으로 TMS 수질의 급격한 변화는 유입부하 및 운영효율의 변동이나 TMS 장치 오류 및 점검 등의 원인으로 발생할 수 있다. 그러나 이러한 수질 변화는 발생 빈도가 높지 않아 모델 학습을 위해 충분한 데이터 확보에 한계가 있다. 갑작스러운 수질변동을 포함한 데이터가 충분하지 않으므로, 한정된 데이터 내에서 모델 예측성능을 향상시키는 방법으로 데이터 학습 기간 최적화를 검토 및 제안하고자 하였다.

방류 수질 예측 모델은 입력항목의 데이터 분포 및 특성을 학습하여 예측을 수행하므로 모델 예측성능은 입력데이터의 영향을 크게 받는다. 즉, 모델 학습 기간에 따라 예측성능 또한 달라질 수 있다.

따라서, 학습 데이터 기간 변화에 따른 예측성능을 비교하여 예측성능 향상을 위한 최적 학습 데이터 기간을 도출하고자 하였다. 모델 학습 데이터 기간은 모델 학습 시점을 기준으로 최근 3일(D3), 7일(D7), 14일(D14), 21일(D21), 28일(D28)간의 시간 평균자료를 사용한 5가지로 설정하였다.

학습 데이터 기간에 따른 예측은 2023년 3월 1일부터 2023년 5월 17일까지 진행하였으며, 해당 기간에 측정된 TMS COD와 T-N은 Fig. 4와 같은 트랜드를 가진다. 두 항목에서 공통적으로 3월 17일, 3월 22일~24일, 4월 7일~10일 등 세 차례 데이터가 누락된 기간이 존재하며, 이는 TMS 점검 및 수리 등에 의한 것으로 판단된다. T-N 항목(Fig. 4(b))의 경우 4월 29일~30일, 5월 6일~7일 두 차례 측정값이 크게 감소하였

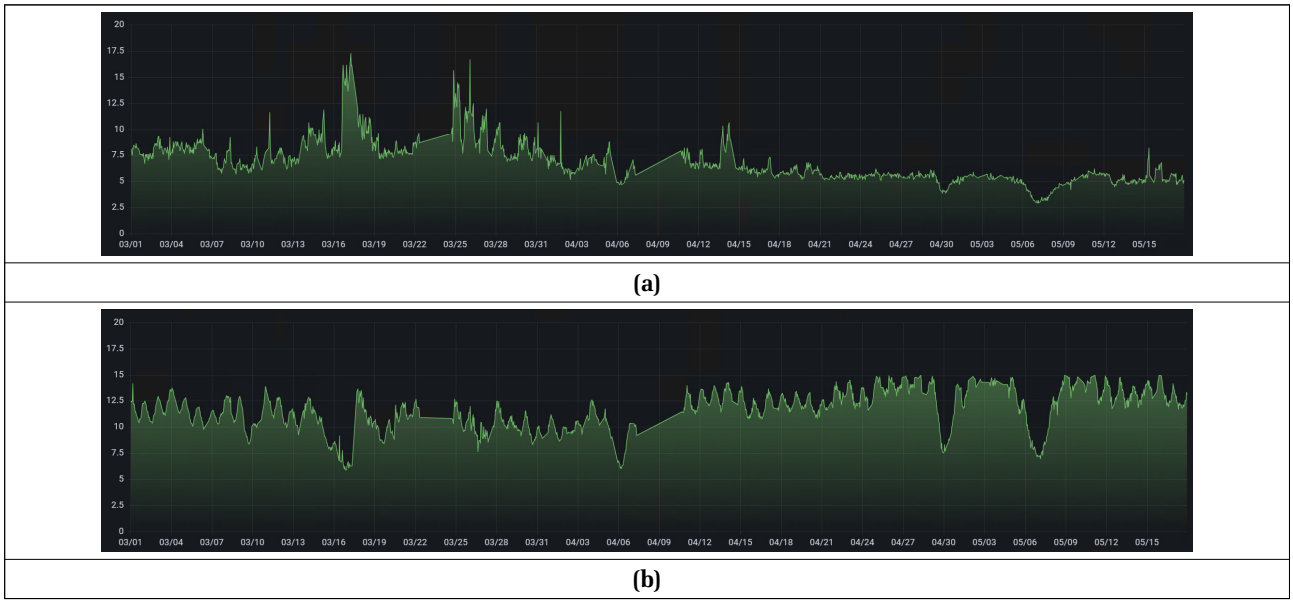


Fig. 4. Effluent (a) COD and (b) T-N trend of total analysis period.

으며, 동일한 기간에 COD 항목(Fig. 4(a)) 또한 소폭 감소하는 것을 확인할 수 있다. 이는 해당 기간에 발생한 강우의 영향으로 보인다.

각 예측 항목에 대하여 모델 학습 데이터 기간을 달리하여 예측 모델 학습 및 예측 수행하였으며, 예측을 수행한 총 기간을 5개의 분석 기간으로 나누어 예측오차를 계산 및 비교하였다 (Table 6).

수질 예측 모델 입력 항목 그룹별 예측 결과에서 선정한 스마트센서 자료, 전계열 생물반응조/유입유량/슬러지인발유량 자료 사용 모델에 각 학습 데이터 기간에 대하여 1시간, 3시간, 6시간, 12시간 후 방류 COD, T-N 예측을 수행하였다.

Fig. 5, Fig. 6은 각 학습 데이터 기간별 1시간 후 COD, T-N 예측 결과이며, 분석 기간별 계산된 상대오차 평균을 바탕으로 예측오차가 작은 학습 데이터 기간을 최적 학습 기간으로 선정하고자 하였다. 3시간,

Fig. 5. Prediction results of effluent COD after 1 hour using learning data period (a) D3, (b) D7, (c) D14, (d) D21 and (e) D28.

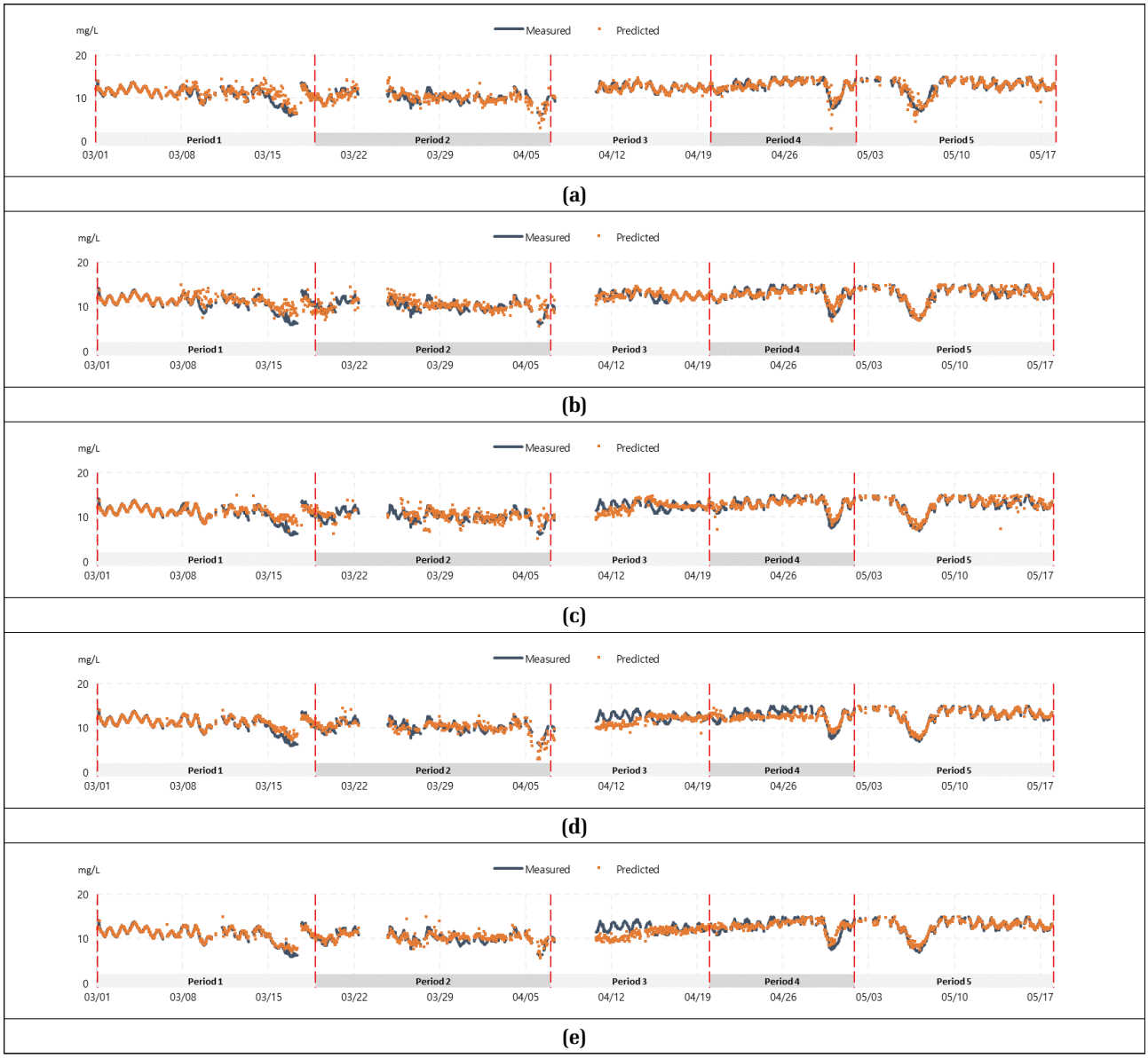


Fig. 6. Prediction results of effluent T-N after 1 hour using learning data period (a) D3, (b) D7, (c) D14, (d) D21 and (e) D28.

6시간, 12시간의 최적 학습 기간도 동일한 방법으로 선정하였다.

T-N 예측 결과, Table 7의 기간 1의 경우 COD, T-N 항목 모두 대부분 21일 이상의 학습 데이터 사용 시 오차가 작게 나타났으며, 1시간, 3시간 후 예측 모델의 경우 모두 28일간의 장기데이터 사용 시 예측성능이 높게 나타났다. 1시간 후 COD 예측 모델에서 7일간, 6시간 후 T-N 예측 모델에서 3일간의 학습 데이터 사용 시 예측오차가 낮게 나타났으나 21일간의 데이터를 사용했을 경우와 유사한 수준으로 나타났다.

기간 2에서는 COD의 경우 모든 예측 항목에서 28일간의 장기데이터를 사용하여 학습했을 때 예측성능이 가장 우수하게 나타났으며, 1시간, 3시간, 6시간 후 예측 모델에서는 21일간의 학습 데이터 사용 시 두 번째로 낮은 오차를 보였다. 12시간 후 예측 모델에서는 3일간의 학습 데이터를 사용하였을 때 오차가 두 번째로 낮았다. T-N의 경우, 6시간 후 예측 모델을 제외한 모든 시간의 예측 모델에서 3일간의 학습 데이

Table 7. COD, T-N prediction relative error of effluent water by learning data period - (Period 1)

Unit : %	COD				T-N			
Learning data	After 1 hr	After 3 hr	After 6 hr	After 12 hr	After 1 hr	After 3 hr	After 6 hr	After 12 hr
3 day	5.3±5.4	8.7±8.9	11.8±11.6	12.4±13.3	8.3±11.2	11.5±12.3	12.1±12.6	14.1±13.6
7 day	4.7±4.9	6.7±7.0	9.4±9.1	10.0±9.9	9.2±11.7	12.2±14.8	14.5±15.1	13.0±12.6
14 day	4.8±5.4	6.6±7.4	8.7±9.2	10.3±11.3	7.7±11.5	10.7±12.9	12.6±13.2	12.5±12.1
21 day	4.5±5.0	6.4±7.1	7.9±8.3	9.8±10.2	7.2±12.6	9.7±12.8	12.5±13.0	12.7±12.3
28 day	4.5±4.6	6.3±6.6	8.0±8.1	9.9±9.4	6.2±13.8	9.2±15.2	14.0±15.2	14.0±12.8

Table 8. COD, T-N prediction relative error of effluent water by learning data period - (Period 2)

Unit : %	COD				T-N			
Learning data	After 1 hr	After 3 hr	After 6 hr	After 12 hr	After 1 hr	After 3 hr	After 6 hr	After 12 hr
3 day	8.6±9.4	11.0±14.7	15.5±33.1	13.6±10.6	7.5±6.6	9.2±9.0	11.3±10.4	12.3±12.8
7 day	20.2±82.3	15.9±36.6	14.6±16.1	27.5±91.2	10.0±9.4	9.9±9.9	11.4±11.4	12.8±12.4
14 day	14.7±27.1	14.0±12.9	14.5±15.6	15.6±16.8	10.3±8.8	10.4±8.6	11.7±11.4	14.5±15.2
21 day	6.5±8.2	8.7±10.0	14.1±18.8	21.0±85.6	8.5±8.6	9.8±7.8	11.2±9.9	12.4±11.0
28 day	5.8±6.7	7.5±7.7	10.2±9.1	12.6±10.5	6.6±6.2	9.4±7.7	11.0±9.3	11.1±9.9

Table 9. COD, T-N prediction relative error of effluent water by learning data period - (Period 3)

Unit : %	COD				T-N			
Learning data	After 1 hr	After 3 hr	After 6 hr	After 12 hr	After 1 hr	After 3 hr	After 6 hr	After 12 hr
3 day	3.7±4.2	5.7±5.8	6.5±6.5	6.7±9.1	3.2±2.9	4.7±3.6	5.6±3.8	5.1±4.7
7 day	2.1±1.5	3.3±2.1	4.8±2.6	6.8±5.7	4.8±4.7	5.0±4.4	5.9±5.1	6.6±6.0
14 day	9.6±5.7	9.6±5.9	8.7±5.8	8.7±5.0	8.8±6.4	9.5±7.0	9.1±6.3	9.3±6.7
21 day	9.3±5.4	9.6±5.3	10.1±5.9	9.8±5.8	10.5±7.4	10.0±6.7	10.3±6.8	11.3±7.1
28 day	10.5±5.9	10.0±6.6	9.9±6.9	10.9±6.0	12.1±8.3	12.2±8.5	13.6±9.8	13.5±9.2

터와 28일간의 학습 데이터를 사용했을 때 예측성능이 우수하였다. 6시간 후 예측 모델의 경우 21일 이상의 학습 데이터를 사용했을 때 오차가 낮게 나타났으나, 5가지 학습 데이터 적용 결과 오차 최소값이 11.0%(D28), 최대값이 11.7%(D14)로 학습 데이터 간 성능의 차이가 크게 나타나지 않았다 (Table 8).

Table 9와 같이 기간 3에서는 기간 1, 2와는 달리 COD, T-N 모두 모든 시간의 예측 모델에서 7일 이하의 단기간 학습 데이터를 사용할 경우 예측성능이 우수하게 나타났으며, 기간 4에서도 기간 3의 결과와 유사하게 대부분의 예측 모델에서 7일 이하의 학습 데이터 사용한 경우 모델 예측오차가 가장 낮았다 (Table 10). T-N의 경우, GRP3-2 그룹의 6시간 및 12시간 후 예측 결과에서만 28일간의 장기데이터 사용한 경우 두 번째로 낮은 오차를 보였다.

한편 기간 5에서는 기간1, 2와 유사하게 T-N, COD 항목 모두 14일 이상의 데이터를 사용하여 학습한 모델이 예측오차가 가장 작게 나타났다 (Table 11).

Fig. 7과 Table 12에서 확인할 수 있듯이 각 분석 기간의 결과를 함께 고려하였을 때 T-N, COD 항목 모두 기간 1, 기간 2, 기간 5에서는 14일 이상의 장기 학습 데이터를 사용하는 것이 더 우수한 예측성능을 보였고, 기간 3, 기간 4에서는 7일 이하의 단기 학습 데이터를 사용한 모델의 예측성능이 더 우수하게 나타났다.

분석 기간의 방류 T-N 측정값의 분산과 표준편차는

Table 10. COD, T-N prediction relative error of effluent water by learning data period - (Period 4)

Unit : %	COD				T-N			
Learning data	After 1 hr	After 3 hr	After 6 hr	After 12 hr	After 1 hr	After 3 hr	After 6 hr	After 12 hr
3 day	3.3±6.8	3.9±13.0	4.2±8.0	5.1±7.9	4.1±3.3	5.1±4.2	6.3±6.2	7.8±6.7
7 day	2.1±1.6	2.8±2.6	3.0±2.5	4.7±6.3	4.6±4.9	6.1±5.1	7.4±5.7	8.2±9.5
14 day	3.2±3.3	3.7±3.8	4.7±4.2	5.4±6.0	5.0±5.2	6.9±6.8	10.3±8.2	10.2±12.4
21 day	5.1±4.2	5.8±4.6	6.7±4.9	8.4±6.1	7.6±5.2	9.6±7.1	9.6±7.8	11.0±13.5
28 day	7.0±5.0	6.2±4.6	6.9±5.1	7.6±5.5	6.3±6.0	6.5±5.9	6.8±5.6	7.3±5.6

Table 11. COD, T-N prediction relative error of effluent water by learning data period - (Period 5)

Unit : %	COD				T-N			
Learning data	After 1 hr	After 3 hr	After 6 hr	After 12 hr	After 1 hr	After 3 hr	After 6 hr	After 12 hr
3 day	3.7±5.9	5.8±25.6	5.6±6.9	5.8±16.6	4.7±5.3	6.3±5.4	9.3±9.4	15.0±18.4
7 day	4.2±6.2	6.3±8.7	6.5±8.9	6.6±8.1	3.9±3.6	8.1±8.8	9.9±10.7	8.3±8.0
14 day	2.2±2.6	3.5±5.4	3.5±3.3	4.2±3.8	4.6±4.9	7.7±7.1	7.7±6.0	7.0±5.8
21 day	2.1±2.3	2.9±3.0	3.5±3.3	3.9±3.4	3.5±3.0	5.5±4.5	7.5±6.3	8.2±7.0
28 day	2.1±2.3	2.7±3.0	3.8±3.4	4.7±3.9	3.8±3.4	5.7±6.2	8.9±8.9	8.8±8.2

※[Orange] The lowest prediction error, [Green] The second lowest prediction error

Table 12. COD and T-N data statistics of TMS effluent by prediction period

Unit : mg/L	T-N					COD				
Classification	Period 1	Period 2	Period 3	Period 4	Period 5	Period 1	Period 2	Period 3	Period 4	Period 5
Minimum	5.9	6.1	9.2	7.6	7.0	5.7	4.7	5.2	3.9	2.9
Maximum	14.2	12.7	14.3	15.0	15.0	17.3	16.7	10.7	6.8	8.2
Dispersion	2.940	1.520	0.825	2.850	3.830	3.740	3.670	0.968	0.255	0.669
Standard deviation	1.710	1.230	0.910	1.690	1.960	1.940	1.920	0.980	0.505	0.818
Average	10.9	10.2	12.4	12.8	12.6	8.3	8.1	6.5	5.4	5.0

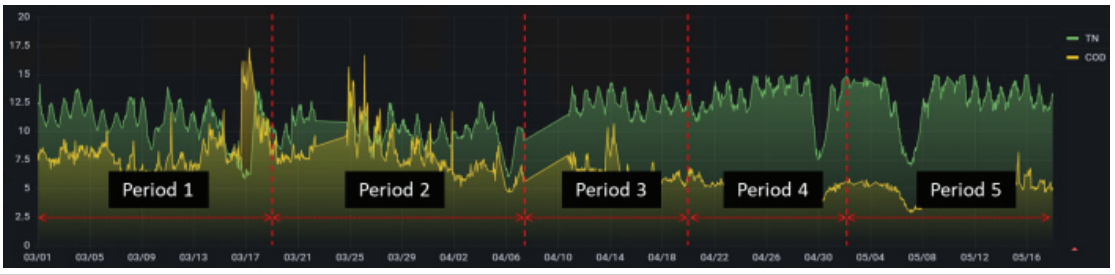


Fig. 7. COD and T-N trend of effluent by analysis period.

기간3에서 크게 감소하였다가 기간 4, 5에서 다시 기간 1, 2와 유사한 수준 또는 그 이상으로 증가하여 데이터 분포의 변화가 발생하였음을 확인할 수 있었다. COD의 경우, 기간 1, 2에서는 COD의 평균값이 8.3,

8.1로 유사하게 나타났으나, 기간 3, 4, 5에서는 6.5, 5.4, 5.0으로 조금 감소하였다. 분산과 표준편차 또한 기간 3 이후부터 기간 1, 2에 비해 크게 감소하였고, 기간 4에서는 기간 3에 비하여 다소 감소한 후 기간 5에서 다시 증가하여 데이터의 분포가 많이 변화되고 있는 것을 알 수 있다.

기간 3은 이전 기간의 데이터 분포와 다른 특징을 가지므로, 7일 이하의 짧은 학습 데이터를 사용한 경우, 데이터 변화를 모델에 잘 반영되어 예측성능이 우수하게 나타난 것으로 판단되며, 기간 4에서 다시 발생한 데이터 분포 변화로 인하여 7일 이하의 단기 학습 데이터를 사용한 모델에서 지속적으로 우수한 예측성능을 나타낸 것으로 판단된다.

T-N의 경우 기간 5에서는 기간 4와 유사한 평균을 가지지만 분산 및 표준편차가 다소 증가하여, 21일 이상의 장기데이터를 사용하여 학습한 모델이 우수한 예측성능을 가지는 것으로 나타났다. COD의 경우, 기간 5에서 기간 3과 기간 4 사이의 데이터 특성을 가지면서 두 기간의 자료를 모두 포함한 데이터를 학습한 경우 우수한 예측성능을 보여주었다.

Fig. 7의 데이터 트랜드에서 기간 1, 기간 2 등에서 측정값의 변동이 크고, 기간 2와 기간 3 초반에 TMS 점검 및 수리 등으로 인한 데이터 누락이 발생한 것으로 보아 이와 같은 데이터 분포의 변화는 측정 장비 운영 시 발생 가능한 측정 오차가 분석에 영향을 미친 것으로 판단된다. 해당 점검 이후 기간 3의 후반 및 기간 4에서는 비교적 안정적으로 수질이 측정된 것으로 나타났으며, 그에 따라 7일 이하의 단기간의 학습데이터에도 우수한 예측성능을 유지했을 것으로 사료된다. 반면, 기간 4 후반, 기간 5 초반 측정값 감소는 해당 기간 발생한 강우의 영향으로 보이며, 그로 인한 데이터 분포 변화의 영향으로 기간 5에서는 예측성능 향상을 위해 기간 3, 4에 비해 필요로 한 학습데이터 기간이 다소 증가한 것으로 판단된다.

이와 같이 분석을 수행한 기간에 따라 입력데이터 분포가 변화하면서 최적으로 선정된 학습데이터 기간이 지속적으로 변화하므로, 학습에 사용되는 데이터 특성에 맞추어 적절한 학습데이터 기간을 자동으로 선정하는 알고리즘을 추가하여 변화하는 데이터 분포를 반영할 수 있다면 예측 모델의 성능을 좀 더 개선할 수 있을 것으로 판단된다.

전체 입력 항목 중 예측 대상과 상관성이 높은 6개의 항목을 선정하여 모델 학습 및 실제 예측을 수행하도록 알고리즘을 구성하였다. 최근 14일간의 데이터를 이용해 학습하고 예측을 수행한 GRP3-2 그룹에 대한 전체 기간 동안 각 입력 항목의 선정 비율은

Table 13. Input item selection rate for GRP 3-2 prediction model

Unit : %	Input items	T-N				COD			
Group		After 1hr	After 3hr	After 6hr	After 12hr	After 1hr	After 3hr	After 6hr	After 12hr
TMS	TMS TN	55	44	21	21	-	-	-	-
	TMS COD	-	-	-	-	57	45	36	22
Group 1	Influent COD	16	12	15	25	23	17	21	21
	Influent EC	38	40	41	26	33	37	35	45
	Influent TSS	37	40	42	40	46	45	39	40
	Influent TN	27	30	38	41	28	34	35	26
	Effluent COD	46	45	30	42	43	48	43	39
	Effluent EC	18	26	33	22	24	25	20	28
	Effluent TSS	61	58	58	63	57	58	62	64
	Effluent TN	26	28	38	42	36	32	38	36
Group 2	Bioreactor DO	26	28	25	21	29	35	42	46
	Bioreactor MLSS	30	38	46	42	36	34	36	44
Group 3	Influent quantity	33	42	47	56	36	35	28	27
	Drawing quantity of sludge	40	39	35	32	41	36	31	34

※ More than 45% of item selection count for prediction model (yellow background)

Table 13과 같다. 생물반응조 유출 TSS 스마트센서 측정 데이터는 모든 T-N, COD 예측 모델에서 45% 이상 입력 항목으로 선정되어 예측 결과에 영향력이 크게 작용하는 것으로 나타났다. T-N, COD 1시간 후 예측 모델에서는 실제 예측 대상인 TMS 최근 측정 데이터가 각각 55, 57%의 높은 비율로 예측에 사용되었다. 이는 단기 예측의 경우, 예측 목표 값이 최근 측정데이터와 비교하여 변동이 크지 않기 때문으로 판단된다. 3시간 후 예측 모델에서도 T-N, COD 두 가지 항목 모두 생물반응조 유출 COD 스마트센서 측정 데이터가 영향력이 높은 항목으로 선택되었다. 그 외, T-N 항목의 경우 6시간 후, 12시간 후 예측 모델에서 유입 유량이 높은 비율로 예측에 사용되어 처리 유량 및 부하 데이터가 예측 결과에 영향력이 높았으며, COD 항목의 경우, 12시간 후 예측 모델에서 생물반응조 DO가 영향력이 높은 입력 변수로 선정되어 예측 성능에 작용하는 영향력이 크게 나타난 것을 확인할 수 있었다.

4. 결 론

본 연구는 하수처리시설 TMS 방류 수질을 예측을 위해 스마트센서 측정 수질 데이터와 기존 SCADA 운영 데이터의 필요성을 확인하였으며, 인공지능 모델을 이용한 미래 수질 예측의 가능성을 확인하였다. AI 기반 방류 수질 예측 모델을 실제 하수처리시설에 적용하여 예측을 수행하고, 예측 모델의 입력 항목, 학습 기간 등에 따른 예측성능을 비교 및 분석하였다.

방류 수질 항목인 COD, T-N, T-P의 1시간, 3시간, 6시간, 12시간 후의 수질 예측을 위해 스마트센서를 사용하여 생물반응조로 유입되는 수질과 최종침전지 유출 수질을 측정하였다. 또한, 생물반응조 유입 유량과 MLSS, DO, 최종침전지 슬러지 인발 유량은 기존 SCADA 시스템과 연계를 통해 확보하였으며, 예측 오차를 최소화 시키기 위해 SCADA 연계 항목은 모든 계열의 평균값을 사용하였다.

스마트센서 자료와 전 계열 생물반응조, 유입유량/슬러지 인발 유량 데이터를 포함한 입력자료 사용 모델을 대상으로 학습 기간을 변경하여 예측을 수행하고 예측성능이 우수한 최적 학습 기간을 도출하였다.

2023년 3월 1일부터 5월 17일까지 약 78일간 예측을 수행하였으며, 총 기간을 5개의 세부 기간으로 나누어 예측 결과를 확인하였다. 기간 1, 2에서는 T-N, COD 항목 모두 대부분 14일 이상의 학습데이터 사용 시 오차가 작게 나타났으나, 기간 3, 4의 경우에는 7일 이하의 단기 학습데이터 사용 시 예측성능이 우수하게 나타났다.

이와 같이 수질 예측 모델의 오차가 학습데이터 기간에 따라 다르게 나타났는데, 이는 예측 항목의 시계열 자료 변동에 따라 영향을 받는 것으로 판단되어 학습데이터 기간을 최적화하는 로직을 적용할 경우, 예측 오차를 줄일 수 있을 것으로 판단된다.

그럼에도 불구하고 학습데이터별 예측성능 비교 결과, T-N의 경우 전체 예측 기간에 대하여 수질 예측 모델의 평균 상대오차가 3.2～15.0%, COD의 경우 기간 2를 제외한 대부분의 기간에 대하여 모든 학습모델의 평균 상대오차가 2.1～12.4%로 나타났으며, 고정된 학습데이터 기간을 사용하여도 신뢰성 있는 예측성능을 확인할 수 있어 향후 인공지능 기반 방류 수질 예측에 대한 가능성을 확인할 수 있었다.

본 연구에서는 AI 기반 방류 수질 예측 모델을 통해 유의한 신뢰도의 예측 수질을 도출할 수 있음을 확인하였다. 이를 활용하여 갑작스러운 유입 부하변동이나 공정 운전이상 등의 원인으로 TMS 수질기준을 초과하는 상황을 사전 감지할 경우, 운영자는 유입 수질 및 공정 운전 상태를 확인하여 원인을 파악하고 그에 따른 적절한 대응을 통해 방류 수질 초과를 사전에 예방할 수 있을 것으로 기대된다.

References

Adnan, R.M., Petroselli, A., Heddam, S., Santos, C.A.G., and Kisi, O. (2021). Comparison of different methodologies for rainfall‐runoff modeling: machine learning vs conceptual approach, Nat. Hazards, 105, 2987-3011.

Ahmed, A.N., Othman, F.B., Afan, H.A., Ibrahim, R.K., Fai, C.M., Hossain, M.S., Ehteram, M. and Elshafie, A. (2019). Machine learning methods for better water quality prediction, J. Hydrol., 578, 124084.

BioWin, https://envirosim.com/products (August 23, 2023).

GPS-X, https://www.hydromantis.com/GPSX-innovative.html (August 23, 2023).

Haghiabi, A.H., Nasrolahi, A.H., and Parsaie, A. (2018). Water quality prediction using machine learning methods, Water Qual. Res. J., 53(1), 3-13.

Hamed, M.M., Khalafallah, M.G., and Hassanien, E.A. (2004). Prediction of wastewater treatment plant performance using artificial neural networks, Environ. Model. Softw., 19(10), 919-928.

Henze, M., Gujer, W., Mino, T., and Van Loosedrecht, M. (2006). Activated sludge models ASM1, ASM2, ASM2d and ASM3, IWA Publishing, London.

Hong, Y.S.T., Rosen, M.R., and Bhamidimarri, R. (2003). Analysis of a municipal wastewater treatment plant using a neural network-based pattern analysis, Water Res., 37(7), 1608-1618.

Jun, H.D. (2021). Developments of a real-time simulation model based on the cyber physical system and a decision support system for management and maintenance for urban water resources, J. Korean Soc. Environ. Eng., 108-108.

Mjalli, F.S., Al-Asheh, S., and Alfadala, H.E. (2007). Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance, J. Environ. Manag., 83(3), 329-338.

Mosavi, A., Ozturk, P., and Chau, K.W. (2018). Flood prediction using machine learning models: Literature review, Water, 10(11), 1536.

MasFlow, http://www.unu-inc.com/massflow (August 23, 2023).

Ribeiro, D., Sanfins, A., and Belo, O. (2013). "Wastewater treatment plant performance prediction with support vector machines, In Advances in Data Mining. Applications and Theoretical Aspects": *13th Industrial Conference, ICDM 2013*, July 16-21, 2013. Springer Berlin Heidelberg. New York, USA.

Usman, S., Kim, J.R., Pak, G.J., Rhee, G.H., and You., K.T. (2022). Investigating Machine Learning Applications for Effective Real-Time Water Quality Parameter Monitoring in Full-Scale Wastewater Treatment Plants, Water 14, 19: 3147. Switzerland.

Water World, https://www.waterworld.com/home/article/16202870/spanish-wastewater-smart-plant-cuts-energy-sludge-and-chemical-use (August 23, 2023).

Yun, Z.W. Water Journal, http://www.waterjournal.co.kr/news/articleView.html?idxno=40300 (August 21, 2023).

You, K.T. KEITI (2020). Development of optimal and smart energy management solution based on IoT for wastewater treatment plant, 2019002210001.

You, K.T. Ministry of Environment. (2021). Machine learning-based water treatment process diagnosis and integrated operation system development final report, 2018002110001.