Journal of the Korean Institute of Industrial Engineers
Vol. 49, No. 1, pp. 95-106, February 2023.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2023.49.1.095

하수처리시설 수질인자의 머신러닝 예측 모델

주형구 · 임준묵†

한밭대학교 창의융합학과

Machine Learning Prediction Model of Water Quality Factors in Sewage Treatment Facilities

Hyeong-gu Joo · Joon-mook Lim

Department of Creative Convergence Eng., Hanbat National University

The purpose of this study is to develop a machine learning-based prediction model for the value of COD, a key factor that measures the quality of sewage flowing from a sewage treatment facility. Considering that the inflowing sewage water quality data has a time-series characteristic, a machine learning model using ARIMAX, RNN, and LSTM was developed as a predictive model for COD. For each model, after learning based on big data collected from domestic J sewage treatment facility, the prediction performance was evaluated by RMSE. ARIMAX model showed an accuracy of 7.83% compared to the average, RNN was 3.62%, and LSTM was 3.56%. Overall, the LSTM model was evaluated as the best performing predictive model. If our model is used in a sewage treatment facility, it is possible to predict the sewage water quality in real time, which is expected to greatly contribute to improving the efficiency of sewage treatment.

Keywords: Sewage Treatment Facilities, Water Quality Factors, Prediction Model, Machine Learning, Bigdata, Deep Learning

1. 서 론

1.1 연구의 배경

UN은 2019년에 한국을 '물 스트레스(water-stressed)' 국가로 분류했다(OECD, 2012). 물 스트레스 국가란 매년 1인당 이용 가능한 수자원량의 기준이 1000~1700㎥인 국가를 의미한다. 한국의 강수량은 비교적 풍부한 수준이지만 높은 인구밀도와 산악지형, 이상기후의 발생이 물 부족의 원인으로 뽑힌다. 물을 효율적으로 사용하는 방법 중 하나는 하수 처리수를 재이용 하는 것이다. 하수 처리장에서 효율적인 공정을 위해 많은 연구들이 지속적으로 진행되어 왔다. 과거의 하수처리 공정은 운전자의 경험과 지식에 의존하는 공정으로, 공정의 객관성과 정확성을 보장받기 힘들다는 문제점이 있다. 이를 해결하기 위해서는 데이터에 기반한 모델을 활용한 객관적인 공정이 필요하다. 기존의 모델들은 공정의 성능을 예측하는 목적으로 개발된 경우가 대부분이기 때문에 공정 상태의 이상을 미리 감지하거나 사전에 대응하기 위한 목적으로 활용하는 것에는 어려움이 있다.

따라서, 하수처리 공정에서의 효율적인 처리를 위해서는 향후 유입수의 정보를 예측하여 발생할 수 있는 문제를 미리 감지하고 조치를 취하는 사전 예방전략이 요구된다. 본 연구에서는 실시간 계측이 어려운 화학적 산소 요구량(Chemical Oxygen Demand, 이하 COD)의 값을 빅데이터와 머신러닝기법을 활용하여 예측할 수 있는 모델을 개발하고자 한다.

1.2 연구의 필요성

우리나라에는 400여 개의 크고 작은 규모의 하수 종말 처리장이 있다. 하수처리장에 유입되는 하수의 정수처리 절차는

이 논문은 한밭대학교 교내학술연구비의 지원을 받아 수행되었음.
† 연락저자 : 임준묵 교수, 대전시 유성구 동서대로 125 한밭대학교 창의융합학과, Tel : 042-821-1972, E-mail: jmlim@hanbat.ac.kr
2022년 10월 19일 접수; 2022년 12월 4일 수정본 접수; 2022년 12월 12일 게재 확정.

일반적으로 [유입하수→(수질측정1)→여과→약품투입, 침전→(수질측정2)→방류]의 과정을 거친다. 오염된 하수가 하수종말 처리장에 유입되면 수질을 측정(수질측정1)하고, 여과한 후 수질의 정도에 따라 적절한 약품을 투입하고, 침전 등의 조치를 취하며, 방류 전 수질을 다시 한 번 측정(수질측정2)하고 방류하는 것이 하수를 처리하는 기본 절차이다. 일반적으로 하수의 수질은 수소이온농도(pH), 탁도(SS), 총인(TP), 총질소(TN), 화학적산소요구량(COD), 생화학적산소요구량(BOD) 등의 6가지 요소를 측정하여 파악한다. 따라서, 하수종말처리장에 유입되는 하수의 수질 상태를 실시간으로 신속하게 정확히 측정할 수 있어야 효과적인 하수처리가 가능하다. pH 및 SS는 수질센서를 활용하여 실시간 측정이 가능하지만, 현재까지 국내·외 기술로는 나머지 수질 측정 요소들의 실시간 계측은 불가능하다(KEITI, 2018).

수자원공사에서는 수질 오염의 정도를 파악하기 위해서 각 단계에서 시료를 채취한 후, 검사키트를 활용하여 수질검사를 수행하는데, 그 값을 알기까지 상당한 시간(4～8시간)을 요한다(NIER, 2015). 결국, 정체불명의 하수가 유입될 경우, 방류 뒤에나 오염수의 정체를 파악할 수밖에 없는 실정이다.

현재까지 알려진 국내 상용화된 실시간 측정 가능한 하수수질계측기는 없다. 외국산 장비(일본)의 경우, TN(총질소) 한 항목만을 측정하는 장비도 5,000～8,000만 원의 예산이 소요되며, 수자원공사의 현장 설치만 가능하고 신뢰도는 높지 않은 실정이다(NIER, 2017). 최근 국내기술로 pH, SS, EC(전기전도도), 온도 등을 실시간으로 계측할 수 있는 수질측정센서가 개발되어 저렴한 비용으로 매우 손쉽게 자료획득이 가능하다. 하지만 여전히 핵심 수질인자인 COD, BOD는 실시간 측정이 불가능하다(WaterEyes Co., Ltd., 2020).

기존의 연구에서는 실시간 계측이 가능한 (pH, SS, EC, 온도)인자와 나머지 수질인자(TN, TP, COD, BOD)간의 상관관계를 고려하여 TN, TP, COD 등의 주요인자를 예측할 수 있는 모델의 개발을 시도하였으나, 단순 선형관계만을 고려한 연구결과로, 인자들 간의 복잡한 비선형관계로 인해서 발생하는 큰 오차로 인해 정확한 예측에는 한계가 있었다.

머신러닝(딥러닝)을 활용한 모델은 이러한 방대한 수질데이터에 대해서 비선형적인 문제를 효과적으로 해결할 수 있는 것으로 알려져 있다. 그러므로 최근 한계에 다다른 하수처리 문제를 해결하기 위해서, 복잡한 비선형 관계를 가지는 수질데이터를 고려하여 하수의 수질을 정확하게 예측하고 실시간 측정이 힘든 데이터의 예측결과를 제공할 수 있는 인공신경망 딥러닝 예측모델의 개발이 필요한 실정이다.

2. 기존의 하수 수질 예측 연구 및 한계점

수질인자들간의 상관관계로부터 수질 인자들을 예측하기 위한 연구는 지속적으로 이루어져 왔다. 하지만, 기존의 연구결과들은 다음의 두 가지 측면에서의 한계점을 가지고 있어, 하수 수질인자의 실시간 예측에 직접 활용하기에는 어렵다.

첫째, 그동안 하수처리시설에서의 COD 등을 추정하기 위한 여러 가지 시도가 있었으나, 수질인자들간에는 비선형적으로 복잡하게 얽혀 있음에도 불구하고, 선형관계만을 가정하여 예측모델을 구성하고 있어 예측의 정확도가 낮아 활용성이 떨어진다(Kim and Park, 1982; Seo *et al.*, 2013; Cho *et al.*, 2014). 또한 그 밖의 연구는 하수처리시설이 아닌 강유역, 저수지 등에서의 추정연구로 하수처리시설에 직접 활용하는데 어려움이 있으며, 역시 선형적인 가정을 하고 있다(Park *et al.*, 2014; Beom *et al.*, 2019).

둘째, 이러한 비선형성을 극복하기 위해서 최근 기계학습 모델을 적용한 국내외적인 연구가 이루어지고 있으나, 여전히 대부분의 연구에서 하수처리시설의 특성 고려없이 일반적인 수질 자료를 활용하고 있고, 실시간 측정이 불가능한 인자들을 예측을 위한 독립변수로 사용하고 있어서, 제시한 모델을 하수처리시설의 수질값(COD) 예측에 직접적으로 활용하는 데는 한계가 있다(Jung *et al.*, 2018; Lim *et al.*, 2021; Zifei *et al.*, 2019).

본 연구에서는 이러한 한계점을 극복하기 위해서 비선형성을 극복할 수 있는 딥러닝 모델을 적용하였으며, 실시간 측정이 어렵지만, 핵심 수질인자인 COD를 예측변수(종속변수)로 하여 미래의 COD를 예측할 수 있는 모델을 제안하였다. 단, BOD 역시 실시간 예측이 필요한 주요 수질인자이지만 현장으로부터 학습데이터의 획득에 한계가 있어서 본 연구에서는 COD만을 예측대상으로 하였다.

3. 예측을 위한 알고리즘

본 연구에서는 비선형적이며 시계열적인 특성을 가지는 하수 수질데이터의 예측에 적합한 알고리즘으로 자기회귀누적이동평균(Autoregressive Integrated Moving Average, 이하ARIMA)에 외생변수 X를 포함할 수 있도록 변형한 ARIMAX(Autoregressive Integrated Moving Average Exogenous) 모델과 딥러닝을 활용한 알고리즘으로 순환신경망(Recurrent Neural Network, 이하 RNN)과 LSTM(Long Short Term Memory, 이하 LSTM) 등을 사용하였다. 알고리즘의 특성을 간단히 요약하면 다음과 같다.

3.1 외생변수를 포함한 자기회귀누적이동평균(ARIMAX)

ARIMA는 시계열 데이터 기반 분석 기법으로 자신의 과거 정보를 활용하는 자기 회귀(AR) 모델과 과거 오류정보를 활용하는 이동 평균(MA) 모델을 결합한 ARMA에 추세까지 고려한 모델이다(Park and Jun, 1984). AR, MA, ARMA 모델이 시계열 데이터가 정상성을 만족한다는 가정의 상황에서 분석을 진행했다면, ARIMA 모델에서는 비정상성인 경향을 가지는 데이터도 차분(differencing)을 통해 예측이 가능하다는 특징이

있다(Hyndman and Athanasopoulos, 2018). ARIMA 모델은 보통 ARIMA(p,d,q)로 표시되며 p는 AR 모델의 차수, d는 차분한 횟수, q는 MA 모델의 순서를 뜻한다. ARIMAX는 AR 모델과 MA 모델을 동시에 포함하는 모델로, 일반적인 ARIMA 모델은 단변량 시계열을 표현하는데 적절한 모형이지만 ARIMAX 모델은 추가적인 외생변수(Explanatory variable)를 포함함으로써 다변량 시계열 데이터를 활용하기에 적절한 모형이다. ARIMAX는 다양한 외생변수의 조건을 고려하여 종속변수와의 관계성을 반영하여 모형을 개선하는 특징이 있고, 다수 입력변수들의 결합적인 영향력을 고려하므로 뛰어난 장기 예측력을 가지고 있다(Suhermi *et al.*, 2019).

3.2 순환신경망(RNN)

인공 신경망의 한 종류인 RNN은 출력으로 나온 output값이 다시 input으로 들어가는 순환구조를 가지고 있다(Bengio *et al.*, 2013). 이러한 구조를 통해 신경망 내부에 현재 상태를 저장할 수 있고, 일반적인 순방향 신경망과 달리 내부의 메모리를 활용하여 시퀀스 형태의 입력값을 처리 할 수 있다. RNN은 과거 상태를 반영하여 가중치를 갱신할 수 있으며 시계열 데이터 처리에 특화된 모델이다. 이때 일어나는 계산은 식 (1)과 같다(Telab, 2018).

$$f_t = \delta(W_t x_t + U_f h_{t-1} - b_f) \quad (1)$$
$$s_t = f(Ux_t + Ws_{t-1})$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

여기서, x_t, s_t는 각각 시점 t에서의 입력값, σ는 활성화 함수

학습 데이터의 길이가 길어지면 순환 신경망이 가중치를 업데이트 할 때 기울기 소멸(vanishing gradient)이 발생하여 학습이 제대로 되지 않는 단점이 있다. 이를 장기 의존성(Long-Term Dependency)이라 한다.

3.3 LSTM(Long Short-Term Memory)

LSTM은 RNN에서 발생하는 장기 의존성 문제를 해결하기 위해 고안된 딥러닝 모델이다(Hochreiter and Schmidhuber, 1997). LSTM의 구조는 <Figure 2>와 같다.

LSTM은 입력 게이트(Input Gate), 출력 게이트(Output gate), 망각 게이트(Forget Gate)가 존재한다. 입력 게이트는 입력받은 정보를 얼마나 반영할지 결정하는 게이트이다. 입력 게이트의 출력값은 활성화 함수가 sigmoid 일 때 식 (2), tanh 일 때 식 (3)에 의해 결정된다(Olah, 2015).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

출력 게이트는 최종적으로 얻어진 상태값에서 얼마나 빼낼지 결정하는 게이트이다. 활성화 함수가 sigmoid 일 때 식 (4), tanh 일 때 식 (5)에 의해 결정된다.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o] \quad (4)$$

$$h_t = o_t * \tanh(C_t) \quad (5)$$

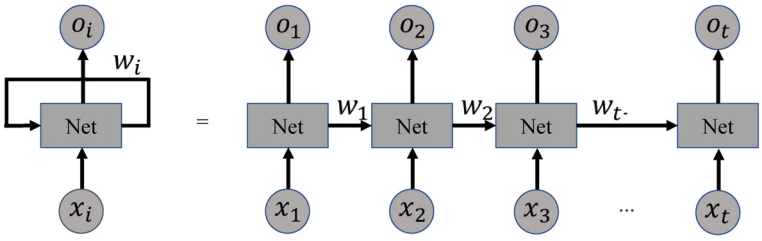


Figure 1. Structure of RNN (Lei, 2016)

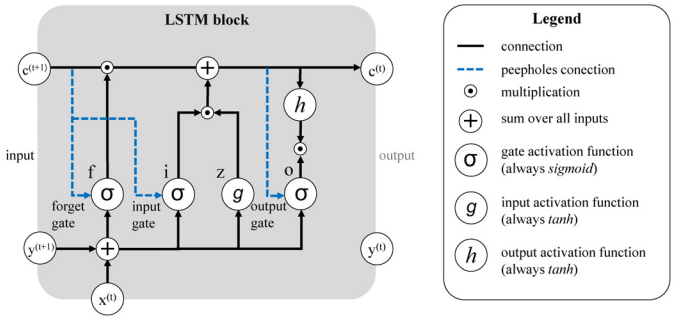


Figure 2. Structure of LSTM (Houdt *et al.*, 2020)

4. 수질 빅데이터의 전처리 및 평가지표

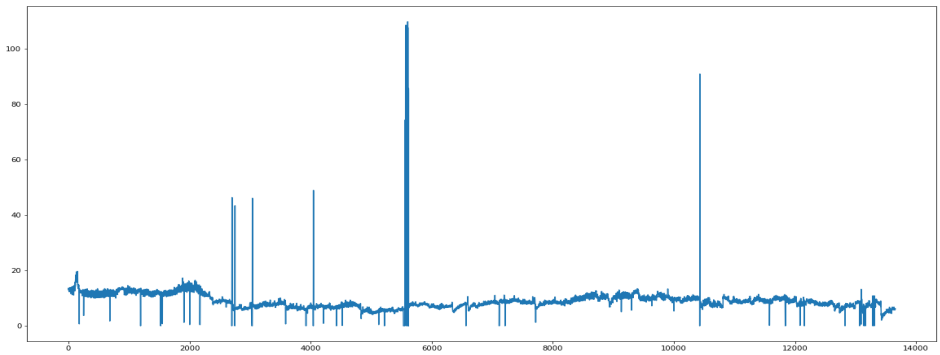
본 절에서는 수집한 데이터의 예측 알고리즘의 적용, 머신러닝과 분석에 앞서 실행하는 전처리 과정, 모델의 학습을 위해 전체 데이터를 훈련데이터와 테스트 데이터로 구분하는 절차, 그리고 모델의 학습결과를 평가하는 평가지표에 대해 설명한다.

4.1 학습데이터의 수집 및 이상치 제거

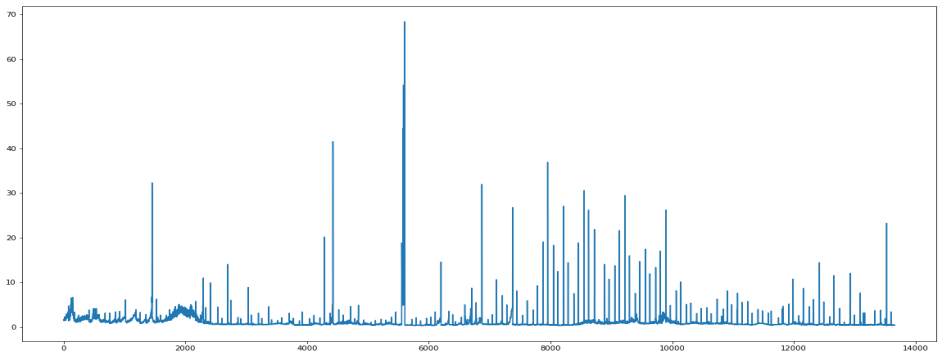
본 연구에서는 국내 J 하수처리장에서 수집한 수질 데이터를 활용하여 향후 24시간의 COD값을 예측하고자 한다. 수집된 수질관련 빅데이터에서 본 연구에서 예측을 위한 데이터로 사용한 자료는 2019년 1월부터 18개월간 기록된 수질관련 측정 자료이다. 데이터는 1시간 단위로 하루에 24회 기록되어 있다. 수집된 데이터는 수소이온농도(pH), 탁도(SS), 총인(TP), 총질소(TN), 폭기조내혼합오니농도(MLSS), 용존산소(DO), 유입유량(incoming flow), 화학적산소요구량(COD)으로 총 8가지이다. 수집된 데이터의 예는 <Table 1>과 같으며 변수별로 각각 약 13,140개의 수치자료로 구성되어 있다.

Table 1. Dataset Sample

M-D H:M	pH	SS (mg/L)	TN (mg/L)	TP (mg/L)	MLSS (mg/L)	DO (mg/L)	incoming flow	COD (mg/L)
01-01 0:04	6.41	1.73	13.41	0.13	2595.63	1.72	150	13.4
01-01 1:04	6.40	1.61	13.40	0.14	2532.5	2.05	50	13.6
01-01 2:04	6.40	1.58	13.15	0.14	619.38	0.04	162	13.2
01-01 3:04	6.43	1.57	13.12	0.13	1895.25	0.04	29	13.3
01-01 4:04	6.40	1.46	12.69	0.13	2608.75	1.78	144	13.2
01-01 5:04	6.40	1.41	12.24	0.13	2514.38	2.66	30	13.3
01-01 6:04	6.33	1.55	11.85	0.12	1442.5	0.04	89	12.8
01-01 7:04	6.30	1.46	11.32	0.12	107.5	0.04	68	13.2
01-01 8:04	6.30	1.53	11.34	0.13	2508.75	1.98	136	13.0
01-01 9:04	6.33	1.47	11.04	0.14	2314.38	2.72	128	12.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

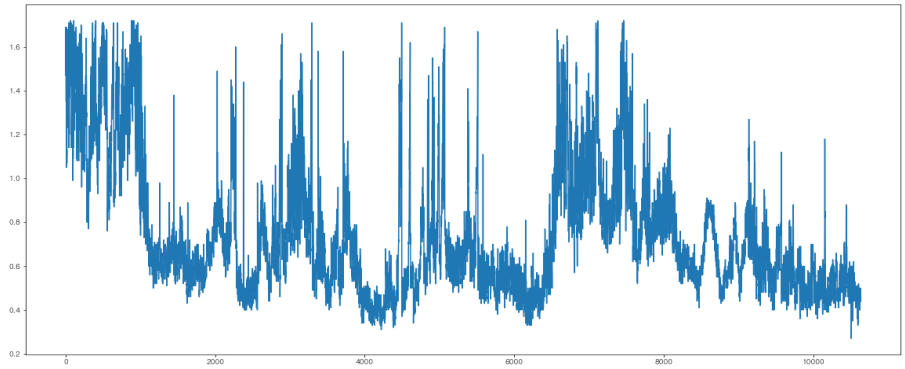


<SS>

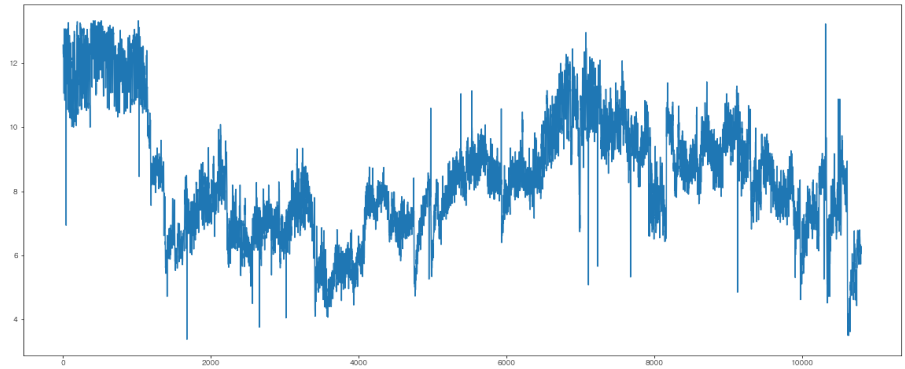


<COD>

Figure 3. SS and COD Data with Outliers



<SS>



<COD>

Figure 4. SS and COD Data with Outliers Removed

확보한 데이터에는 계측기의 오류와 같은 이상치(outlier)가 존재할 수 있고, 이는 과대적합(overfitting)과 같은 문제를 발생시켜 데이터의 신뢰성을 떨어뜨린다. 이상치를 제거하기 위한 기법으로 IQR 탐색 기법을 사용하였다.

$$IQR = Q3 - Q1 \tag{6}$$
$$MINIMUM = Q1 - 1.5 * IQR$$
$$MAXIMUM = Q3 + 1.5 * IQR$$

본 연구에서는 식 (6)에서 MINIMUM보다 작거나 MAXIMUM보다 큰 값을 이상치로 설정하고 제거하였다. 이상치를 제거하기 전의 SS와 COD 데이터의 예를 그래프로 나타내면 <Figure 3>과 같다.

또한 이상치를 제거한 후의 SS와 COD 데이터의 예를 그래프로 나타내면 <Figure 4>와 같다.

4.2 입력 변수의 선택

선행 연구 검토 결과, 다양한 연구들에서 COD 예측을 위해 입력 변수로 기상 데이터, 유입유량, SS, TN, TP, BOD, COD 등을 사용한 것을 알 수 있었다. 또한 입력 변수의 선정법으로는 주로 예측하고자 하는 변수인 COD와 높은 상관관계를 가지는 변수들을 선택한 것을 확인할 수 있었다(Moon, 2008; Beom *et al.*, 2019; Lim *et al.*, 2021; Zifei *et al.*, 2019). 본 연구에서는 이를 참고하여 사용할 예측 알고리즘의 입력변수를 설정하기 위해서 예측하고자 하는 수질인자 변수인 COD와 나머지 변수들간의 Pearson 상관계수를 계산하여 유의미한 상관관계를 가지는 변수들을 고려하였다.

<Table 2>로부터 COD와 유의미한 상관관계를 보이는 변수인 pH, SS, TN, TP와 유입유량 및 COD를 입력변수로 결정하였고, MLSS와 DO는 상관계수가 너무 낮아 입력변수에서 제외하였다.

4.3 데이터 정규화

이상치를 제거했지만, 각 변수들의 데이터 값의 범위가 서로 차이가 심하기 때문에 이를 동등한 범위로 조정해주는 과정이 필요하다. 본 연구에서는 데이터 값에서 평균을 빼고 표준편차로 나누는 식 (7)과 같은 표준화 과정을 적용하였다.

$$Z = \frac{X - X_{\min}}{X_{std}} \tag{7}$$

Table 2. Pearson Correlation Coefficient between COD and Other Variables ($p<0.05$)

Identifier	pH	SS	TN	TP	MLSS	DO
COD	-0.25	0.37	0.56	0.18	0.06	0.05

4.4 데이터 분할과 평가지표

예측 알고리즘과 머신러닝 모델에 데이터를 훈련시킨 후, 모델의 성능을 검증하기 위해 전체 데이터셋을 훈련 데이터와 테스트 데이터로 나누는 과정이 필요하다. 본 연구에서는 전체 데이터셋의 임의의 80% 데이터를 훈련데이터로, 나머지 20%를 테스트 데이터로 분할하고 모델 성능평가에 사용하였다.

본 연구에서는 모델의 성능을 평가하기 위한 지표로 모델이 예측한 값과 실제 데이터와의 오차의 차이를 제곱한 값의 평균에 제곱근을 적용한 평균제곱근오차(Root Mean Square Error; RMSE)와 오차와의 차이에 대한 값을 예측값으로 나눈 절대값의 평균인 MAPE(Mean Absolute Percentage Error)를 사용하였다.

$$RMSE = \sqrt{mean((Y_{predict} - Y)^2)} \tag{8}$$

$$MAPE = mean(\left|\frac{Y_{predict} - Y}{Y_{predict}}\right|) \tag{9}$$

일반적으로 RMSE와 MAPE가 낮은 모델이 오차가 적어 우수한 모델이라고 판단할 수 있다.

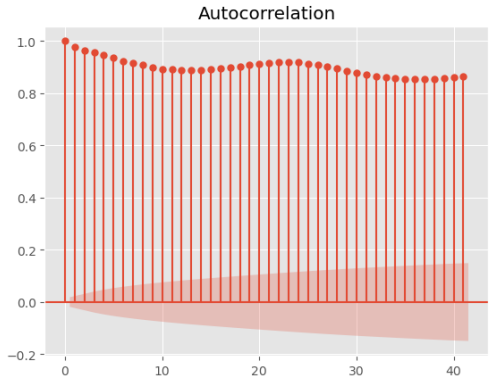


Figure 6. ACF Graphs

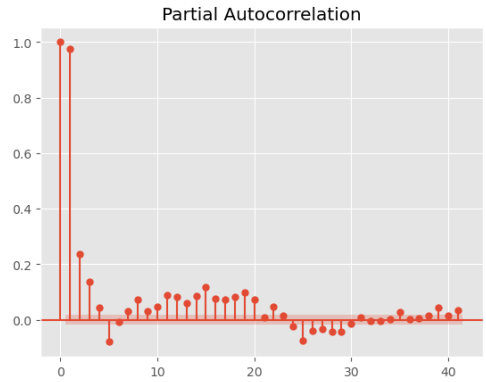


Figure 7. PACF Graph

5. 하수처리시설 수질 예측 모델의 구성 및 예측

5.1 ARIMAX

ARIMAX는 시계열 데이터의 예측에 사용하는 다변량 예측 모델로 본 연구에서는 pH, SS, TN, TP와 유입유량 및 COD를 입력변수로 하여 미래의 COD를 예측하게 된다. COD를 시계열 분해하여 Trend, Seasonal, 및 Residual로 나누어 살펴보면 <Figure 5>와 같다.

우선 ARIMAX모델을 사용하기 위해서는 데이터의 정상성(stationarity)을 확인해야 하는데, 자기상관함수(AutoCorrelation Function; ACF)의 그래프를 그려보면 알 수 있다.

<Figure 6>으로부터 ACF가 시간에 따라 매우 느리게 감소하는 현상을 확인하고 COD 데이터가 정상성을 보이지 않음을 고려하여 차분을 진행하였다.

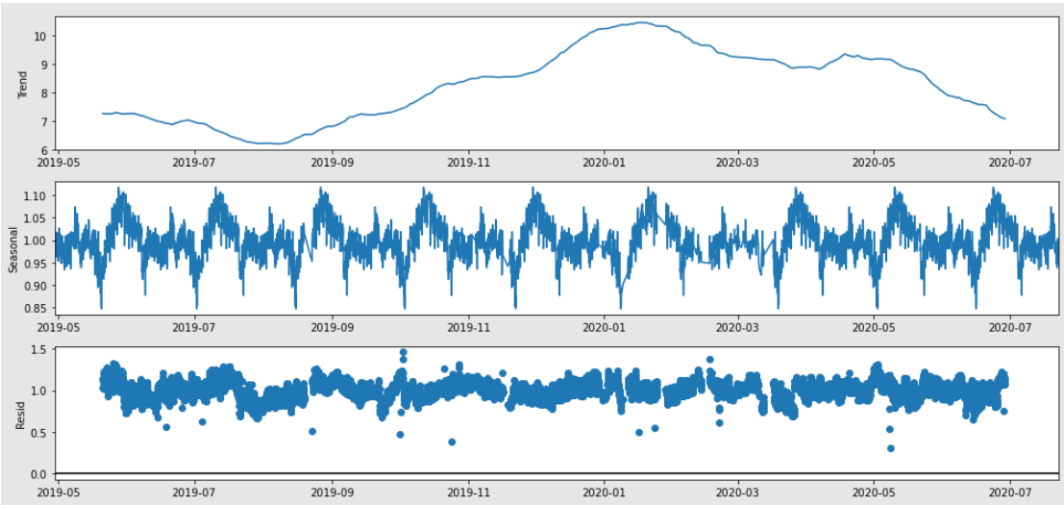


Figure 5. Time Series Decomposed COD

Table 3. ARIMAX Model Summary

Identifier	coef	std err	z	P> | z |
ar.1	1.914	0.003	649.525	0.000
ar.2	-0.983	0.003	-326.23	0.000
ma.1	-2.274	0.011	-210.59	0.000
ma.2	1.5188	0.026	60.106	0.000
ma.3	-0.0639	0.026	-2.504	0.012
ma.4	-0.1582	0.011	-14.502	0.000

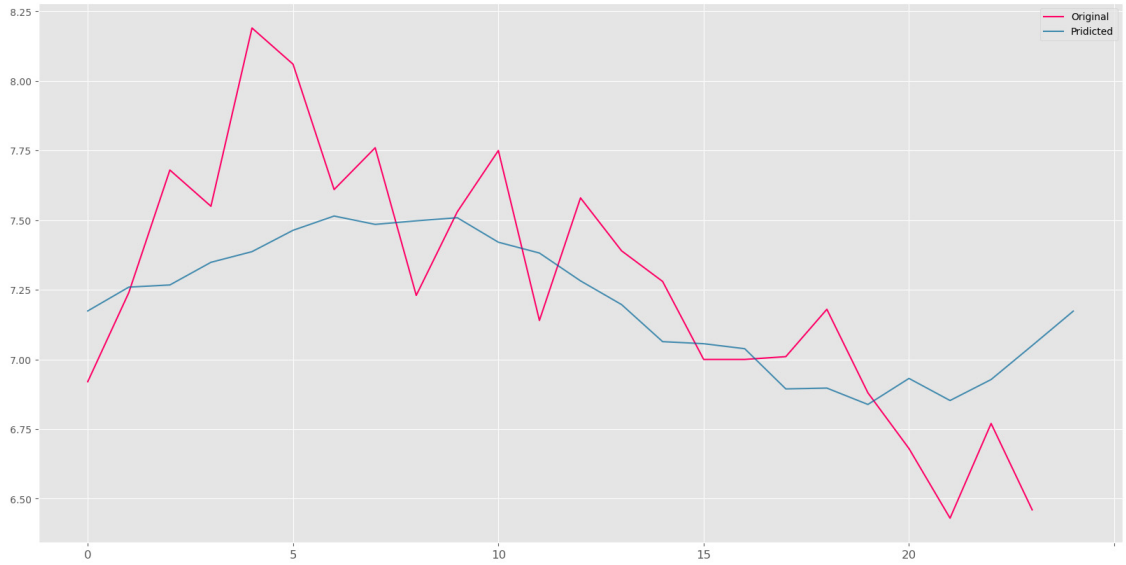


Figure 8. Prediction of COD in ARIMAX Model

또한 <Figure 7>에서 보는 바와 같이 부분자기상관함수(Partial ACF; PACF) 그래프의 LAG가 4번째 이후에 0으로 수렴함을 확인하고, 반복적인 아카이케(Akaike)의 정보기준(AIC; Akaike's information Criterion)을 확인하여 최적의 (p,d,q)의 값을 (2,1,4)로 확정한 후, 외생변수 X(pH, SS, TN, TP, 유입유량)를 포함하는 ARIMAX 모델로 분석을 진행하였다. ARIMAX 모델의 실험결과 요약은 <Table 3>과 같다.

구축한 모델로부터 예측한 향후 24시간의 COD값을 <Figure 8>과 같이 그래프로 나타내고 이를 실제 COD 값과 비교하였다.

<Figure 8>의 그래프에서 꺾은선이 실제값을 의미하고 부드러운 굵은선이 ARIMAX에 의한 예측값을 나타낸다. ARIMAX가 예측한 COD값은 전체적으로 실제 COD값의 평균과 비슷한 양상을 보이며, 시간에 따른 실제 COD값의 증감을 따라가는 경향을 보여준다. 하지만 이러한 성질로 인해서, ARIMAX모델을 사용하여 변화가 심한 COD의 구체적인 값을 예측하는 것에는 한계가 있음을 확인하였다.

ARIMAX 모델의 평균 학습시간은 431초이고, 평균 예측시간은 0.01s이며, RMSE는 0.6622, MAPE는 3.7201이다. RMSE는 COD의 평균인 8.45와 비교했을 때 약 7.83%에 해당하는 수치이다.

5.2 RNN 모델

본 연구에서 COD 예측을 위해 구축한 RNN 모델은 2개의 은닉층으로 설계하였으며 각 층은 64, 128, 256, 512, 1024개의 노드 개수 중 하나로 설정하고 실험을 진행하였다. 예측에 사용할 과거 데이터 개수는 과거 15일 데이터로 설정하였다. 각 은닉층 사이에는 활성화 함수로 일반적으로 RNN 모델에서 사용하는 tanh를 사용하였고 optimizer로는 Adam을 사용하였다. 모델의 학습 횟수인 epoch는 과적합을 방지하기 위해 30회로 하였고 batch size는 64로 설정하였다.

<Table 4>는 예측 모델의 은닉층 노드 수에 따른 학습과 예측실험의 결과를 보여준다. 과거 15일의 데이터를 사용한 예측 모델의 출력값에 따른 RMSE와 MAPE값을 은닉층의 노드 수에 따라 통계처리한 결과를 보여준다. <Table 4>에서 괄호 속의 값은 RMSE의 COD 평균대비 비율(%)을 나타낸다.

실험 결과 은닉층 노드수(256, 1024)의 모델을 사용하여 과거 15일의 데이터로 예측했을 때, 모델의 평균 학습시간은 661초, 예측시간은 0.04초, RMSE 값은 0.3063로 가장 낮게 나타났다. RMSE는 COD의 평균대비 약 3.62%에 해당하는 수치이다. 또한 MAPE도 은닉층 노드수(256, 1024)에서 1.9982로 가장 낮은 값을

보여주어, RMSE와 같은 결과를 보임을 확인할 수 있었다.

은닉층(256, 1024)을 사용한 RNN 모델의 예측결과에 대한 그래프 예는 <Figure 9>, <Figure 10>, <Figure 11>과 같다.

Figure 9, Figure 10, Figure 11 에서 후반부의 굵은 실선이 RNN 모델에 의한 COD 예측값을 나타내고, 가는 실선은 실제 COD값을 나타낸다. RNN 모델은 데이터의 증감추세 뿐만 아니라 실제 데이터에도 상당히 근사한 예측을 하는 것을 확인할 수 있다.

Table 4. RMSE Value, Average Ratio and MAPE Predicted by Data from the Past 15 Days of the RNN Model

Identifier			hidden layer I				
			64	128	256	512	1024
hidden layer II	64	RMSE	0.4133	0.4699	0.3672	0.6739	0.8512
			(4.89%)	(5.56%)	(4.35%)	(7.98%)	(10.07%)
		MAPE	2.1137	2.1109	2.0055	3.7209	4.2764
	128	RMSE	0.7391	0.4903	0.4119	0.5573	0.4557
			(8.75%)	(5.80%)	(4.87%)	(6.60%)	(5.39%)
		MAPE	4.0101	2.1991	2.1305	2.2133	2.1365
	256	RMSE	0.7066	0.4356	0.7409	0.3294	0.7257
			(8.36%)	(5.16%)	(8.77%)	(3.90%)	(8.59%)
		MAPE	3.9841	2.1329	4.0163	2.0006	4.0081
	512	RMSE	0.4699	0.4986	0.8887	0.3825	0.8174
			(5.56%)	(5.90%)	(10.52%)	(4.53%)	(9.67%)
		MAPE	2.1109	2.2188	4.2899	2.0146	4.2542
	1024	RMSE	0.4458	0.3488	**0.3063**	0.6379	0.7246
			(5.28%)	(4.13%)	**(3.62%)**	(7.98%)	(8.58%)
		MAPE	2.1350	2.0031	**1.9982**	3.7168	4.0079

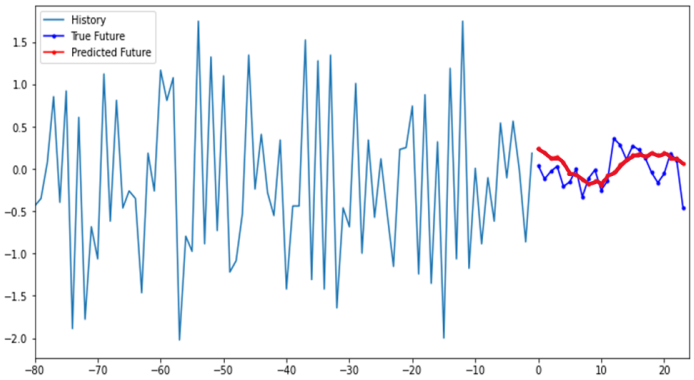


Figure 9. Prediction Example 1 of the RNN Model

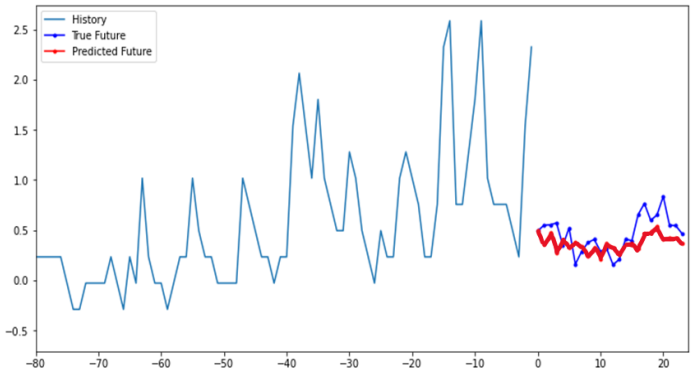


Figure 10. Prediction Example 2 of the RNN Model

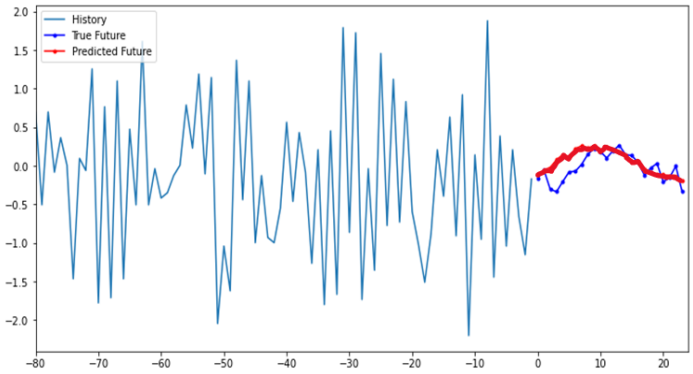


Figure 11. Prediction Example 3 of the RNN Model

5.3 LSTM 모델

LSTM 모델에서는 RNN 모델과 마찬가지로 훈련데이터의 COD와 상관관계가 가장 높았던 pH, SS, TN, TP와 유입유량, COD값을 입력값으로, 미래의 COD를 출력값으로 하는 모델을 구성하였다. 은닉층의 구성 및 기타 설정값은 RNN 모델과 동일한 값으로 하였으며 같은 환경에서 실험을 진행하였다.

<Table 5>는 과거 15일의 데이터를 사용한 예측 모델의 출력값에 따른 RMSE와 MAPE값을 보여주며, 괄호 속의 값은 COD 평균대비 RMSE의 비율(%)을 나타낸다.

Table 5. RMSE Value, Average Ratio and MAPE Predicted by Data from the Past 15 Days of the LSTM Model

Identifier			hidden layer I				
			64	128	245	512	1024
hidden layer II	64	RMSE	0.4946	0.7478	0.3768	0.3727	0.4863
			5.85%	8.85%	4.46%	4.41%	5.54%
		MAPE	2.2196	4.0261	2.0049	2.0041	2.2175
	128	RMSE	0.4073	**0.3004**	0.3897	0.3638	0.5240
			4.82%	**3.56%**	4.61%	4.31%	6.20%
		MAPE	2.0174	**1.9927**	2.0138	2.0046	2.2462
	256	RMSE	0.4546	0.3164	0.5545	0.4424	0.4307
			5.38%	3.74%	6.56%	5.24%	5.10%
		MAPE	2.1587	1.9998	2.2684	2.1335	2.1316
	512	RMSE	0.3486	0.6060	0.3495	0.4344	0.7663
			4.13%	7.17%	4.14%	5.14%	9.07%
		MAPE	2.0030	2.3286	2.0032	2.1323	3.4260
	1024	RMSE	0.5574	0.5352	0.6092	0.4770	0.5589
			6.60%	6.53%	7.21%	5.64%	6.61%
		MAPE	2.2686	2.2689	2.3290	2.2072	2.2687

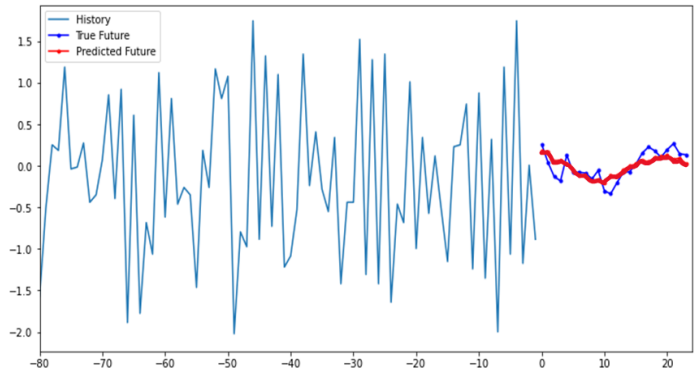


Figure 12. Prediction Example 1 of the LSTM Model

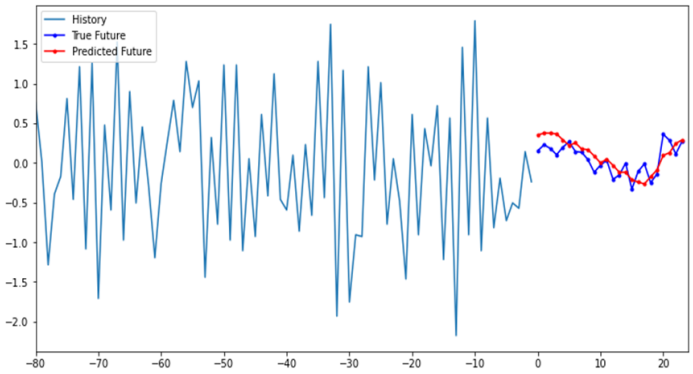


Figure 13. Prediction Example 2 of the LSTM Model

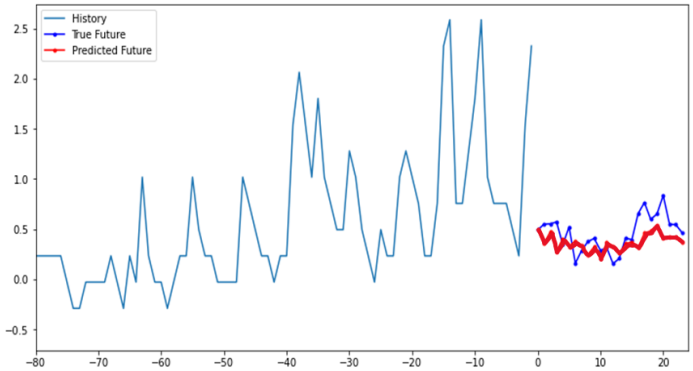


Figure 14. Prediction Example 3 of the LSTM Model

실험 결과 은닉층(128,128)의 모델을 사용하여 과거 15일의 데이터로 예측했을 때, 모델의 평균 학습시간은 794초이고 예측시간은 0.04초, RMSE 값은 0.3004로 가장 낮게 나타났다. RMSE는 COD의 평균대비 약 3.56%에 해당하는 수치이다. 또한 MAPE의 경우도 은닉층(128,128)에서 1.9927로 가장 작은 값을 나타냈다.

과거 15일의 데이터를 사용한 은닉층(128,128) LSTM 모델의 예측 그래프 예는 <Figure 12>, <Figure 13>, <Figure 14>와 같다.

LSTM 모델도 RNN 모델과 비슷하게 데이터의 경향뿐만 아니라 실제 데이터에 가까운 예측을 하는 것을 확인할 수 있다. 또한 RNN모델에 비해서 근소한 차이지만 보다 더 낮은 RMSE 및 MAPE값을 보여주는 우수한 예측을 수행할 수 있음을 알 수 있다.

6. 실험 결과의 분석 및 평가

본 연구에서는 핵심 하수 수질인자인 COD를 예측하기 위한 알고리즘과 머신러닝 모델로 ARIMAX, RNN, LSTM을 활용한 예측모델을 개발하였다. 그리고 머신러닝 모델의 학습을 위해서, 현재 실제로 운영 중인 국내의 J하수 처리장에서 기록된 18개월간의 하수 수질관련 데이터를 수집하여 사용하였다. 예측 모델에 성능평가 실험 결과에 따른 평가지표인 RMSE와 MAPE의 결과값을 요약하면 <Table 6>과 같다.

Table 6. Evaluation Indicators for Models

Model	RMSE	Ratio of RMSE to mean(%)	MAPE	Average learning time(s)	Average prediction time(s)
ARIMAX (2,1,4)	0.6622	7.83	3.7201	431	0.01
RNN (256,1024) hidden layer	0.3063	3.62	1.9982	661	0.04
LSTM (128,128) hidden layer	0.3004	3.56	1.9927	694	0.04

<Table 6>으로부터 각각의 예측모델의 결과를 살펴보면 다음과 같다.

ARIMAX 모델은 데이터의 증감 경향에 대한 예측은 가능하였지만 구체적인 값의 예측에는 어려움을 보였으며 0.6622의 RMSE값을 나타냈다. 이는 평균대비 약 7.83%이며 분석 모델 중 가장 높은 수치를 보였으며, MAPE 역시 3.7201로 가장 높은 수치를 보였다. ARIMAX 모델과 다른 모델의 차이점은, 다른 모델에 비해 학습속도가 가장 빠르다는 특징이 있다.

RNN 모델은 데이터의 증감 추세 뿐만 아니라 실제값에 매우 가까운 예측을 보여주었다. RNN 모델의 RMSE는 0.3063이며 이는 평균 대비 약 3.62%이고, 1.9982의 MAPE를 가졌다. 이를 통해 ARIMAX 보다 상대적으로 높은 정확도를 보임을 확인할 수 있었다.

LSTM 모델은 RNN과 동일한 환경에서 실험을 진행하였으며 0.3004의 RMSE를 나타냈다. 이는 평균 대비 약 3.56%, 1.9927의 MAPE로 가장 정확도가 높은 결과를 보였다. 하지만 학습시간면에서는 LSTM이 가장 긴 시간이 걸렸다. LSTM 모델은 RNN 모델과 마찬가지로 향후 데이터의 증감 추세와 실제값에 가장 가까운 예측값을 보여줌을 알 수 있었다.

전체적으로, 정확도면에서는 LSTM > RNN > ARIMAX순으로 나타났으며, 학습시간에서는 ARIMAX > RNN > LSTM순으로 오래 걸림을 확인할 수 있었다. 각 모델의 예측시간은 평균적으로 0.04초 이내로서 세 모델 모두 실시간 예측기로 활용하는데 문제가 없는 것을 확인할 수 있었다.

7. 결 론

본 연구에서는 하수처리장에서 수집된 수질 데이터를 활용하여 향후 24시간의 COD를 머신 러닝 모델을 활용하여 예측을 하였다. 예측을 위한 알고리즘으로 ARIMAX, RNN, LSTM 모델을 사용하였다.

분석 결과 수집 데이터의 평균 COD와 모델의 RMSE를 비교하였을 때, LSTM(3.56%), RNN(3.62%), ARIMAX(7.83%) 순으로 높은 성능을 보여주었으며, MAPE로 비교하였을 때도 LSTM(1.9927), RNN(1.9982), ARIMAX(3.7201) 순으로 같은 결과를 나타냈다. 종합적으로 봤을 때, 정확도가 가장 높은 LSTM 모델을 사용한 예측이 가장 효과적이라고 할 수 있다.

또한 학습시간은 세 모델 모두에서 평균적으로 10분 내외로 합리적인 시간 내에서 학습이 가능하며, 예측시간은 각 모델 모두 평균 0.04초를 넘지 않아서 매우 효율적인 실시간 예측기로서 활용할 수 있음을 확인하였다.

하수처리시설에서 개발된 모델을 활용한 COD의 사전 예측 결과를 하수처리 시스템의 운영을 위한 통제변수의 예측값으로 활용한다면, 유입수의 상태를 미리 파악할 수 있어 공정에서 발생하는 문제를 대처하는데 크게 기여할 것으로 예상된다.

하지만 본 연구에서는 하수의 수질에 영향을 미칠 수 있는 기상관련 데이터나 주변 지역 공장의 방류 데이터와 같은 하수처리장 외의 데이터를 분석에 포함시키지는 못하였다. 이러한 다양한 정형 및 비정형적 빅데이터를 포함하여 예측에 활용하는 것과, 공간적 해석 능력이 뛰어난 CNN모델과 결합한 CNN-LSTM 모델 등을 활용하지 못한 것은 연구의 한계로 판단된다. 이는 추후 연구방향으로 남겨둔다.

참고문헌

Bengio, Y., Courville, A., and Vincent, P. (2013), Representation Learning: A Review and New Perspectives, *IEEE Trans. PAMI*, **35**(8), 1798-1828.

Beom, J., Seo, D. H., Park, M. K., and Yoon, K. S. (2019), Correlation of BOD, COD, and TOC by Land Use in the Pungyeongjeongcheon Basin, *Proceedings of the Korean Society of Agricultural Engineers Conference*, 255-255.

Cho, Y. B., Oh, Y. K., Shin, D. C., and Park, C. H. (2014), Distribution of Total Organic Carbon and Correlations between Organic Matters of Sewage Treatment Plants, *Journal of Environmental Analysis, Health and Toxicology*, 17(4), 207-214.

Hochreiter, S. and Schmidhuber, J. (1997), Long Short-Term Memory, *Neural Computation*, **9**(8), 1735-1780.

Houdt, G. V., Mosquera, C., and Nápoles, G. (2020), A Review on the Long Short Term Memory Model, *Artificial Intelligence Review*, **53**, 5929-5955.

Hyndman, R. J. and Athanasopoulos, G. (2018), 2nd ed. Otexts. Forecasting: Principles and Practice, 85-86.

Jung, S. H., Cho, H. S., Kim, J. Y., and Lee, K. H.(2018), Prediction of Water Level in a Tidal River Using a Deep-learning based LSTM Model, *Journal of Korea Water Resources Association*, **51**(12), 1207-1216.

Kim, S. J. and Park, C. G. (1982), A Study on the Correlation among BOD, COD, TOD, and TOC Values for Food - Processing Wastewaters, *Journal of Korean Society of Environmental Engineers*, **4**(1), 8-22.

Korea Environmental Industry & Technology Institute: KEITI(2018), *Water Pollutants (TOC, TN, TP) Automatic Measuring Instrument Commercialization Final Reports*, Humas Co., Ltd.

Lei Tai (2016), Deep-learning in Mobile Robotics - from Perception to Control Systems: A Survey on Why and Why not, J*OURNAL OF LATEX CLASS FILES*, 14(8).

Lim, H. S., An, H. U., Song, I. H., and Yu, S. H. (2021), Prediction of Pollution Loads in Agricultural Reservoirs using Machine Learning, *Proceedings of the Korean Society of Agricultural Engineers Conference*, 68-68.

Moon, T. S., Choi, J. S., Kim, S. H., Cha, J. W., Yoom, H. S., and Kim, C. W. (2008), Prediction of Influent Flow Rate and Influent Components using Artificial Neural Network (ANN), *Journal of Korean Society on Water Quality*, **24**(1), 91-98.

National Institute of Environmental Research: NIER (2015), Study on Management Strategy to Improve the Accuracy of Real-time Monitoring Data for Water Quality, *Research Report*.

National Institute of Environmental Research: NIER (2017), Water Pollution Process Test Standards(Ministry of Environment Notice No.2017-4).

OECD (2012), OECD Environmental Outlook to 2050, 67-70.
Olah, C. (2015), Understanding LSTM Networks, https://colah.github.io/posts/2015-08-Understanding-LSTMs/.
Park, J. W., Moon, M. J., Han, S. W., Lee, H. J., Jung, S. J., Hwang, K. S., and Kim, K. S. (2014), Application of Regression Analysis Model to TOC Concentration Estimation: Osu Stream Watershed, *Journal of Environmental Impact Assessment*, **23**(3), 187-196.
Park, S. J. and Jun, T. J. (1984), Introduction to Forecasting Techniques Box-Jenkins, *Korean Management Science Review*, **1**(1), 68-80.
Seo, I. S., Kim, Y. K., Kim, H. S., and Kim, J. Y. (2013), A Study on Availability of SS, Conductivity and OUR for the Real Time Prediction of Influent COD Concentration in STP, *Journal of Korea Society of Water Science and Technology*, **21**(6), 27-36.
Suhermi, N., Permata, R. P., and Rahayu, S. P. (2019), Forecasting the Search Trend of Muslim Clothing in Indonesia on Google Trends Data Using ARIMAX and Neural Network, *International Conference on Soft Computing in Data Science*, 272-286
Telab, A. (2018), Time Series Forecasting Using Artificial Neural Networks Methodologies: A Systematic Review, *Future Computing and Informatics Journal*, 3(2), 334-340.
WaterEyes Co., Ltd. (2020), *Integrated Information Provision System Using Multi-item Water Quality Measurement for Smart Water City Construction*, Patent.
Zifei, W., Yi, M., Yusha, H., Jigeng, L., Mengna, H., and Peizhe, C. (2019), A Deep Learning Based Dynamic COD Prediction Model for Urban Sewage Electronic Supplementary Information (ESI) Available, *Environmental Science: Water Research & Technology*, **5**(12), 2210-2218.

저자소개

주형구 : 한밭대학교 창의융합학과에서 2022년 학·석사학위를 취득하고, 현재 창의융합학과 지능형자료분석실에서 연구원으로 재직 중이다. 연구분야는 빅데이터분석 및 데이터마이닝이다.

임준묵 : 서울대학교 산업공학과에서 1988년 학사, KAIST 산업공학과에서 1990년 석사학위, 1994년 박사학위를 취득하였다. 강릉원주대학교 산업경영공학과 조교수를 역임하고, 1997년부터 한밭대학교 산업경영공학과 교수로 근무하였으며 현재는 창의융합학과에서 교수로 재직하고 있다. 연구분야는 빅데이터분석, 인공지능 및 빅데이터 융합, 텍스트마이닝 등이다.