J. Korea Water Resour. Assoc. Vol. 57, No. 3 (2024), pp. 151-164
doi: 10.3741/JKWRA.2024.57.3.151
pISSN 2799-8746
eISSN 2799-8754

Study on water quality prediction in water treatment plants using AI techniques

Lee, Seungmin[a] · Kang, Yujin[b] · Song, Jinwoo[c] · Kim, Juhwan[d] · Kim, Hung Soo[e] · Kim, Soojun[f]*

[a]Master's Course, Program in Smart City Engineering, Inha University, Incheon, Korea
[b]Doctor's Course, Program in Smart City Engineering, Inha University, Incheon, Korea
[c]Official, Jemulpo Renaissance Planning Division, Incheon Metropolitan City, Korea
[d]Research Professor, Department of Civil Engineering, Inha University, Incheon, Korea
[e]Professor, Department of Civil Engineering, Inha University, Incheon, Korea
[f]Professor, Department of Civil Engineering, Inha University, Incheon, Korea

Paper number: 23-102
Received: 27 December 2023; Revised: 18 February 2024; Accepted: 19 February 2024

Abstract

In water treatment plants supplying potable water, the management of chlorine concentration in water treatment processes involving pre-chlorination or intermediate chlorination requires process control. To address this, research has been conducted on water quality prediction techniques utilizing AI technology. This study developed an AI-based predictive model for automating the process control of chlorine disinfection, targeting the prediction of residual chlorine concentration downstream of sedimentation basins in water treatment processes. The AI-based model, which learns from past water quality observation data to predict future water quality, offers a simpler and more efficient approach compared to complex physicochemical and biological water quality models. The model was tested by predicting the residual chlorine concentration downstream of the sedimentation basins at Plant, using multiple regression models and AI-based models like Random Forest and LSTM, and the results were compared. For optimal prediction of residual chlorine concentration, the input-output structure of the AI model included the residual chlorine concentration upstream of the sedimentation basin, turbidity, pH, water temperature, electrical conductivity, inflow of raw water, alkalinity, NH_3, etc. as independent variables, and the desired residual chlorine concentration of the effluent from the sedimentation basin as the dependent variable. The independent variables were selected from observable data at the water treatment plant, which are influential on the residual chlorine concentration downstream of the sedimentation basin. The analysis showed that, for Plant, the model based on Random Forest had the lowest error compared to multiple regression models, neural network models, model trees, and other Random Forest models. The optimal predicted residual chlorine concentration downstream of the sedimentation basin presented in this study is expected to enable real-time control of chlorine dosing in previous treatment stages, thereby enhancing water treatment efficiency and reducing chemical costs.

Keywords: AI, Multivariate Regression, Pre-Residual Chlorine, Residual Chlorine Concentration Prediction Model, Chlorine Control

AI 기법을 활용한 정수장 수질예측에 관한 연구

이승민[a] · 강유진[b] · 송진우[c] · 김주환[d] · 김형수[e] · 김수전[f]*

[a]인하대학교 스마트시티공학과 석사과정, [b]인하대학교 스마트시티공학과 박사과정, [c]인천광역시 제물포르네상스계획과 주무관,
[d]인하대학교 사회인프라공학과 연구교수, [e]인하대학교 사회인프라공학과 교수, [f]인하대학교 사회인프라공학과 교수

요 지

상수도 공급을 위한 정수장에서 전염소 또는 중염소 공정이 도입된 수처리 공정의 염소농도 관리에 필요한 공정제어를 위하여 AI 기술을 활용한 수질예측 기법이 연구되고 있다. 본 연구에서는 정수장 수처리 공정에서 실시간으로 관측, 생산되고 있는 수량·수질자료를 이용하여 염소소독 공정제어 자동화를 목적으로 침전지 후단의 잔류염소 농도를 예측하기 위한 AI 기반 예측모형을 개발하였다. AI 기반 예측모형은 과거 수질 관측자료를 학습하여 이후 시점의 수질에 대한 예측이 가능한 기법으로, 복잡한 물리·화학·생물학적 수질모형과 달리 간단하고 효율적이다. 다중회귀 모형과 AI 기반 모형인 랜덤포레스트와 LSTM을 이용하여 정수장의 침전지 후단 잔류염소 농도를 예측하여 비교하였다. 최적의 잔류염소 농도 예측을 위한 AI 모형의 입출력 구조로는 침전지 전단의 잔류염소 농도, 침전지 탁도, pH, 수온, 전기전도도, 원수의 유입량, 알칼리도, NH_3 등을 독립변수로, 예측하고자 하는 침전지 유출수의 잔류염소 농도를 종속변수로 선정하였다. 독립변수는 침전지 후단의 잔류염소에 영향이 있는 정수장에서 확보가 가능한 관측자료중에서 분석을 통해 선별하였으며, 분석 결과 연구대상 정수장인 정수장에서는 중회귀모형, 신경망모형, 모델트리 및 랜덤포레스트 모형을 비교한 결과 랜덤포레스트에 기반한 모형오차가 가장 낮게 도출되는 결과를 얻을 수 있었다. 본 연구에서 제시하는 침전지 후단의 적정 잔류염소 농도 예측값은 이전 처리단계에서 염소주입량의 실시간 제어가 가능토록 할 수 있어 수처리 효율 향상과 약품비 절감에 도움이 될 것으로 기대된다.

핵심용어: AI, 다중회귀, 전염소, 잔류염소 농도 예측모형, 염소제어

***Corresponding Author.** Tel: +82-32-860-7563
E-mail: sk325@inha.ac.kr (Kim, Soojun)

1. 서 론

최근 세계적으로 인구, 기후, 생활환경 등의 여건이 변화하면서 물, 대기, 생태계 등 다양한 환경분야에서 문제가 대두되고 있다. 특히 국민의 삶과 건강에 직접적인 영향을 가지는 상수도분야는 환경, 사회·경제 등 전 분야에서 다양한 문제가 발생하고 있다.

물 공급 역할을 하는 상수도시설의 공정은 크게 단위공정, 전체공정으로 분류한다. 상수도시설의 모든 공정에서는 수질관리, 약품의 주입, 정수처리설비 운영 등에 대한 평가 및 개선을 위한 제어기술이 필요하다. 식수의 수질기준 강화로 관리 기술의 과학화, 기술능력 향상 및 기능별 전문화 등이 요구됨에 따라 운영 전문성을 높이는 것이 필요하다.

현재 정수장의 수질관리 기술은 응집 및 여과 공정이나 오존 등 일부 단위공정에 대한 인공지능 기반 제어기술 개발이 추진되고 있으나, 아직 전체 공정의 지능화는 도입에 성공하지 못하였다. 실제로 인공지능 기술을 활용한 공정제어 기술은 일부 도입에 성공한 하수처리 분야와 달리 정수처리 분야에서는 적용된 사례가 적은 실정이다.

국내 상수도 분야의 연구는 상수관망을 위주로 수행되어 왔으며, 정수 시설은 일부 공정에 대해서만 연구되었다. 그러나 최근에 수돗물의 적수 사태 및 유충 등의 문제가 발생함에 따라 정부에서는 취수원부터 수도시설까지 모든 과정을 실시간으로 파악하고, 사고가 발생할 경우 자동으로 대응할 수 있는 스마트 체계의 구축을 목표로 한다. 이에 따라 정수장에 인공지능 기술을 적용하여 물 공급의 핵심 시설로 강화하고자 하는 시도가 있으나, 현재에는 일부 공정만을 인공지능이 제어할 수 있기 때문에 더 많은 공정에 인공지능을 적용하는 시도가 필요하다(ME and KEITI, 2021).

Zhang and Stanley (1999)은 정수처리공정의 침전 및 응집공정의 효율성을 높이기 위하여 인공신경망 기반의 모형을 사용하였다. 해당 연구에서는 2,000개의 공정 제어 자료를 이용하여 신경망 공정 모형을 구축하였다. 해당 연구에서 개발한 모형은 약간의 입력자료 및 매개변수 수정으로 다른 수질관리 공정에 대한 제어 모형을 구축할 때 사용이 가능하다(Zhang and Stanley, 1999).

Lee *et al*. (2007)은 침전지에서 유출수의 잔류염소 농도를 유지하기 위한 전염소 투입량 예측모형을 개발하였다. 이 모형은 다중회귀 모형과 신경망 모형을 이용하여 개발하였으며, 염소의 분해속도에 영향을 가지는 인자들을 활용한다. 해당 연구에서 개발한 모형은 침전지 유출수 잔류염소의 농도를 예측할 수 있으며, 최적의 전염소의 투입량을 결정하는데에 있어 유용한 모형으로 활용할 수 있다(Lee *et al*., 2007).

Lowe *et al*. (2022)은 수질관리 분야에서 머신러닝을 이용한 수처리 공정이 기존 방법론의 복잡성을 개선하고 비용을 절감 수 있을 것으로 기대하였다. 그는 인공지능 기반의 방법론이 제어 및 최적화, 모델링 등과 같은 문제점들을 효과적으로 해결할 수 있다고 주장한다. 그러나 이러한 지능형 모형을 성공적으로 구현하기 위해서는 부실한 데이터의 관리, 모형의 낮은 재현성 등의 문제점이 있음을 지적하였다(Lowe *et al*., 2022).

정수처리공정에 있어 응집제 등 약품들의 주입량을 결정하는 것은 중요한 요소이며, 현재 이를 위한 수질 관리 및 잔류염소 제어 등을 위한 모형의 개발에 대한 연구가 진행되고 있다(Gulzar *et al*., 2022; Jun *et al*., 2001; Kim *et al*., 2001, 2014, 2016a, 2016b, 2022b; Kyoung *et al*., 2006; Lee *et al*., 2014; Yoon *et al*., 2001). 본 연구에서는 이러한 선행연구들을 바탕으로 다양한 인공지능 알고리즘을 통해 정수처리과정에서 중요한 잔류염소를 예측하고, 약품투입량을 자동제어하는 방법론을 개발하여 기존 정수처리공정의 인력 및 관리 방법으로 인한 문제점을 개선한다.

AI 기반의 예측모형은 수질자료의 입력 및 출력이 간단하며 물리적, 화학적, 생물학적 과정을 고려하지 않고 수집한 자료를 이용하여 수질을 예측할 수 있다. 실제로 AI는 문제를 해결하는 도구로 다양한 분야에서 적용되어 큰 성과를 이루고 있다. 수질관리 기술에 적용되는 인공지능은 수처리 과정에서 발생하는 비용을 절감하고, 약품 사용을 최적화하는 데에 큰 효과가 있을 것으로 기대된다.

부평 정수장은 수돗물 생산을 담당하는 시설로 10,000개가 넘는 태그로 수많은 인자들을 계측하고 있다. 그러나 2~3명정도의 적은 인력으로 운영되기 때문에 소수의 중요한 데이터만을 관리하며, 이벤트 알람을 통해 사람이 육안으로 확인하기 어려운 상황을 인지한다. 이러한 문제를 해결하기 위하여 인천광역시 부평구에 위치한 정수장의 자료를 활용하여 해당 정수장에 적용이 가능한 기술을 개발하여 빅데이터를 활용한 시스템을 도입하고자 한다.

인공지능 기반의 빅데이터 처리기술은 정수처리공정에 있어 매우 유용하다. 정수처리공정은 수질·유량이 자동으로 계측되는 시스템으로, 비교적 명확한 경계조건을 가진다. 해당 자료들을 분석하여 정수처리공정의 관리 및 제어, 예측, 최적화 등을 할 수 있다. 정수처리공정은 취수장에서 공급받은 물에 응집제를 투입하고, 응집·침전공정에서 용해되지 않은 화학 물질을 제거한다. 이후 여과 및 정수공정을 수행한 후 배수지 및 관로로 시민들에게 물을 공급하는 과정을 가진다.

정수처리 공정에 있어서 중요한 과정중 하나인 염소처리 공정은 염소의 주입 위치에 따라 전염소, 후염소 처리로 구분된다. 전염소 처리공정은 취수장에서 바로 염소를 주입하여 원수 단계에서부터 미생물을 제거하는 공정이며, 후염소 처리공정은 정수장에서 염소를 주입하여 용수를 소독을 하는 공정이다. 과거에는 일반적으로 후염소 처리 방법을 많이 사용하였으나, 염소가 수중에서 암모니아성 질소 또는 암모늄 이온과 반응하면서 유해한 부산물이 발생하는 문제가 발생하였다. 현재에는 이러한 문제점을 해결하기 위하여 전염소 처리공정이 도입되었다. 전염소 처리공정은 취수량 및 원수의 수온·수질, 날씨와 계절 등에 영향을 받는다. 이러한 요소들을 종합적으로 고려하여 전문가가 수동으로 염소의 주입을 결정하는 경우가 많다.

국민의 생활용수 급수를 시행하는 상수도시설은 정수장과 도수, 송수, 배수, 급수시설로 구성된다. 정수장은 취수한 원수를 착수, 혼화, 침전, 여과, 고도정수, 염소소독 등 일련의 정수처리 과정을 거친 후 생활용수로 공급한다.

공급되는 생활용수에는 미생물 등과 같은 생물학적 오염을 방지하기 위하여 정수과정에 염소를 주입한다. 염소주입 공정은 용수의 생물학적 안전성에 중요한 과정이다. 먹는물의 수질 기준에 따르면 물 1 L당 할당되는 잔류염소량의 범위는 4.0 mg 이하이다(WHIM, 2023). 또한 염소주입공정은 염소의 주입 위치에 따라 전염소, 후염소 등의 처리공정으로 분류하며, 이는 원수의 수질에 따라 결정된다. 최종적으로 각 공정에서의 주입 제어는 잔류염소 농도 목표치를 유지하기 위해 설정되며, 송수 및 배수과정에서 잔류염소의 감소를 고려하여 목표 잔류염소 농도를 조정한다.

전염소 처리공정에서는 원수가 응집, 침전 등 다른 공정을 거치기 전에 염소를 주입한다. 원수의 조류 및 기타 병원균, 암모니아성 질소 등을 제거하고 원수에 함유되어 있는 철이나 망간을 산화시켜 침전을 쉽게 할 수 있도록 처리하는 것이 주요 목적이다. 염소와 유기물질이 오랜 시간동안 접촉할 경우, 트리할로메탄 등과 같이 유해한 부산물이 생성될 수 있기 때문에 염소의 용량 및 잔류시간을 적절하게 조절해야 한다.

후염소 처리공정은 수질오염에 따른 병원성 미생물을 억제하며, 정수 및 송수 등 이후 과정에서의 재오염을 방지하기 위하여 실시한다. 정수지에서 염소를 주입하여 살균한 후 잔류염소를 적정 수준으로 유지하여 용수의 품질을 보장한다.

잔류염소는 유량의 변화에 따라 적절한 염소주입량을 결정하여 목표치를 유지하도록 관리해야 한다. 염소의 주입량은 수온 및 원수의 수질 등의 요인들을 종합적으로 고려하여 잔류염소가 목표값을 유지할 수 있도록 관리한다. 잔류염소의 목표설정지점의 잔류염소 농도를 정기적으로 측정·검증하여 염소 주입량의 효과를 평가하고 개선해야 한다.

전·후염소처리 공정은 외부 환경 등과같이 실시간으로 변화하는 요인으로 인하여 예측하여 제어하기에는 상당히 어려

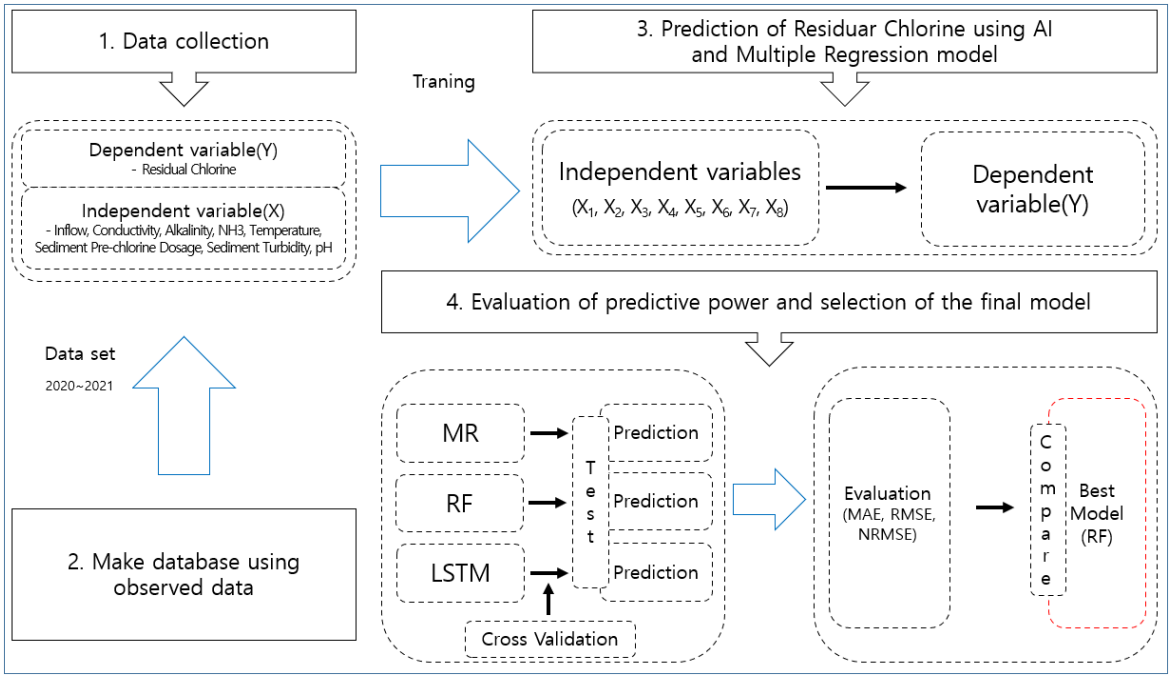


Fig. 1. Flow chart of the study

운 실정이다. 수동제어를 할 경우에도 관련 전문가가 24시간 동안 염소주입 설정량을 조절해야 하기 때문에 2~7시간의 체류시간동안 온전히 제어하기에는 큰 어려움이 있다.

본 연구에서는 기존 연구에서 수행된 AI 알고리즘의 구조를 구체화 및 보완하고, 매개변수의 최적화 방법론으로 유전알고리즘을 적용하여 최적의 잔류염소 예측모형을 개발하였다. 현장에서 측정이 가능하고 염소의 분해속도에 영향력을 가지는 인자들을 선정하고, 자료를 수집하였다. 랜덤포레스트(Random Forest, RF), LSTM (Long-Short Term Memory)과 같은 AI 기반 모형과 다중회귀 모형을 활용하여 전염소 주입량을 제어할 수 있는 모형을 개발하고, 그 성능을 비교 및 평가하였으며, 그 과정을 Fig. 1과 같이 나타내었다. 연구에 사용된 모형들은 모두 다른 방식으로 데이터를 예측하며, 최종적으로 잔류염소 농도 예측에 있어 어떤 모형이 가장 적합한지 평가하였다. 최종적으로 정수처리 공정에서는 랜덤포레스트가 가장 우수한 정확도를 보였으며, 의사결정 트리구조를 사용하는 인공지능 모형을 활용하는 것이 유용하다는 것이 도출되었다.

2. 연구 방법

2.1 자료 및 연구 대상

연구 대상으로는 인천광역시 부평구에 위치한 부평 제1정수장으로 선정하였다. 인천광역시 부평에는 1, 2, 3정수장이 있으며, 실제로 가동되고 있는 정수장은 1, 3정수장이다. 그 중 관측자료의 종류가 많고, 자료의 품질이 준수한 정수장인 제1정수장의 자료를 수집하여 본 연구에 사용하였다. 정수장에서는 관측지점별로 탁도, 응집제 투입량, 잔류염소 농도, 알칼리도, 유입량, pH 등과 같은 자료를 시간단위로 관측한다. 본 연구에서는 그 중 실제로 현장에서 잔류염소를 제어할 때 고려하는 인자인 유입량(inflow), 전기전도도(conductivity), 알카리도(alkalinity), 암모니아 농도(NH_3), 수온(temperature), 침전지 전단 염소농도(sediment pre-chorine), 침전지 탁도(sediment turbidity), pH, 침전지 유출수의 잔류염소(residuar chlorine) 등 총 9가지의 관측자료를 사용하였다. 침전지 유출수의 잔류염소를 종속변수로, 그 외 유입량, 전기전도도 등의 관측자료를 독립변수로 활용한다. 자료는 2020년부터 2022년까지의 10분단위의 시계열 관측자료가 있다. 해당 자료에서 결측치가 있는 시간의 자료는 제외하고, 시간단위로 변환하여 입력자료를 구축하였다. 일반적으로 정수장에서 자료를 관측하는 지점과 정수처리과정을 Fig. 2와 같이 나타내었다.

2.2 정수장 잔류염소 농도 예측 지능화

현재 정수처리 공정에 인공지능 알고리즘을 적용하기 위한 많은 시도가 있다. 인공지능 기반 모형으로는 랜덤포레스트(Random Forest, RF), 서포트 벡터 머신(Support Vector Machine, SVM), XGBoost, 인공신경망(Artificial Neural Network, ANN), 장단기 기억(Long Sort-Term Memory, LSTM) 등과 같은 모형들이 있다. 본 연구에서는 그 중 랜덤포레스트, LSTM 기반 예측모형과 다중회귀모형을 사용하여 분석하고 비교하였다.

2.2.1 다중회귀모형

다중회귀모형(Multiple Regression model, MR)은 다수의 독립변수들을 활용하여 종속변수와의 연관성을 분석하는 분석하는 통계적 방법이다. 이는 하나의 독립변수만을 고려하는 단순회귀분석과 달리 두 종류 이상의 독립변수들을 활용하여 종속변수와 독립변수 간 선형관계를 추정할 수 있다. 가장 일반적으로 Eq. (1)과 같은 회귀식을 사용하여 회귀계수를 산정한다(Kang *et al.*, 1992; Lim, 2019; Ngo and La Puente, 2012).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \epsilon_i \quad (1)$$

여기서 β_0는 상수항, x_i는 독립변수, β_i는 독립변수의 회귀계수, ϵ_i는 회귀식에서 종속변수를 추정할 때 발생하는 오차를 의미한다.

다중회귀모형에서는 오차가 정규분포를 따르며 서로 독립

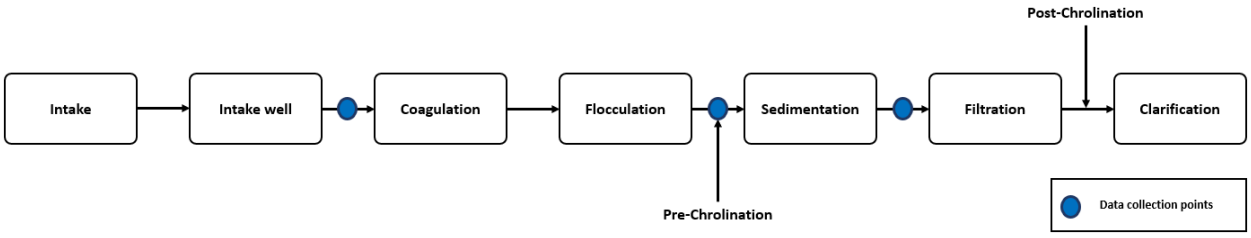


Fig. 2. Water treatment plant and data collection points (Kim *et al.*, 2022b)

적인 관계를 가진다는 가정을 하는 것이 일반적이며, 최소제곱법(Ordinary Least Square, OLS)을 활용하여 오차의 제곱합계를 최소화함으로써 회귀계수를 추정하는 방법이 사용되고 있다(Ngo and La Puente, 2012). 독립변수 간 상관성이 없을 경우에는 회귀모형의 성능이 향상된다. 그러나, 독립변수 간에 상관성이 있거나 표본의 개수가 제한적일 경우에는 다중공선성 문제가 발생할 가능성이 있다. 다중공선성은 회귀계수 추정치의 분산을 증가시켜 모형의 신뢰성을 저하시키는 현상이다(Kwon *et al.*, 2017). 다중공선성의 명확한 판단기준은 존재하지 않는다. 그러나, 일반적으로 다중공선성을 판단할 때에는 입력자료 간의 상관계수나 분산팽창지수(Variance Inflation Factor, VIF)값을 활용한다. 상관계수의 경우 입력자료 간 값이 0.9 이상일 경우 다중공선성이 있다고 보고, 0.7~0.8 사이의 값이 나타날 경우 다중공선성을 검토한다. 상관계수를 통한 다중공선성 검토를 실시한 후 검토가 필요할 경우 VIF 값을 산정하여 다중공선성을 판단할 수 있다. 여기에도 다중공선성의 여부를 명확하게 구분하는 기준은 없으나, 일반적으로 VIF 값이 10보다 클 경우 다중공선선성이 있다고 판단한다(Diamantopoulos and Siguaw, 2006; Kwon, 2015; Petter *et al.*, 2007).

2.2.2 랜덤포레스트

랜덤포레스트는 1990년대 후반에 Amit and Geman (1997), Ho (1998)가 최초로 제안한 모형이다. 이후 Breiman (2001)이 발전시켜 현재 사용중인 형태가 완성되었다. 랜덤포레스트는 다수의 의사결정나무(Decision Tree, DT)를 통합한 앙상블 형태의 모형으로, 각 결정트리를 배깅(bagging) 방법으로 학습한다. 이 방법을 통해 각 트리가 선택하는 변수를 무작위로 설정하여 트리간의 상관성을 줄여 모형의 과적합을 방지할 수 있다. 랜덤포레스트는 CART (Classification And Regression Tree) 방법을 토대로 구현되며, 분류 및 회귀분석에 널리 적용되고 있다(Amit and Geman, 1997; Breiman, 2001; Ho, 1998).

랜덤포레스트는 배깅방법을 활용해 학습 데이터를 무작위로 추출한 후 일반적으로 약 500개의 결정 트리를 생성하여 그 결과를 집계(voting)하여 결과를 제시한다. 대표적인 랜덤포레스트의 초 매개변수(hyper parameter)는 무작위 분류기의 수를 나타내는 mtry가 있다(Kim, 2022; Liaw and Wiener, 2002; Lee *et al.*, 2023a). 최종적으로 주어진 데이터를 사용해 모형의 매개변수를 조절하여 최적의 결과를 도출한다(Fig. 3).

2.2.3 장단기 메모리

장단기 메모리(Long Short-Term Memory, LSTM)는 순환신경망(Recurrent Neural Network, RNN)에서 파생된 모형이며, 셀(cell) 및 입력(input), 출력(output), 망각(forget) 게이트를 활용하여 기존 순환신경망의 기울기 소멸 문제를 해결하기 위한 방안으로 개발되었다(Hochreiter and Schmidhuber, 1997; Kim, 2022). 일반적인 순환신경망이 과거 자료에 대한 은닉층의 결과만을 업데이트하는 한계를 가진 반면, LSTM은 이를 개선하여 장기적인 데이터 저장을 가능하게 하는 추가적인 셀의 구조를 은닉층에 추가한 모형이다(Le *et al.*, 2019;

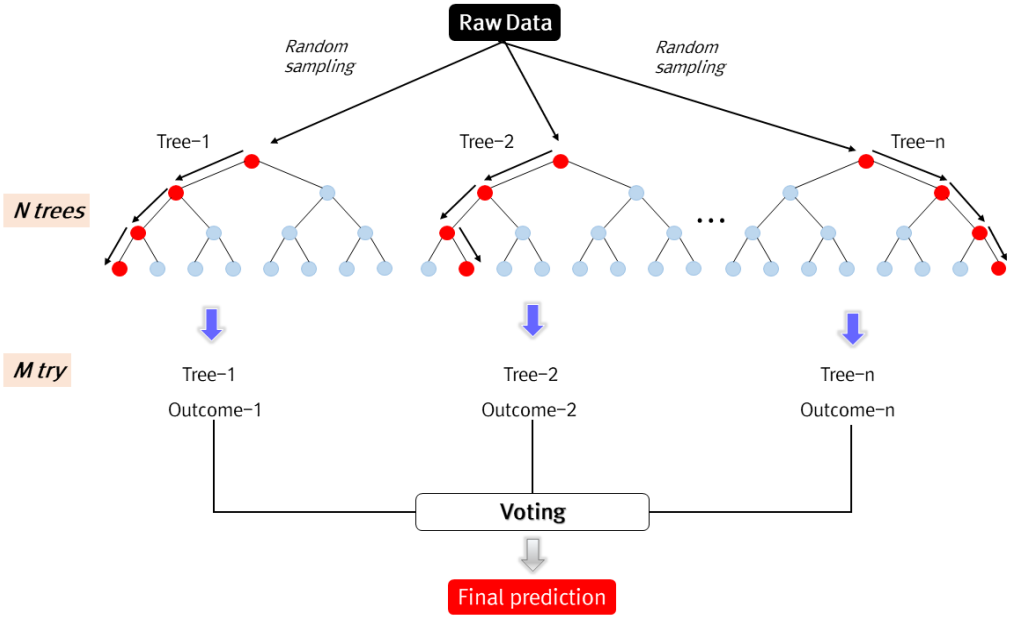


Fig. 3. Conceptual diagram of random forest (Kim, 2021)

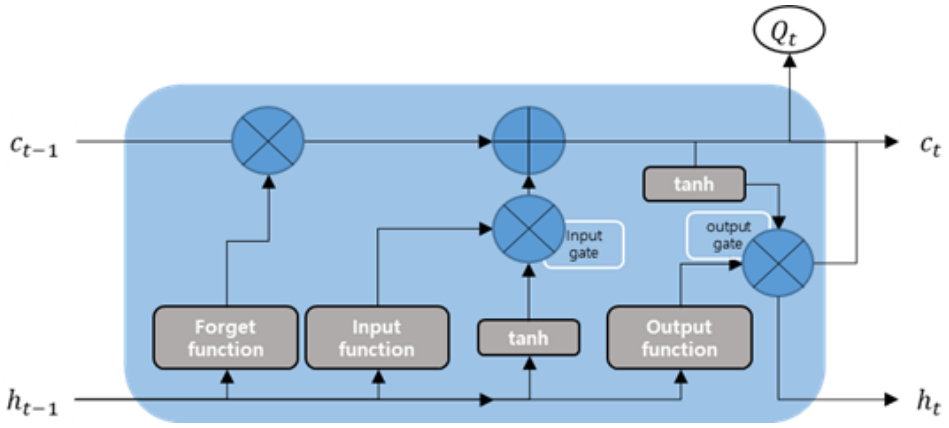


Fig. 4. Conceptual diagram of LSTM (Kim, 2022)

Han *et al*., 2021). 대표적인 LSTM의 초 매개변수는 레이어(layer)의 뉴런 수를 의미하는 units, 학습 시 한번에 처리되는 데이터 샘플의 수를 의미하는 batch size가 있으며, LSTM의 기본적인 개념도를 Fig. 4와 같이 나타냈다.

2.3 AI 기반 모형의 매개변수 최적화

AI 기반 모형에는 조정해야 할 매개변수가 있다. 일반적으로 LSTM 모형의 ADAM 모형 같이 자체적으로 탑재된 알고리즘을 통해 최적화하는 매개변수가 있고, 사용자가 조정해야 하는 초 매개변수가 있다. 본 연구에서 조정해야 하는 초 매개변수는 랜덤포레스트의 mtry, LSTM의 units와 batch size 등이 있다. 본 연구에서는 기본적으로 모형에 포함되어 있는 최적화 알고리즘을 사용하되, 초 매개변수의 최적화 방법론은 유전 알고리즘을 사용하였다.

유전 알고리즘은 최적화 방법론의 일종으로, 자연 세계에서 생명체가 환경에 적응하며 진화하는 과정에서 영감을 얻어 개발되었다. 유전 알고리즘은 환경에 적합한 생명체가 더 많은 후손을 남기고, 세대를 거듭하면서 유전자 변이를 통해 적응성을 강화하는 생물학적 메커니즘을 기반으로 한다. 유전 알고리즘의 주된 특징은 뉴턴법과 같은 전통적인 최적화 방식과 달리 단일 해답이 아닌, 다수의 가능한 해결책들을 포함하는 최적해 집단을 탐색한다는 것이다(Holland, 1975; Kim and Kim, 2007a, 2007b; Lee *et al*., 2023b).

유전 알고리즘은 유전자가 세대를 거쳐 특성을 전달하는 개념을 바탕으로 구축된다. 유전 알고리즘에서 최적화하고자 하는 목표는 목적함수로 정의되며, 최적화할 변수의 계수는 유전자로 간주된다. 이 과정은 종속변수의 적합도가 만족스러울 때까지 진행된다.

2.3.1 초기 유전자 집단 생성

유전 알고리즘은 초기 유전자 집단 생성(Initial Population)으로 시작한다. 해당 과정에서는 목적함수(objective function) 및 유전자로 활용할 초기 변수를 결정한다. 본 연구에서는 초기 유전자 집단을 생성하기 위하여 주어진 개체 수와 유전자의 수에 따른 무작위 값으로 결정되었다. 생성된 초기 유전자 집단은 알고리즘의 시작 변수로 활용되며, 이후 과정에서 개선된다.

2.3.2 적합도 계산

적합도 계산(Fitness Evalution)은 각 개체의 성능을 평가하는 단계이다. 본 연구에서는 목적함수에 AI 모형을 탑재하고, 최적화하고자 하는 초 매개변수를 유전자로 적용하여 결과 값을 계산한 후 실제 값과의 제곱오차평균(Mean Square Error, MSE)을 적합도로 사용하였다. 이렇게 계산된 적합도는 그 값이 작을수록 개체의 성능이 좋다는 것을 의미한다. 좋은 적합도를 가진 개체가 높은 확률로 다음 세대에 전달되며, 해당 단계는 재생산과정 이후에도 동일하게 진행된다. 본 연구에서는 유전 알고리즘에서 매 세대별로 적합도를 평가하고, 적합도의 변화가 100세대동안 변화가 없을 경우 알고리즘을 종료하도록 설계하였다.

2.3.3 재생산

재생산(reprouduction) 단계는 기존 세대의 유전자를 바탕으로 새로운 세대의 개체들을 생성하는 과정이며 선택, 교배, 돌연변이 등의 세부단계로 구성된다.

① 선택(selection): 더 높은 적합도를 지닌 개체들을 선택. 이때 적합도는 유전자의 선택 확률로 활용되며, 개체의 적합도가 높을수록 후속 세대로 전달될 확률이 증가

Fig. 5. Conceptual diagram of cross validation

② 교배(crossover): 두 개의 개체에서 유전자를 서로 교차시켜 새로운 개체를 생성하는 과정

③ 돌연변이(mutation): 유전자에 임의의 변화를 추가함으로써 생물학적 다양성을 유지하는 방식으로, 설정된 돌연변이 확률에 따라 유전자 값을 임의로 변형

2.4 모형의 검증 방법

검증을 위해 사용되는 방법은 일반적으로 학습구간과 별개로 평가구간을 분류하여 예측 성능을 측정해야 하지만, 이러한 방법은 과적합(overfitting)이 발생할 가능성이 있다(Kim *et al*., 2022a). 과적합은 모형이 학습 자료를 과하게 학습하는 것을 의미한다. 과적합이 발생할 경우, 학습구간에서는 모형의 성능이 매우 높게 나오지만 평가구간에서는 모형의 성능이 낮게 나온다. 즉, 실제로 사용하기 어렵고, 범용성이 없는 모형이 되는 것을 의미한다. 이러한 문제를 해결하기 위하여 본 연구에서는 k-겹 교차검증(k-fold cross validation) 방법을 활용하였다. 교차검증 방법은 예측모형의 성능을 판단하는 방법 중 하나로, 사용한 입력자료를 여러 폴드(fold)로 분류하여 한 부분을 검증용 집합으로 활용한 후 나머지 부분을 모형의 학습에 활용한다. Fig. 5와 같이 모든 입력자료에 대하여 해당 과정을 반복하여 모형의 평균 예측 오류(average prediction error)를 측정함으로써 모형의 성능을 평가한다(Bates *et al*., 2023; Jung *et al*., 2018).

3. 잔류염소 예측을 위한 AI 기반 예측모형 적용 및 평가

3.1 정수장 수질 관측자료 수집

부평 정수장은 하천수를 사용하는 주요 정수 처리 공정을 운영하며, 다른 정수장들과 비교할 때 관측 데이터가 풍부하여 본 연구의 대상으로 선정되었다. 연구에 사용된 데이터는 2020년 1월 1일부터 2022년 12월 31일까지 수집된 최근 3년간의 데이터이다. 본 연구에서 수집한 입력자료별 시계열 그래프를 Fig. 6과 같이 나타내었다.

정수장에서의 수질 데이터의 수집과 분석에는 여러 가지 어려움이 존재한다. 각 정수장의 처리 방법의 차이, 센서와 통신의 오류, 청소 작업 등의 요인으로 인해 데이터에는 종종 결측치와 이상치가 포함된다. 이를 효과적으로 활용하기 위해, 데이터 전처리는 반드시 필요하다. 수질관측자료의 경우 사람이나 시설, 계측 기계, 기후 등으로 인한 변수가 많은 것이 특징이다. 이러한 특징으로 인해 기존에 개발된 통계적 분석기법을 활용한 결측치 및 이상치 보간은 유효하지 않기 때문에 본 연구에서는 다음과 같은 순서로 데이터의 전처리를 수행하였다.

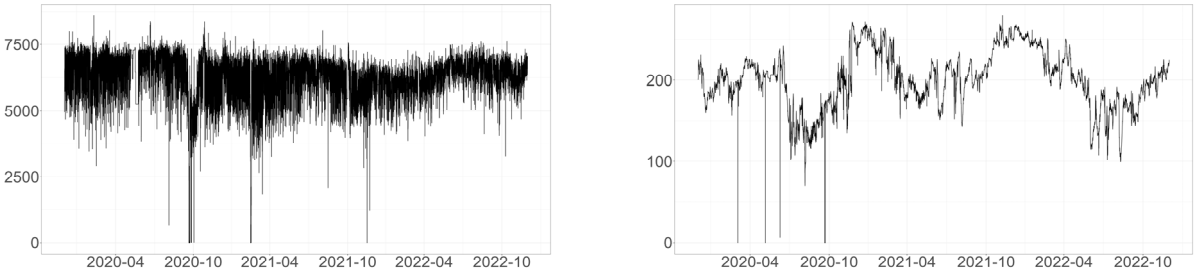
(1) 수집된 데이터셋에서 각 변수별로 발견된 결측치는 선형 보간 방식을 사용하여 결측치를 보간

(2) 각 변수에 대한 시계열 그래프를 작성하여 자료의 패턴 분석 후 이상치가 확인되는 부분을 제거

(3) 각 변수에서 이상치를 찾아내고, 이상치가 집중된 구간을 데이터셋에서 배제

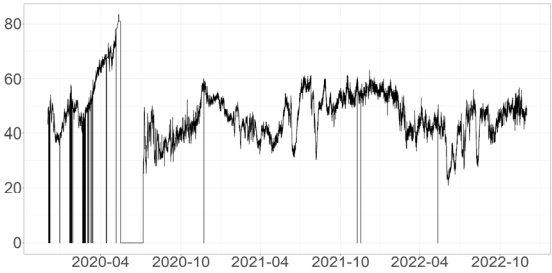
(4) 10분마다 기록된 자료 중 정시에 해당하는 자료만을 선택하여, 이를 1시간 단위로 재구성

본 연구는 정수 처리 과정에서의 염소 수준 변화를 이해하기 위해 다양한 수질 지표들을 조사했다. 이러한 지표들로는 유입량(inflow), 전기전도도(conductivity), 알칼리도(alkalinity), 암모니아의 농도(NH_3), 물의 온도(temperature), 착수정의 염소 농도(sediment Cl_2), 탁도(turbidity), pH, 침전지 유출수의 잔류염소가 포함된다. 분석된 기간은 2020년부터 2022년까

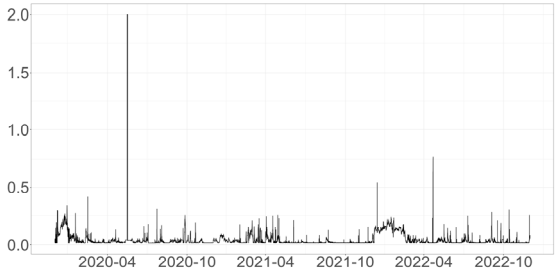


(a) Inflow

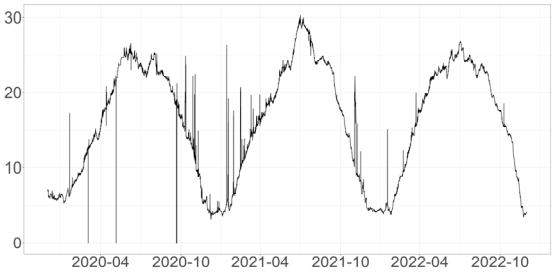
(b) Conductivity



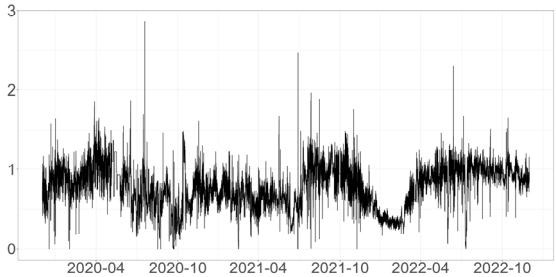
(c) Alkalinity



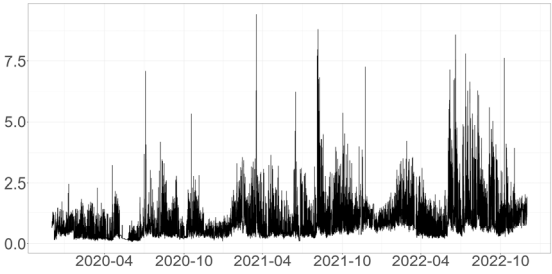
(d) NH_3



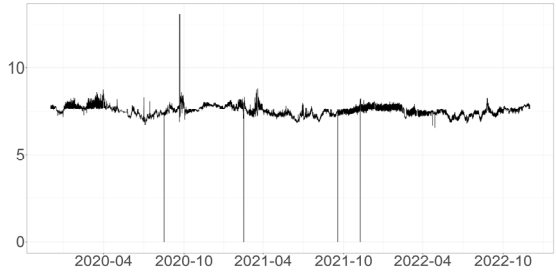
(e) Temperature



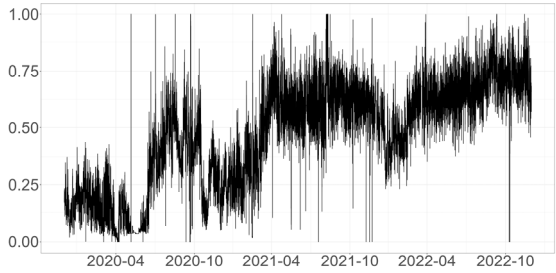
(f) Sediment Cl_2



(g) Turbidity



(h) pH



(i) Settle Cl_2

Fig. 6. Time series graphs by input data

Table 1. Basic statistics of input data

Parameter	Average	Max	Min	Standard Deviation	Coef of Variation
Inflow	6230.28	8616.88	2659.38	743.83	0.12
Conductivity	200.65	272.50	70.13	35.54	0.18
Alkalinity	46.99	83.29	2659.38	9.07	0.19
NH_3	0.04	0.77	0.02	0.04	1.00
Temperature	15.65	30.24	3.15	7.51	0.48
Sediment Pre. Cl_2	0.78	2.86	0.00	0.26	0.33
Turbidity	0.91	9.43	0.12	0.73	0.81
pH	7.56	8.83	6.51	0.27	0.04
Settle Cl_2	0.49	1.00	0.00	0.22	0.44

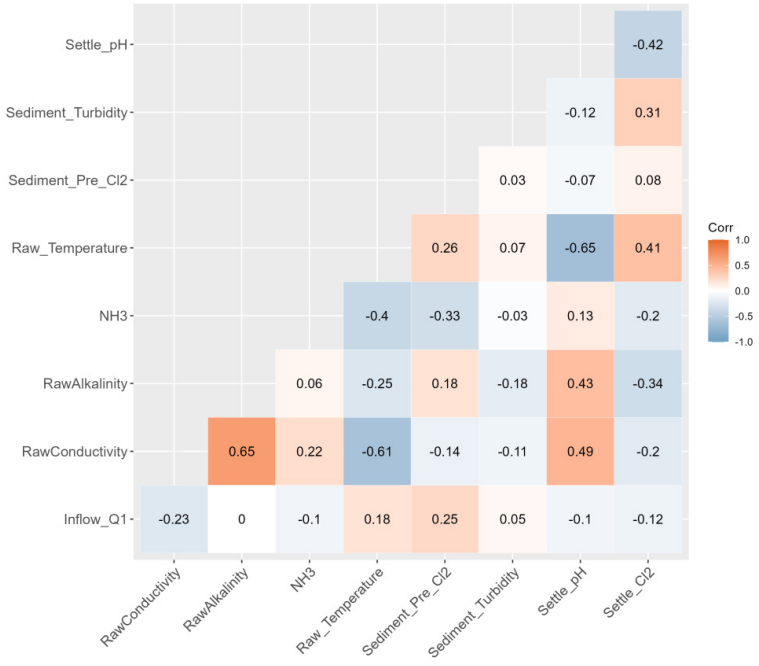


Fig. 7. Correlation coefficient between input data

지로, 대상 기간 동안 수집된 10분 간격의 약 160,000개 데이터 중에서 결측치 보간 및 이상치를 제외한 후 약 25,000개의 시간 단위 자료를 사용하였다. 최종적으로 전처리가 완료된 입력자료의 기초통계량을 Table 1과 같이 나타내었다.

3.2 입력변수간 상관성 분석

분석에 사용한 자료의 상관성을 분석하기 위하여 피어슨(Pearson) 상관분석을 수행하였다. 피어슨 상관분석은 두 변수 간의 선형 관계를 판단할 때 유의미한 지표이다. 해당 분석을 통해 나오는 상관계수는 -1에서 1 사이의 값을 가지며, 0에 수렴할수록 두 변수 간 상관성이 약하다고 볼 수 있고, -1이나 1에 수렴할수록 두 변수 간에 강한 상관성이 있다는 것을 의미한다(Cohen *et al.*, 2009). 본 연구에서 적용되는 분석 방법론의 경우 독립변수와 종속변수 간 상관관계가 높을 경우 다중공선성 또는 과적합 문제가 발생할 가능성이 있다. 이에 따라 분석에 앞서 각 입력자료 간 상관성을 검토하였다(Fig. 7).

분석 결과 종속변수로 사용되는 침전지 잔류염소 농도와의 상관계수는 원수 유입량에 -0.12, 전기전도도에 -0.2, 알카리도에서 -0.34, 암모니아 농도에 -0.2, 수온에 0.41, 침전지 전단 염소농도에 0.08, 침전지 탁도에 0.31, pH에 -0.42로 나타났다. 종속변수와 모든 독립변수 간의 상관계수는 0.7 미만으로, 일반적으로 다중공선성이 발생하지 않을 것으로 판단하는 기준에 부합한다. 따라서, 본 연구에서 사용한 변수들을 다중회귀모형에 적용할 경우 다중공선성 문제는 발생하지 않

는 것으로 확인되었다.

3.3 모형 적용 및 평가

본 연구에서는 정수장 침전지 후단의 잔류염소 농도를 예측하는 모형을 구축하고자 여러 모형을 활용하였다. 기반 모형에는 다중회귀모형, LSTM, 랜덤포레스트 모형이 포함된다. 모형의 성능평가를 위하여 자료의 80%를 학습용으로, 그 외의 20%를 검증용으로 분류하였다. 모형의 성능은 검증용 자료를 대상으로 실제 관측값과 학습 후에 도출되는 예측값간의 여러 지표를 비교하였다. 평가 지표로는 절대오차평균(Mean Absolute Error, MAE), 결정계수(R-sqared score, R^2), 평균 제곱근 오차(Root Mean Squared Error, RMSE), RMSE를 정규화한 NRMSE (Normalized Root Mean Squared Error)를 사용하였다.

3.3.1 다중회귀모형 적용 결과

다중회귀분석의 함수식으로는 가장 일반적으로 사용하는 함수식을 사용하였으며, 사용한 함수식을 Eq. (2)과 같이 나타내었다.

$$aX_1 + bX_2 + cX_3 + dX_4 + eX_5 + fX_6 + gX_7 + hX_8 + \epsilon = Y \quad (2)$$

여기서 a, b, ..., h는 각 독립변수별 회귀계수, X_i는 독립변수, Y는 종속변수, ϵ은 회귀식의 절편(intercept), 즉 오차보정계수를 의미한다. 다중회귀모형을 적용할 때에는 모든 입력변수에 대하여 min-max 정규화 방법을 통한 무차원화를 진행한 후 분석하였으며, 그 결과를 Table 2와 같이 나타내었다.

본 연구에서는 다중회귀모형을 활용하여 얻은 실측 데이터와 예측 데이터를 비교 분석하였다. 분석 결과 실측값과 예측값 사이의 MAE는 0.0966, R^2는 0.6712, RMSE는 0.1228, NRMSE는 0.1224로 나타났으며, 다중회귀모형에 대한 실측값과 예측값의 산포도를 Fig. 8과 같이 나타내었다.

3.3.2 랜덤포레스트 모형 적용 결과

랜덤포레스트는 딥러닝 알고리즘인 LSTM에 비해 간단한 구조를 가진다. 실제로 조정해야 하는 매개변수가 상당히 많은 LSTM에 비해 랜덤포레스트는 조정해야 하는 매개변수가 무작위 분류기 수를 의미하는 mtry만이 있다. 정수장의 수질 예측을 위한 랜덤포레스트 알고리즘을 유전 알고리즘의 적합도 검정 과정에 탑재하였다. 유전 알고리즘에서 최적화하고자 하는 랜덤포레스트 매개변수의 최소값과 최대값은 일반적으로 매개변수의 값으로 나타나는 범위에서 확장된 0과 100으로 설정하였다.

각 세대별로 가장 적합도가 양호한 매개변수를 기록하여 적합도를 최적화하였다. 최적화 결과 랜덤포레스트 기반 예측모형에서 최적의 초매개변수 mtry는 1.17로 적용되었다. 최적의 초매개변수를 적용하여 예측한 결과 실측값과 예측값 사이의 MAE는 0.0474, R^2는 0.9147, RMSE는 0.0629, NRMSE는 0.0630으로 나타났으며, 랜덤포레스트 기반 예측모형에 대한 실측값과 예측값의 산포도를 Fig. 9과 같이 나타내었다.

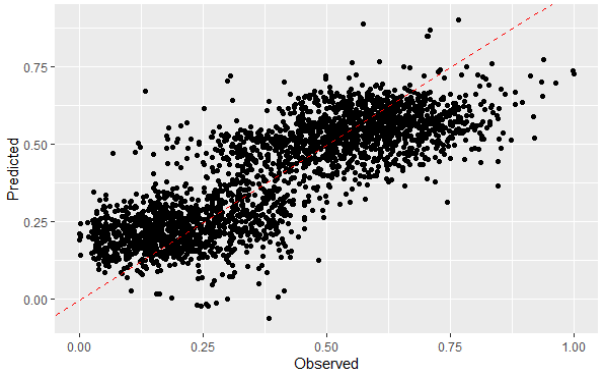


Fig. 8. Compare predicted and observed values from multiple regression models

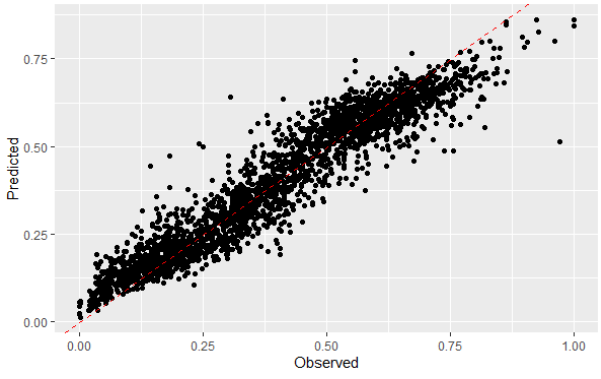


Fig. 9. Compare predicted and observed values from random forest

Table 2. Results of the multiple regression model

Inflow (a)	Conductivity (b)	Alkalinity (c)	NH_3 (d)	Temperature (e)	Sediment Pre Cl_2 (f)	Sediment Turbidity (g)	pH (h)	Intercept (ϵ)
-0.424	0.532	-0.702	-0.230	0.591	0.251	0.298	0.298	0.041

3.3.3 LSTM 모형 적용 결과

LSTM은 복잡한 AI 모형인 만큼 초매개변수와 같은 내부 구조의 조정에 상당히 민감하다. LSTM에는 본 연구에서 유전 알고리즘으로 최적화하고자 하는 units와 batch size 외에도 활성화 함수(activation function), 최적화 함수(optimization function) 등이 있다.

활성화 함수는 신경망 기반 모형에서 뉴런의 입력 신호를 다음 layer의 출력 신호로 변환하는 방식을 결정하는 함수로, 여기에는 Sigmoid, SeLU, ReLU, tanh 등이 포함된다. 입력자료가 충분할 경우 활성화 함수로 ReLU 또는 SeLU를 고려할 수 있다. 그러나 본 연구의 입력자료는 2년동안의 수질관측자료로, 수문·수리 분야에서 예측과 같은 통계적 분석을 수행할 때 충분하다고 판단하는 기준인 30년에 미달되기 때문에 LSTM에서 가장 기본적이고 범용적으로 사용하는 tanh를 적용하였다.

최적화 함수는 모형이 훈련하는 과정에서 network의 가중치를 업데이트하면서 손실을 최소화하는 역할을 하며, 대표적인 최적화 함수에는 Adam (Adaptive moment estimation)과 Nadam (Nesterov-accelerated adaptive moment estimation) 등이 포함된다. Adam은 경사하강법(Stochastic Gradient Descent, SGD)에서 매개변수 갱신 방향이 과도하게 변하는 것을 완화한 방법인 모멘텀 방법과 매개변수의 갱신 크기를 조절하는 RMSProp 방법을 혼합한 기법으로, 가장 일반적으로 사용되는 방법이다. Nadam은 Adam에 모멘텀 방법의 문제점을 개선한 NAG (Nesterov Accelerated Gradient)를 혼합한 방법이다. 본 연구에서는 최적화 함수로 관련 연구에서 가장 많이 사용하는 Adam 알고리즘을 적용하였으며, 이를 통해 학습율(learning rate)과 drop out 값을 최적화하였다.

또한, LSTM에서는 예측에 사용할 데이터의 시간 길이를 의미하는 Input length와 예측을 하고자 하는 결과의 시간을 의미하는 Output length를 정의할 수 있다. 즉, 입력자료로 활용한 각 인자들의 영향 시간을 고려하여 모형을 구축할 수 있다. 본 연구에서 사용한 수질 인자의 경우, 각 수질 인자의 변화가 다른 인자에 영향을 가지기까지의 시간이 1시간보다도 짧은 것이 특징이다. 이에 따라 Input length와 Output length를 1시간으로 설정하였다.

본 연구에서 사용된 LSTM의 모든 과정을 유전 알고리즘의 적합도 검정 과정에 구현하여 모형의 최적화를 실시하였다. 유전 알고리즘에서 units, batch size 매개변수를 최적화할 때 각 매개변수의 최소값과 최대값 범위는 각각 일반적으로 매개변수가 나타나는 범위에서 확장된 0~100으로 설정하였다. Fig. 10는 유전 알고리즘 구동 과정에서 한 세대의 epoch별 적합도 변화를 도식화 한 그래프이다. 여기서 x축은 epochs 값, loss는 모형의 학습구간에 대한 적합도, val_loss는 모형의 평가구간에 대한 적합도를 나타낸 것이다. 각 세대별로 가장 적합도가 양호한 매개변수들을 기억하고 적합도가 가장 좋은 매개변수를 적용하여 LSTM 기반 예측모형을 개발하였다.

LSTM 기반 예측모형에서 최적의 초매개변수는 units가 53.2, batch size가 16으로 적용되었다. 최적의 초 매개변수를 적용하여 예측한 결과 실측값과 예측값 사이의 MAE는 0.1019, R^2는 0.8416, RMSE는 0.1230, NRMSE는 0.1296으로 나타났으며, LSTM 기반 예측모형에 대한 실측값과 예측값의 산포도를 Fig. 11과 같이 나타내었다.

3.3.4 각 모형별 성능 비교

각 모형별 성능 지표들을 비교하였다(Table 3). 모형별 비교 결과 랜덤포레스트 기반의 예측모형이 가장 좋은 성능을 보였다. 의외로 LSTM의 성능이 가장 낮게 나온 것을 확인했

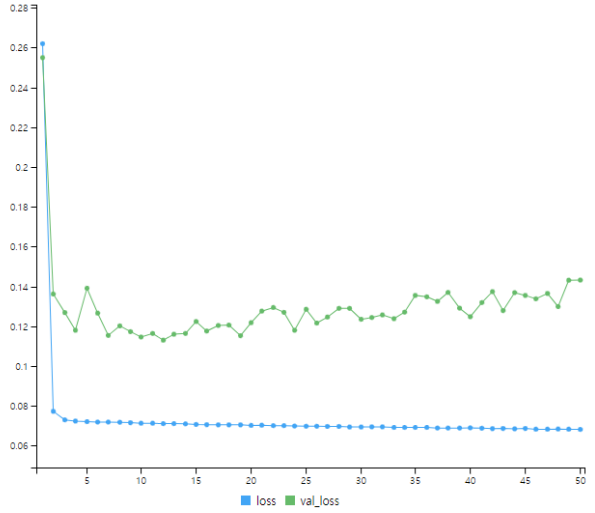


Fig. 10. Driving example for LSTM

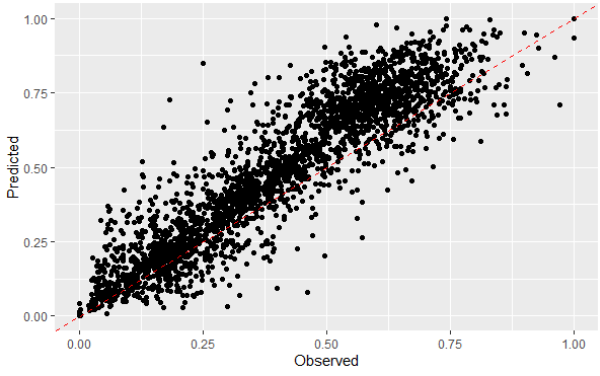


Fig. 11. Compare predicted and observed values from LSTM

Table 3. Compare performance index for each model

Model	MAE	R-Squard	RMSE	NRMSE	Rank
Multiple Regression	0.0966	0.6712	0.1228	0.1224	2
Random Forest	0.0474	0.9147	0.0629	0.0630	1
LSTM	0.1019	0.8416	0.1230	0.1296	3

는데, 이는 데이터의 특성이 원인일 것으로 판단된다. LSTM은 일반적으로 시간적 순서가 중요한 시계열 자료에서 좋은 성능을 보인다. 예를들어, 강수량과 수위와 같이 시간의 흐름에 따른 변화가 중요할 경우에 높은 성능을 보인다. 그러나 본 연구의 입력자료는 전처리 과정에서 시간 연속성의 훼손되었으며, 이러한 요인이 LSTM의 성능을 낮춘 원인이라고 판단하였다.

최종적으로 본 연구에서는 여러 인공지능 기반 예측모형과 다중회귀모형을 비교 분석한 결과, 수질 데이터와 같이 데이터 분산이 크고 차원 및 규모의 차이가 클 때, 랜덤포레스트와 같은 의사결정 트리 구조의 적용이 적절하다는 결론이 도출되었다.

4. 결 론

정수장 수처리 공정에서는 여러 단계를 거쳐야만 수돗물이 생산되고 있으며, 공정마다 수질의 농도관리가 필요하다. 특히 잔류염소는 암 유발 물질의 형성을 억제하고 미생물을 살균하는 데 필수적이지만 원수 수질, 온도, 유량, 지속시간 등의 환경적 요인에 따라 변동하므로 이를 예측하는 일은 매우 복잡하다. 높은 성능을 가지는 예측모형을 개발하기 위해서는 관측자료의 품질관리가 필수적이다. 그러나 정수장 시설에서 현재 운영하고 있는 수질관측장비는 여전히 결측치와 이상치가 많이 기록되고 있는 것이 실정이다. 관측자료의 특성으로 인해 자료의 전처리 과정에서 이상치와 결측치를 제거할 때 발생하는 시간의 연속성 훼손은 LSTM과 같은 시계열 예측모형에서 이러한 점은 치명적이다. 해당 부분은 추후 개선되어야 할 중요한 해결과제이다. 본 연구에서는 정수장 수처리 효율의 향상과 수질관리를 위한 공정제어를 위하여 처리단계별로 변화하는 잔류염소 농도예측에 인공지능 기술을 적용하여 침전지 유출수의 잔류염소 농도를 예측하는 모형을 개발하였으며, 최종적으로 도출된 결론은 다음과 같이 요약할 수 있다.

(1) 본 연구에서는 정수장에서 관리되고 있는 수질자료를 수집, 분석하여 잔류염소 관리에 필요한 침전지 유출수의 잔류염소 농도 예측에 요구되는 영향인자를 선정하였으며 이를 토대로 랜덤포레스트, 인공신경망, 그리고 다중회귀모형을 적용하여 MAE, RMSE, NRMSE 등의 평가지표를 통해 예측성능을 비교, 평가하였다. 결과에서는 랜덤포레스트 기반 예측모형이 가장 우수한 예측력을 보였으며, 이어서 다중회귀모형과 LSTM 순으로 나타났다. 이는 랜덤포레스트 기반 예측모형이 예측 목표값인 침전지 유출수 잔류염소 농도에 영향을 주는 다양한 요소들을 효과적으로 반영하고 있음을 나타낸다.

(2) LSTM 기반 예측모형은 예측부분에서 강력한 성능을 보일 것으로 예상한것과 달리 성능이 떨어지는 것을 확인하였다. LSTM은 강수량이나 수위와 같이 시간적 순서가 중요한 시계열 자료를 예측할 때 높은 성능을 보인다. 그러나, 본 연구의 입력자료 전처리 과정에서 시계열의 시간 연속성 훼손이 발생하였기 때문에 LSTM의 성능이 떨어지는 것으로 판단하였다. 이러한 결과는 정수장 수질 관측 자료에 대한 품질관리가 상당히 중요하며, 추후 해결되어야 할 중요한 문제임을 의미한다.

(3) 인공지능을 활용한 예측모델을 효과적으로 구축하고 운영하기 위해서는 적절한 자료의 기간과 시간 간격을 선정하는 것이 중요하다. 특히, 잔류염소 특성을 반영할 수 있는 인자의 선택이 필수적이다. 모형의 성능은 정수처리 과정에서의 수량과 수질 데이터의 처리 방식에 영향을 받으며, 이는 관측자료의 품질에 따른 영향이 크다는 의미이다. 모형의 매개변수 선택 또한 성능에 중요한 영향을 가지며, 본 연구에서는 매개변수 선택에 유전 알고리즘을 사용하였다.

(4) 관측자료의 해석은 수처리 시설의 운영관리에 필수적인 요소로서 이 과정은 복잡성과 시간 소모가 많은 특성을 가지고 있다. 이러한 특성을 AI를 통해 보완하는 것은 충분히 유용하다. 이에 따라 데이터 수집, 정리, 전처리, 알고리즘 선정 및 최적화를 포함한 전 과정의 지능화가 필요한 일이며, 본 연구의 결과를 통해 수처리 시설의 운영관리 효율 향상과 수돗물 품질의 개선을 기대할 수 있을 것으로 판단된다.

감사의 글

이 논문은 행정안전부 재난피해 복구역량강화 기술개발사업의 지원을 받아 수행된 연구임(2021-MOIS36-002).

Conflicts of Interest

The authors declare no conflict of interest.

References

Amit, Y., and Geman, D. (1997). "Shape quantization and recognition with randomized trees." *Neural Computation*, Vol. 9, No. 7, pp. 1545-1588.

Bates, S., Hastie, T., and Tibshirani, R. (2023). "Cross-validation: What does it estimate and how well does it do it?." *Journal of the American Statistical Association*, pp. 1-12.

Breiman, L. (2001). "Random forests." *Machine learning*, Vol. 45, No. 1, pp. 5-32.

Cohen, I., Huang, Y., Chen, J., and Benesty, J. (2009). "Pearson correlation coefficient." *Noise Reduction in Speech Processing*, Edited by Benesty, J., and Kellermann, W., Springer, Berlin, Heidelberg, Germany, pp. 1-4.

Diamantopoulos, A., and Siguaw, J.A. (2006). "Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration." *British Journal of Management*, Vol. 17, No. 4, pp. 263-282.

Gulzar, A., Ihsanullah, I., Mu, N., and Mika, S. (2022). "Applications of artificial intelligence in water treatment for optimization and automation of adsorption processes: Recent advances and prospects." *Chemical Engineering Journal*, Vol. 427, No. 1, 130011.

Han, H., Choi, C., Jung, J., and Kim, H.S. (2021). "Application of sequence to sequence learning based LSTM model (LSTM-s2s) for forecasting dam inflow." *Journal of Korea Water Resources Association*, Vol. 54, No. 3, pp. 157-166.

Ho, T.K. (1998). "The random subspace method for constructing decision forests." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, pp. 832-844.

Hochreiter, S., and Schmidhuber, J. (1997). "Long short-term memory." *Neural Computation*, Vol. 9, No. 8, pp. 1735-1780.

Holland, J.H. (1975). *Adaptation in natural artificial systems*. The MIT Press, MA, U.S., pp. 1-19.

Jun, H.B., Lee, Y.J., Lee, B.D., and Lee, J.D. (2001). "Effects of pre-chlorination on diatoms coagulation." *Journal of Korean Society on Water Environment*, Vol. 17, No. 3, pp. 347-355.

Jung, S.H., Lee, D.O., and Lee, K.S. (2018). "Prediction of river water level using deep learning open library." *Journal of Korean Socioty Hazard Mitigation*, Vol. 18, No. 1, pp. 1-11.

Kang, G.W., Park, C.Y., and Kim, J.H. (1992). "Nonlinear prediction of river runoff using pattern recognition methods." *Journal of Korea Water Resources Association Conference*, pp. 196-202.

Kim, B.J., Choi, M.W., Kim, G.H., and Kim, H.S. (2016a). "Evaluation and analysis of characteristics for Hazen-Williams C based on measured data in multi-regional water supply systems." *Journal of Korean Society of Water and Wastewater*, Vol. 30, No. 2, pp. 197-206.

Kim, B.J., Kim, G.H., and Kim, H.S. (2016b). "Statistical analysis of Hazen-Williams C and influencing factors in multi-regional water supply system." *Journal of Korea Water Resources Association*, Vol. 49, No. 5, pp. 197-206.

Kim, D.H. (2022). *Development of flood water level forecasting and flood damage risk assessment method for river basin using AI-based hybrid moded*. Ph. D. Dissertation, Inha University, pp. 34-37.

Kim, D.H., Lee, K.S., Hwang-Bo, J.G., Kim, H.S., and Kim, S.J. (2022a). "Development of the method for flood water level forecasting and flood damage warning using an AI-based model." *Journal of the Korean Society of Hazard Mitigation*, Vol. 22, No. 4, pp. 145-156.

Kim, H.S., Jeong, G.H., Kim, E.S., and Kim, J.H. (2001). "Estimation of mean and variance for NH_3-N data of Puyeo Intake." *Journal of Korea Water Resources Association*, Vol. 34, No. 4, pp. 357-364.

Kim, J.H., Lee, K.H., Kim, S.J., and Kim, K.H. (2022b). "Machine learning model for residual chlorine prediction in sediment basin to control pre-chlorination in water treatment plant." *Journal of Korea Water Resources Assocition*, Vol. 55, No. S-1, pp. 1283-1293.

Kim, J.S. (2021). *Development of prediction and warning technique of heavy rain damage risk based on ensemble machine learning and risk matrix*. Ph. D. Dissertation, Inha University, p. 56.

Kim, J.W., Kim, Y.S., Kang, N.R., Jung, J.W., and Kim, S.J. (2014). "Risk assessment for water quality of a river using QUAL2E model." *Journal of Wetlands Researh*, Vol. 16, No. 3, pp. 441-450.

Kim, S.W., and Kim, H.S. (2007a). "Neural network-genetic algorithm model for modeling of nonlinear evaporation and evapotranspiration time series 1. Theory and application of the model." *Journal of Korea Water Resources Association*, Vol. 40. No. 1, pp. 73-88.

Kim, S.W., and Kim, H.S. (2007b). "Neural network-genetic algorithm model for modeling of nonlinear evaporation and evapotranspiration time series 2. Optimal model construction by uncertainty analysis." *Journal of Korea Water Resources Association*, Vol. 22. No. 2, pp. 149-169.

Kwon, S.D. (2015). "Exploring a way to overcome multicollinearity problems by using hierarchical construct model in structural equation model." *Journal of Information Technology Applications & Management*, Vol. 40. No. 1, pp. 89-99.

Kwon, S.H., Lee, J.W., and Chung, G,H. (2017). "Snow damages

estimation using artificial neural network and multiple regression analysis." *Korean Society of Disaster & Security*, Vol. 17, No. 2, pp. 315-325.

Kyoung, M.S., Kim, S.D., Kim, H.S., and Park, S.K. (2006). "Statistical water quality monitoring network design of Kyung-An Stream." *Journal of Civil Engineering*, Vol. 26, No. 3B, pp. 291-300.

Le, X.H., Ho, H.V., Lee, G., and Jung, S. (2019). "Application of long short-term memory (LSTM) neural network for flood forecasting." *Water*, Vol. 11, No. 7, 1387.

Lee, H.H., Jang, S.B., Hong, S.T., and Chun, M.G. (2014). "Intelligent controller for constant control of residual chlorine in water treatment process." *Journal of Korean Institute of Intelligent Systems*, Vol. 23, No. 1, pp. 147-154.

Lee, K.H., Kim, J.H., Lim, J.L., and Chae, S.H. (2007). "Prediction models of residual chlorine in sediment basin to control pre-chlorination in water treatment plant." *Journal of Korean Society of Water and Wastewater*, Vol. 21, No. 5, pp. 601-607.

Lee, S.M., Baek, S.W., Lee, J.H., Kim, K.T., Kim, S.J., and Kim, H.S. (2023a). "Development of disaster severity classification model using machine learning technique." *Journal of Korea Water Resources Assocition*, Vol. 56, No. 4, pp. 261-272.

Lee, S.M., Wang, W.J., Kim, D.H., Han, H.C., Kim, S.J., Kim, H.S. (2023b) "Establishing meteorological drought severity considering the level of emergency water supply." *Journal of Korea Water Resources Assocition*, Vol. 56, No. 10, pp. 619-629.

Liaw, A., and Wiener, M. (2002). "Classification and regression by randomforest." *R News*, Vol. 12, No. 3, pp. 18-22.

Lim, J.O. (2019). *Estimation of flood damage based on multi-dimensional flood damage assessment and multiple regression analysis: A case study for the PyeongChang River Basin*. Master's Thesis, Inha University, pp. 20-22.

Lowe, M., Qin, R., and Mao, X. (2022). "A review on machine learning, artificial intelligence, and smart technology in water treatment and monitoring." *Water*, Vol. 14, No. 9, 1384.

Ministry of Environment (ME) and Korea Environmental Industry & Technology Institute (KEITI) (2021). *2020 water & wastewater R&D technology trends report*.

Ngo, T.H.D., and La Puente, C.A. (2012). "The steps to follow in a multiple regression analysis." *Proceedings of the SAS Global Forum*, Florida, FL, U.S., pp. 22-25.

Petter, S., Straub, D., and Rai, A. (2007). "Specifying formative constructs in information systems research." *MIS Quarterly*, Vol. 31, No. 4, pp. 623-656.

Waterworks Headquarters Incheon Metropolitan City (WHIM) (2023). *2023 Incheon sky water quality report*, pp. 16-22.

Yoon, J.Y., Byun, S.J., and Choi, Y.S. (2001) "Importance of Pre-chlorination practices and structures of clearwell in estimating disinfection capabilities in water treatment plants." *Journal of Korean Society on Water Environment*, Vol. 17, No. 3, pp. 327-337.

Zhang, Q., and Stanley, S.J. (1999). "Real time water treatment process control with artificial neyral networks." *Journal of Environmental Engineering*, Vol. 125, No. 2, pp. 153-160.