J. Korea Water Resour. Assoc. Vol. 55, No. S-1 (2022), pp. 1283-1293
doi: 10.3741/JKWRA.2022.55.S-1.1283
pISSN 2799-8746
eISSN 2799-8754

Machine learning model for residual chlorine prediction in sediment basin to control pre-chlorination in water treatment plant

Kim, Juhwan[a] · Lee Kyunghyuk[b] · Kim, Soojun[c] · Kim, Kyunghun[d]*

[a]Research Professor, Department of Civil Engineering, Inha University, Incheon, Korea
[b]Head Researcher, Water Use Efficiency Research Center, Korean Water Resources Corp., Daejeon, Korea
[c]Asscoiate Professor, Department of Civil Engineering, Inha University, Incheon, Korea
[d]Ph.D Student, Department of Civil Engineering, Inha University, Incheon, Korea

Paper number: 22-077
Received: 15 September 2022; Revised: 16 November 2022; Accepted: 18 November 2022

Abstract

The purpose of this study is to predict residual chlorine in order to maintain stable residual chlorine concentration in sedimentation basin by using artificial intelligence algorithms in water treatment process employing pre-chlorination. Available water quantity and quality data are collected and analyzed statistically to apply into mathematical multiple regression and artificial intelligence models including multi-layer perceptron neural network, random forest, long short term memory (LSTM) algorithms. Water temperature, turbidity, pH, conductivity, flow rate, alkalinity and pre-chlorination dosage data are used as the input parameters to develop prediction models. As results, it is presented that the random forest algorithm shows the most moderate prediction result among four cases, which are long short term memory, multi-layer perceptron, multiple regression including random forest. Especially, it is result that the multiple regression model can not represent the residual chlorine with the input parameters which varies independently with seasonal change, numerical scale and dimension difference between quantity and quality. For this reason, random forest model is more appropriate for predict water qualities than other algorithms, which is classified into decision tree type algorithm. Also, it is expected that real time prediction by artificial intelligence models can play role of the stable operation of residual chlorine in water treatment plant including pre-chlorination process.

Keywords: Machine learning, Residual chlorine prediction, Random forest, Multilayer perceptron, Long short term memory

정수장 전염소 공정제어를 위한 침전지 잔류염소농도 예측 머신러닝 모형

김주환[a] · 이경혁[b] · 김수전[c] · 김경훈[d]*

[a]인하대학교 사회인프라공학과 연구교수, [b]한국수자원공사 상하수도연구소 수석연구원, [c]인하대학교 사회인프라공학과 부교수,
[d]인하대학교 토목공학과 박사과정

요 지

본 연구는 정수장의 수처리 공정에서 계측되고 있는 수량 및 수질데이터의 활용과 수처리 공정제어의 지능화를 위한 것으로 정수장에서 전염소 공정이 수반되는 처리공정에서 침전지 유출수 잔류염소농도 안정화를 위하여 이를 추정할 수 있는 모형을 구축하고자 하였다. 정수장 침전지 유출수의 잔류염소농도를 예측하기 위하여 중회귀모형과 인공지능 알고리즘 중 다층퍼셉트론 신경망, 랜덤포레스트 및 장단기기억(Long Short Term Memory; LSTM) 모형을 활용하였고 그 결과를 비교, 평가하였다. 모형의 입력변수로는 전염소 공정이 도입된 정수장에서의 잔류염소농도, 수온, 탁도, pH, 전기전도도, 유량, 알칼리도 등이 사용되었고 전염소에 따른 침전지의 안정적 운영을 위해 요구되는 침전지 잔류염소농도를 출력변수로 구성하였다. 적용 결과에서는 랜덤포레스트 모형이 가장 양호한 결과를 보여 주었으며 다음으로 LSTM, 다층퍼셉트론 신경망 순으로 나타났다. 수학적 모형인 중회귀모형은 적합도 측면에서 가장 낮은 결과를 보여 주었는데, 이는 수량과 수질데이터의 수치적인 규모나 차원의 차이뿐만 아니라 계절별 수질특성에 따라 염소소비 특성이 매우 다양하게 반응하기 때문으로 판단된다. 따라서 정수장 수처리 공정에서 인공지능 알고리즘의 적용을 위해서는 랜덤포레스트와 같이 의사결정 트리구조의 도입과 적용이 타당한 것으로 나타났다. 본 연구에서 분석된 결과를 근거로 전염소 공정이 도입된 정수장 수처리 공정에서 염소주입량을 실시간으로 예측 가능하게 함으로써 침전지 유출수에서 잔류염소농도를 일정하게 유지하는데 기여할 수 있을 것으로 기대된다.

핵심용어: 머신러닝, 잔류염소농도 예측, 랜덤포레스트, 다층퍼셉트론 신경망, 장단기 기억모형

***Corresponding Author.** Tel: +82-32-872-8729
E-mail: tgb611@naver.com (Kim, Kyunghun)

1. 서 론

국내 정수장 정수처리의 여러 공정 가운데 중요한 소독공정은 대부분 염소를 이용하여 실시하고 있으며 염소처리 공정은 주입지점을 기준으로 전염소, 중염소, 후염소 공정으로 분류된다. 전염소 처리는 정수처리에서 초기공정으로 물의 잔류염소농도 증가, 생물 멸균, 페수 냄새와 색깔 제거, 철, 망간의 제거 등을 목적으로 원수에 함유된 조류를 취수장이나 도수관 또는 착수정에서 산화시키거나 암모니아성 질소와 같은 피산화성 물질의 제거를 위하여 주입한다. 이는 원수를 취수한 후 응집, 침전과 모래 여과과정을 진행하기 전에 염소를 주입하여 불순물을 제거하기 위한 공정으로, 특히 원수에 유기물이 존재하면 유리 잔류염소와 반응하여 발암물질인 트리할로메탄(Trihalomentahne)이 생성되므로 전염소 처리로서 이를 제거하여야 한다. 유기물이나 세균이 많은 오탁 하천에서 원수를 취수할 때나 철, 망간을 포함하여 여과 전에 석출 분리할 필요가 있을 때, 급속 모래여과에서 생물멸균처리 등에 이용되고 있으며 취수정과 착수정에서는 소독 이외의 목적으로 전염소처리가 행해지며 살균을 목적으로 하는 후염소처리보다 전에 주입되기 때문에 전염소처리라고 한다. 착수정 이후 혼화지, 플록형성지, 침전지 그리고 여과지 등의 구조물에서 수온이 높은 시기에 부착조류의 사멸이나 침전슬러지의 부상억제 등을 목적으로 주입되는 것을 중간염소처리라고 한다. 이와 대조적으로 여과와 같은 최종 입자 제거공정 이후에 살균소독을 목적으로 실시하는 염소처리를 후염소처리라고 한다. 만약, 입자가 충분히 제거되지 않으면 염소살균 효과가 감소되므로 살균을 목적으로 하는 후염소처리는 입자제거 공정 이후에 실시된다.

전염소 및 후염소처리 외에도 정수장의 상황에 따라 중염소처리, 과염소처리, 탈염소처리 및 재염소처리를 시행하는 경우도 있다. 중염소처리의 경우, 정수처리 공정의 중간지점 또는 다양한 공정들의 사이 지점 염소를 주입하는 것으로 주로 여과 전에 염소를 주입함으로써 여과지 세균부하 경감 및 여과 지속시간을 늘려 준다. 또한 주입방법은 간헐적 또는 연속적으로 주입할 수 있는데 연속적으로 주입하는 경우에는 소독부산물 생성의 원인으로 작용하기도 한다.

후염소 공정의 경우 최종 정수 처리된 물에 대한 소독 및 배급수 관망에서 염소의 잔류 농도를 유지하기 위하여 실시한다. 정수장 원수에 대해 염소를 주입하는 전염소 공정의 경우 국내 정수장의 약 43%정도가 실시중인 것으로 조사되었으며 전염소 공정을 실시하는 목적은 정수장의 상황마다 다양하다(Yoon *et al*., 2001). 우선 전염소는 조류의 제거를 위해 적용될 수 있다. 전염소 처리 후 일부 조류에서는 응집, 침전 공정에서 제거율이 높아질 수 있다(Jeon *et al*., 2001). 또한 전염소는 원수 내 철,망간, 암모니아 등의 처리에 적용될 수 있다. 하지만 실제 정수장에서 전염소 공정을 실시하는 주된 목적은 침전지에서 생기는 부착성 조류 및 이끼류를 제거 하는 것이며, 또한 후염소 공정에서 일정한 잔류염소농도를 안정적으로 유지시키기 위해 전염소 공정에서 염소 요구량을 미리 감소시키는 목적이 있다(Maneual and Jean, 1999). 그러나 전염소 공정으로 침전지까지 잔류염소농도를 일정하게 유지하기가 현실적으로 매우 어렵다. 왜냐하면 원수의 수질변화 및 수온 등의 환경의 변화에 의해 염소 소비량이 변하며, 유입 원수의 유량이 일정하지 않을 경우 염소 주입지점부터 침전지 유출부분까지의 수리학적 체류시간 또한 변하기 때문에 이를 예측하기가 어려운 실정이다.

인공지능은 인간과 같이 생각하고 학습하고 판단하는 인간의 사고회로를 모방하여 만든 컴퓨터 모형이며, 인공신경망은 생물학적 뉴런(neuron)의 동작원리와 뉴런간의 연결 관계를 모방하여 만든 알고리즘이다. 인공신경망은 McCulloch와 Pitts (1943)가 수학적 기법으로 처음 신경망의 기초논리를 연구하였고, 이후 Rosenblatt (1957)가 많은 기계학습(Machine Learning)과 심층학습(Deep Learning)의 기초가 되는 신경망 알고리즘인 퍼셉트론(Perceptron)을 제안하였다. 그러나 그 당시 연구됐었던 신경망으로는 복잡한 논리회로 연산이 힘들다는 등의 이유로 Minsky와 papert (1969)의 머신러닝에 관한 연구 이후로 관련 연구가 침체되어 암흑기를 맞았으나 1980년대 Rumelhart *et al*.(1986)에 의해 출력값을 역으로 전파하여 오차의 경사를 줄여나가는 역전파(back propagation) 알고리즘이 개발되면서 다시 각광받기 시작하였다. 이후 빅데이터 수집 및 처리에 관한 다양한 기술들이 개발됨과 동시에 사람의 학습능력을 모방하는 딥러닝 기법이 소개되고, 순환신경망(Recurrent Neural Network, RNN), 합성곱신경망(Convolutional Neural Network, CNN) 등의 새로운 알고리즘의 개발로 영상인식, 음성인식, 자율주행자동차 개발 등 다양한 분야에서 적용성을 인정받고 있다. 인공지능 이론을 정수처리 공정에 적용한 사례는 간헐적으로 이루어져 왔으며 기존 정수처리 공정에서 응집제 등 약품주입량 결정을 위한 것으로 Qing and Stephen (1999)은 2000여개의 정수장 자료를 활용하여 다층퍼셉트론 신경망을 도입함으로써 정수장의 실시간 운영에 활용하고자 하였고, 국내에서는 전염소처리공정이 도입된 동일 정수장을 대상으로 침전지 잔류염소 예측을 위한 중회귀모형과 다층퍼셉트론신경망을 적용한 사례(Lee *et al*., 2007)가 있었으며, 본 연구는 이에 대한 후속 연구로서 현재 다양하게 적용되고 있는 인공지능 알고리즘을 확대, 적용하는 것으로 볼 수 있다. 즉, 전염소 공정의 후속공정인 침전지의

잔류염소농도의 안정적 관리를 위하여 수행되는 전염소 주입량의 제어를 위하여, 현장에서 연속측정이 가능하면서 염소분해속도에 영향을 주는 인자들을 이용, 기존의 결과를 재분석하고 다층퍼셉트론 신경망과 더불어 의사결정트리 형태의 랜덤포레스트와 LSTM 모형 등 인공지능에 이론을 추가적으로 검토, 적용하여 성능을 비교, 평가함으로써 인공지능 정수장에서 활용될 수 있도록 하고자 하였다.

2. 적용이론

2.1 중회귀 모형

중회귀 모형(Multiple Regression Model)은 2개 이상의 독립변수를 사용하여 종속변수 y를 산정하는 방법으로 선형다중회귀모형과 지수형 다중회귀모형 등 여러 가지 형태의 관계가 가능하며 일반식은 다음과 같이 표현된다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

여기서, $\beta_0, \beta_1, \cdots, \beta_n$는 중회귀 분석을 통해 산정되는 매개변수이며, $x_1, x_2, \cdots, x_n$는 독립변수, y는 종속변수를 나타낸다. 또한, 회귀모형에서는 독립변수들 간의 상관관계가 높을 경우 다중공선성의 문제가 발생한다(Jung *et al*., 2018). 다중공선성을 판별하기 위하여 분산팽창요인(Variance Inflation Factor)을 통해 평가하고, 분산팽창요인이 10 이상일 때 다중공선성이 높다고 판단한다(Kim *et al*., 2020). 다중공선성을 해결하기 위하여 통계적 유의성이 높은 변수를 추출할 수 있는 변수선택법을 고려할 수 있고, 변수선택법으로는 가장 예측력이 좋은 단계적 선택법(Stepwise), 전진선택법(Forward), 후방소거법(Backward)이 있다.

2.2 다층퍼셉트론 신경망

신경망은 인간의 두뇌를 수학적으로 모사한 모형이며, 가장 광범위하게 사용되어온 신경망 모형은 Rumelhart *et al*. (1986)에 의해 소개된 다층퍼셉트론(Multi-Layer Perceptron, MLP) 신경망이다. 신경망 이론을 이용한 모형의 개발은 회귀모형과 같은 특별한 구조나 매개변수의 산정 및 자료의 변화 등이 필요치 않고 자료의 축적에 따라 모형의 능력을 향상시킬 수 있는 장점을 가지고 있다(Kim, 1993).

다층퍼셉트론 신경망은 여러 개의 처리소자를 각 계층, 즉 입력자료를 받아들이는 입력층과, 결과를 나타내는 출력층 그리고 두 개의 층 사이에 중간층을 두어 각 패턴마다 필요한 정보를 학습시킬 수 있도록 한 것이다(Kang *et al*., 1992). 각 층을 구성하는 뉴런은 뉴런간의 연결강도를 합해주는 결합함수(Summation Function)와 자료의 범위에 따라 뉴런의 동작범위를 결정해 주는 활성화함수(Activation Function)로 구성되며 일반적으로 시그모이드(Sigmoid) 함수가 사용된다. 입력자료와 출력자료의 비선형 관계를 구축하기 위한 신경망 즉, 다층퍼셉트론 신경망의 학습방법으로는 여러 가지 알고리즘이 있으나 가장 일반적으로 시용되는 학습알고리즘은 오차역전파(Error Back Propagation, BP) 알고리즘이 사용된다(Kumar *et al*., 2005).

역전파 알고리즘은 입력패턴과 출력패턴의 집합으로 구성되는 학습패턴을 신경망에 입력하여 계산값과 관측값의 차이가 설정된 오차의 범위까지 최소화 하도록 모형의 매개변수인 각 층간의 가중치를 변화시켜 가는 방법으로 Eq. (2)와 같이 제곱오차의 합이 최소가 되는 방향으로 반복적으로 진행되어 신경회로망 모형의 연결강도를 최적화시킴으로써 모형을 적응시켜 나가는 것이다(Lisboa, 1992).

$$E_p = \frac{1}{2}\sum_{k=1}^{p}(y_{pk} - o_{pk})^2 \quad (2)$$

여기서, y_{pk}는 p번째 입력패턴에 대한 k번째 처리소자의 관측값이며 o_{pk}는 p번째 입력패턴에 대한 출력층의 k번째 처리소자의 출력값이다. 하나의 패턴 p에 대한 오차 E_p와 모든 패턴에 대한 총오차를 최소화 시킨다. 총오차를 최소화시키기 위한 연결강도의 조정은 p번째 패턴의 입력값 x_p와 출력층의 델타값 δ_{pk}에 의하여 다음과 같이 표현된다.

$$\Delta W = \eta \delta_{pk} x_p \quad (3)$$

각 층간의 연결강도는 위 식으로 표현되는 현재 단계에서의 조정량과 전 단계에서의 조정량 ΔW에 모멘텀(momentum) 상수 α를 곱하여 다음 Eq. (4)와 같이 조정된다.

$$\Delta W(t+1) = \eta \delta_{pk} x_p + \alpha \Delta W(t) \quad (4)$$

여기서, t는 반복회수, η는 학습율, α는 모멘텀 상수를 나타내며 반복과정에서 다음단계의 반복을 위해 조정된 연결강도는 다음 Eq. (5)에 의해 계산된다.

$$W(t+1) = W(t) + \triangle W(t) \quad (5)$$

또한 중간층과 출력층간 연결강도의 조정도 입력층과 중간층간의 조정방식과 유사하게 진행되며 이 과정은 신경망이 안정될 때까지 또는 목적함수가 허용오차의 범위 내에 이를 때까지 반복된다(Tiwari and Chatterjee, 2010).

2.3 랜덤포레스트 모형

랜덤포레스트(Random Forest, RF)는 지도머신러닝 알고리즘으로 의사결정나무 모형 여러 개를 훈련시켜 그 결과를 종합해 예측하는 앙상블 알고리즘으로 정확성, 단순성 및 유연성으로 인해 가장 많이 사용되고 있는 모형 중 하나이며 다양한 데이터의 분류 및 회귀분석에 사용되고 있다. Breiman (1996; 2001)에 의해 제시된 RF모형은 앙상블 학습으로 부트스트랩(Bootstrap)방식을 이용하여 다수의 표본을 생성하고, 의사결정 트리(Decision Trees) 모형을 적용하여 그 결과를 종합하는 방법으로 Fig. 1에서처럼 다수의 표본으로 의사결정트리를 구축하여 이중에서 오차가 가장 작은 모델을 선택함으로써 수행된다.

부트스트랩 표본생성은 무작위로 이루어지며 의사결정나무 모형을 적합 시켜 가는 과정에서 각 노드(node)의 설명변수를 선택하는 과정도 무작위성이 더해져 이를 최대로 주기 위해 부트스트랩 표본을 생성할 뿐만 아니라 각 의사결정마디에서의 설명변수에 대해 무작위로 표본을 수집하는 방식이다. 무작위성 확보를 위하여 다음 2가지 방법이 일반적으로 사용된다. 첫째, 모든 의사결정나무를 부트스트랩 방식으로 추출된 표본에 적합시켜, 의사결정나무 숲(Forest)을 구성한다. 두 번째는 마디를 분기하는 과정에서 변수에 무작위성을 부여한다. 즉 모든 분기과정에서 변수를 선택할 때 p개의 모든 변수를 고려하는 것이 아니고 m개의 변수를 무작위로 사용한다. 이 과정에서 전체 훈련자료 중에서 N개의 부트스트랩 표본을 얻은 후 남는 표본이 있는데 이것을 'OOB (Out-Of-Bag)'자료라 하며, 이 데이터를 이용해서 과적합 문제를 방지하기 위해서 분류오차(Generalization Error)와 변수중요도지수(Variable Importance)를 구한다.

Breiman (2001)은 이러한 무작위성이 최대로 되면 의사결정나무들 간에 상관관계가 줄어들게 되고 이로 인해 예측오차가 줄어드는 특성을 갖게 되어 의사결정나무가 수가 많아도 RF모형은 과적합문제가 발생하지 않는 것을 확인하였다. 다만, 의사결정나무에서처럼 성장, 가지치기 등과 같은 조정모수(Tuning parameter)가 없다는 장점을 가지나 부트스트랩 표본을 몇 개로 할 것인지, 각 마디에서 설명변수의 개수를 몇 개로 할 것인지, 결과를 종합할 때 선형결합을 어떻게 할 것인지 등을 결정하여야 한다.

2.4 LSTM 모형

LSTM (Long Short Term Memory)모형은 스스로를 반복하면서 이전 단계에서 얻은 정보가 지속되도록 하는 시간 의존성을 가진 RNN의 특별한 종류로 긴 의존 기간을 필요로 하는 학습을 수행할 능력을 갖고 있다. LSTM은 Hochreiter and Schmidhuber (1997)에 의해 소개되었고, 그 후에 여러 추후 연구로 계속 발전되어 여러 분야의 문제를 해결해 왔으며 지금도 널리 사용되고 있다. 긴 의존기간의 문제를 피하기 위하여 명시적으로(explicitly) 설계된 LSTM 모형은 기존의 RNN이 출력과 먼 위치에 있는 정보를 기억할 수 없다는 단점을 보완하여 장/단기 기억을 가능하게 설계한 신경망의 구조를 말한다.

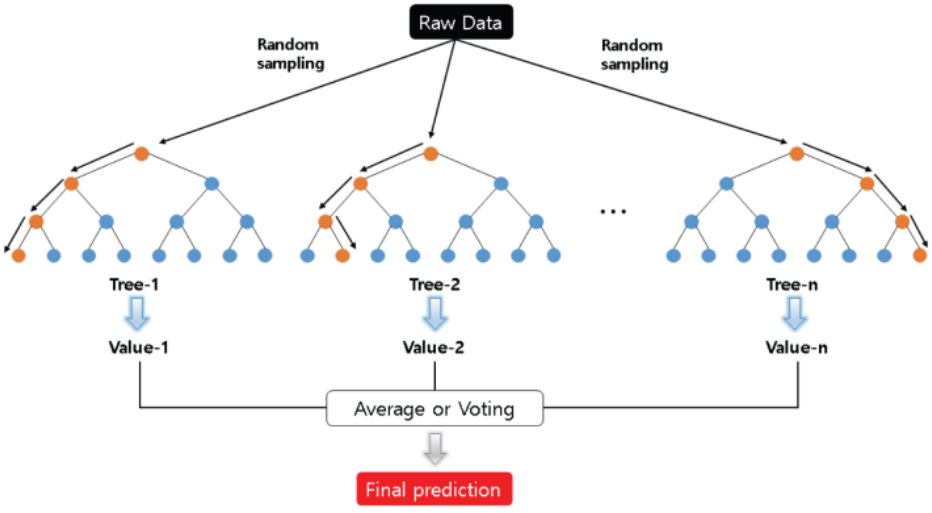


Fig. 1. Concept of random forest

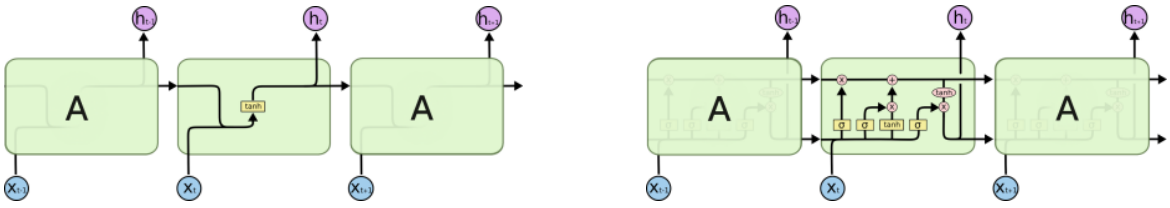
Fig. 2(a)에서 RNN은 뉴런과 유사한 노드들이 이룬 레이어가 연속되어 있는 구조를 가지며 자료가 입력되면 이전 단계에서 얻은 정보가 지속할 수 있도록 하는 체인 구조로 tanh층을 모듈로 사용한다. 각 노드간 연결관계는 가중치로 설명될 수 있으며 신경망 외부로부터 값을 입력받는 입력노드, 결과값을 산출하는 출력노드, 입력노드에서 출력노드 사이에 존재하는 중간노드로 구분된다.

Fig. 2(b)에서는 LSTM의 경우 RNN에는 없는 3개의 게이트(Gate)가 존재하며 이는 현 시점의 입력과 이전 시점의 중간층을 합친 후 tanh를 통과하고 출력되는 현 시점의 출력인 중간층 정보가 다음 시점에 반영됨으로써 과거의 정보에 의존하는 역할을 하게 되는 구조를 갖는다.

3. 대상자료 및 분석

정수장 침전지의 잔류염소는 일정하게 유지하여야 하는데, 이는 전염소 주입량의 영향을 받는다. 침전지의 잔류염소를 일정하게 유지하기 위해서 유량에 따라서 전염소 주입량을 제어하고 있지만, 일정 수준을 유지하기에 어려운 실정이다. 따라서 본 연구에서는 국내 K정수장을 대상으로 전염소 주입 시 수질 및 수량 조건 변화에 따라 침전지 유출수의 잔류염소 농도를 예측하기 위한 것으로 인공지능 이론을 도입 적용함으로써 각 모델별 예측의 정확성을 비교, 평가하고자 하였다. 대상 정수장은 댐 원수를 취수하는 시설용량 700,000 m^3/day 규모이며, 하천에서 취수하는 원수와 비교해 볼 때 상대적으로 원수의 수질변화가 작으며 전염소 공정의 도입으로 침전지에서 잔류염소를 제어하기 위하여 실시간으로 염소 주입량을 정수장 운영자의 경험에 의존하여 주입량을 조절하고 실정이다. 따라서 수중에 잔류하는 염소의 분해에 영향을 미치는 인자는 다양하나 이 중에서 대표적으로 탁도, 유기물 농도, pH, 수온, 체류시간유량, 염소주입량 등의 인자를 들 수 있다(Uber *et al.*, 2003).

본 연구에서는 정수장 현장에서 모니터링이 가능한 데이터를 수집하여 정수처리 공정에서 염소변화 특성을 파악할 수 있는 수질데이터를 분석하였다. 모니터링이 가능한 수질



(a) Structure of RNN (b) Structure of LSTM

Fig. 2. Structures of RNN and LSTM

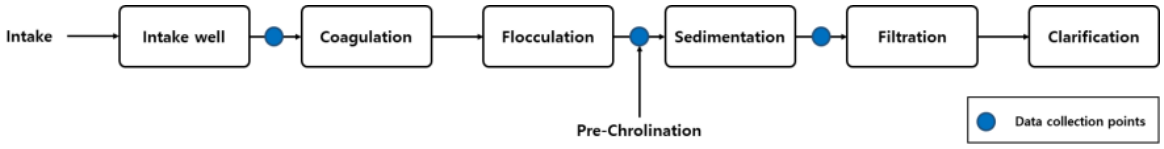


Fig. 3. Data collection points of study water treatment plant

Table 1. Basic statistics of collected data

Parameter	Mean	STD	Variance	Min.	Max.	Range	Correlation
Alkalinity (mg/L as CaCO3)	19.12	2.76	7.62	15.00	25.00	10.00	0.073
Flow Rate (m^3/day)	9211.06	2800.26	7841446.59	3113.21	13423.83	10310.62	0.392
Turbidity (NTU)	3.61	3.71	13.78	0.43	43.62	43.19	-0.345
Temperature (℃)	18.21	4.06	16.48	8.03	23.80	15.77	-0.516
pH	6.78	0.27	0.07	6.08	7.37	1.29	0.376
Conductivity (μmhos/cm)	71.28	9.25	85.53	45.64	99.92	54.28	0.069
Pre-chlorine Doasge (mg/L)	1.63	0.50	0.25	0.68	2.99	2.31	-0.527
Residual Chlorine (mg/L)	0.35	0.15	0.22	0.00	0.98	0.98	1.000

데이터는 1시간 단위의 수온, pH, 탁도, 전기전도도, 알칼리도, 유입유량 및 전염소 주입량과 침전지 염소농도로서 각 데이터는 Fig. 3에 나온 자료 수집 구간들에서 취득하였으며, 통계적 특성은 Table 1에서 볼 수 있다.

수집된 데이터는 2006년 5월 19일부터 2006년 12월 31일까지 계측된 5,448개의 시간데이터로 유지관리 기간, 일부 항목의 결측, 전후 운영과정을 고려하여 비정상적이라고 판단되는 일부를 제외시켜 5,298개의 시간데이터를 활용하였다. Fig. 4에서는 사용된 전체 자료의 전염소 주입농도와 침전지 잔류염소농도의 변화를 볼 수 있도록 도시한 것으로 침전지 잔류염소 농도는 전염소 주입농도에 가장 큰 영향을 받으며 상관계수는 Table 1에서와 같이 (-)0.527로 산정되었다. Fig. 5

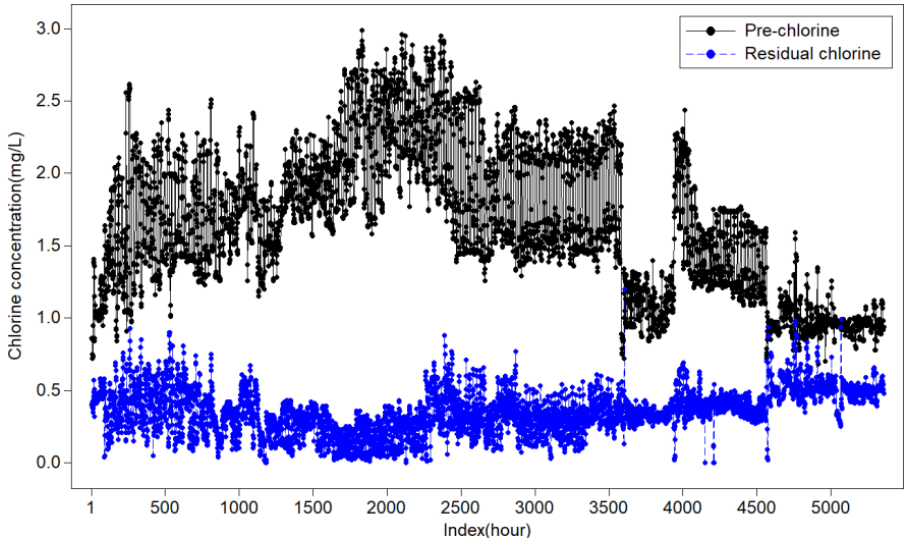


Fig. 4. Time series plot of pre-chlorine and residual chlorine concentration

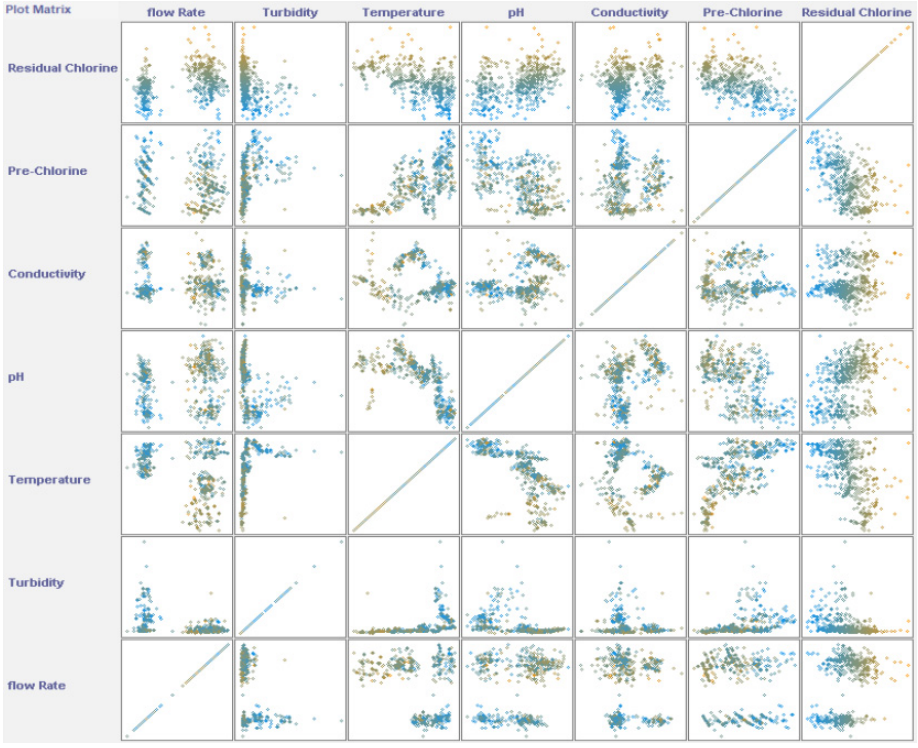


Fig. 5. Plot matrix of correlation between residual chlorine and water quality data

는 침전지 잔류염소농도가 유량 및 수질항목들과 어떤 상관성을 갖는지를 분석한 Plot Matrix로서 유량, 수온, pH, 탁도 등의 순으로 상관성을 보이는 것으로 나타났으며 알칼리도, 유량, pH, 전기전도도는 비례적 상관성을 가지고 있고, 탁도, 수온, 전염소 주입량은 반비례 상관성을 보이고 있다는 것을 알 수 있다.

4. 적용 및 결과

정수장 침전지 유출수의 잔류염소농도 예측을 위하여 사용된 모형은 중회귀모형과 다층퍼셉트론 신경망, 랜덤포레스트 및 LSTM 모형이며, 입력변수는 7개 수질항목 데이터로 구성하였고, 출력변수로는 침전지 유출수의 잔류염소농도로 구성하였다. 또한, 전체 수질 데이터 중 80%에 해당하는 4,238시간의 자료가 학습용 데이터로 사용되었고 검증을 위해서 20%인 1,060시간 데이터가 사용되었다.

모형의 적합도는 검증자료를 대상으로 추정치와의 상관계수 및 RMSE (Root Mean Square Error)를 도출하여 평가하는 것으로 하였으며 각각 중회귀모형, 다층퍼셉트론 신경망, 랜덤포레스트 및 LSTM모형을 상대적으로 비교하는 것으로 하였다. 또한 비학습자료에 의한 검증은 각 모형별로 결정계수(Determination Coefficient)값과 수정 결정계수를 산정하여 모형의 적합도를 평가하였다(Table 2).

Table 2. Results of validation for each model

Model	Validation		R-square
	Correlation	RMSE	
Multiple Regression	0.6261	0.1178	0.392
MLP	0.6247	0.1338	0.442
Random Forest	0.8669	0.0760	0.752
LSTM	0.7620	0.0960	0.580

중회귀모형에서 전체 데이터의 80%인 4,328개의 자료를 사용하여 추정식이 도출되었고 이를 활용하여 나머지 20%에 해당하는 자료에 대한 침전지 잔류염소를 추정하였으며 그 결과 다음과 같이 중회귀식이 도출되었다. 여기서 유량과 pH 항목은 잔류염소농도에 영향이 미비하고 통계적으로 유의하지 않아 식에서 제외된 형태로 나타났으며 상관계수는 0.6261, RMSE값은 0.1178로 산정되었다.

$$\text{Residual Chlorine} = -0.0014*\text{Alkalinity}-0.0077*\text{Temperature}+0.0007*\text{Conductivity}-0.0953*\text{Pre-Chlorine}+0.5136 \quad (6)$$

중회귀분석 모형에 의해 예측된 침전지 잔류염소 농도와 실측값과 비교해 볼 때 가장 낮은 상관계수를 보여 주었는데 이는 침전지내에서 염소의 소비특성이 여러 입력변수들에 대해 상당히 복잡한 관계가 있다는 것을 의미하며 계절별 수질 특성에 따라 염소 소비특성이 다양하게 영향을 받기 때문이다.

다층퍼셉트론 신경망의 경우 중간층과 중간층의 뉴런 수를 변화시켜 예측성능을 평가하였는데 중간층이 1개의 경우 상관계수는 0.6247, RMSE값은 0.1338로 추정되었고 중간층을 2개로 구성한 경우 상관계수 및 RMSE는 0.6732, 0.1394 또한 3개 층으로 증가시킨 경우 상관계수 및 RMSE는 각각

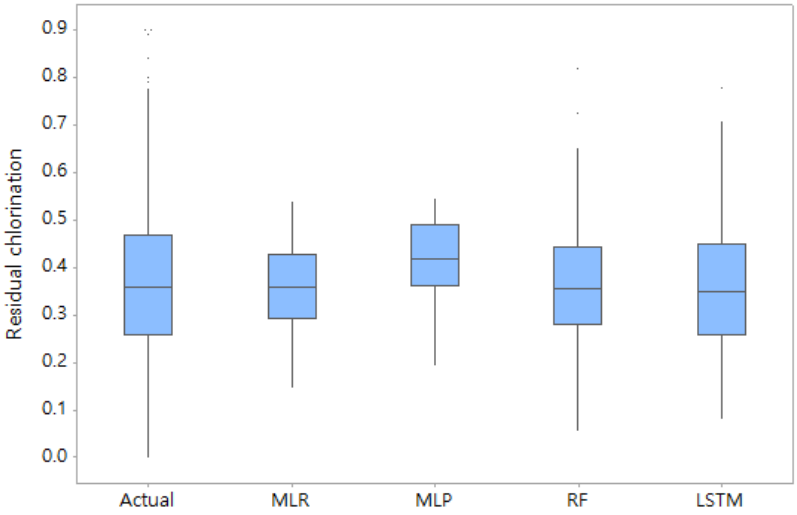


Fig. 6. Box plot of actual data and the results of application models

0.6942, 0.1362로 추정되어 상관계수가 약간 증가하는 것으로 나타나나 유의한 결과를 보이지는 않는 것으로 나타났다.

의사결정나무 구조의 랜덤포레스트의 경우에는 상관계수와 RMSE값은 각각 0.8669과 0.076로 산정되었으며 학습단계에서는 0.98이상으로 계산되어 높은 모형 적합도를 보여주는 것으로 나타났다.

LSTM의 경우에는 모형의 중간층은 4개로 구성하였으며, 각 레이어의 유닛은 256개, 128개, 64개 및 1개로 설정하여 수행하였다. 특히 활성화함수로는 시그모이드(Sigmoid)함수나 Tanh 보다 학습속도가 빠른 것으로 알려져 있는 Relu를 적용하였으며 그 결과 상관계수값은 0.762로, RMSE값은 0.096로서 상관계수는 랜덤포레스트 모형 다음으로 높은 적합도를 보여 주었다.

또한 상자그림(Box plot)을 나타낸 Fig. 6에서 볼 수 있듯이 각 모형에 의한 예측 평균값은 큰 차이를 보이지 않고 있으나 Fig. 7에서는 각 모델 수행 결과에 대한 빈도분석결과 다층퍼셉트론 신경망에 의한 결과가 빈도분포가 벗어나는 경향을 보이고 있어 평균이하의 작은 값을 예측하는데 한계성을 보이고 있음을 알 수 있었다.

Figs. 8~11에서는 각각 중회귀모형, 다층퍼셉트론 신경망, 랜덤포레스트 및 LSTM 모형에 대하여 실제값과 예측값을 이용한 검증결과를 95% 신뢰구간과 예측구간과 함께 도시한 것으로 각각의 그림에서 s값은 표준오차(Standard Error)와 결정계수(Determination Coefficient) R-sq.값과 수정(Adjusted) 결정계수 adj. R-sq.값 을 볼 수 있다.

검증단계에서 중회귀모형, 다층퍼셉트론 신경망, 랜덤포레스트 및 LSTM 모형 각각의 결정계수 값은 0.392, 0.442, 0.752 그리고 0.580로 나타나 랜덤포레스트 모형에 의한 결과

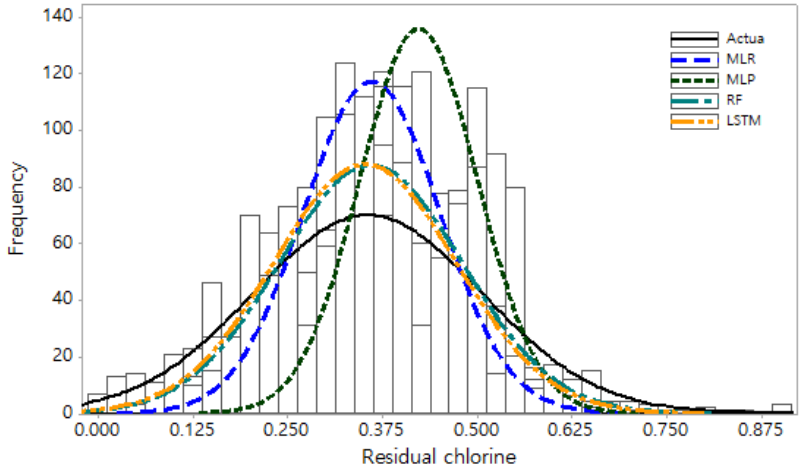


Fig. 7. Histogram comparison of actual data and predicted data from application models

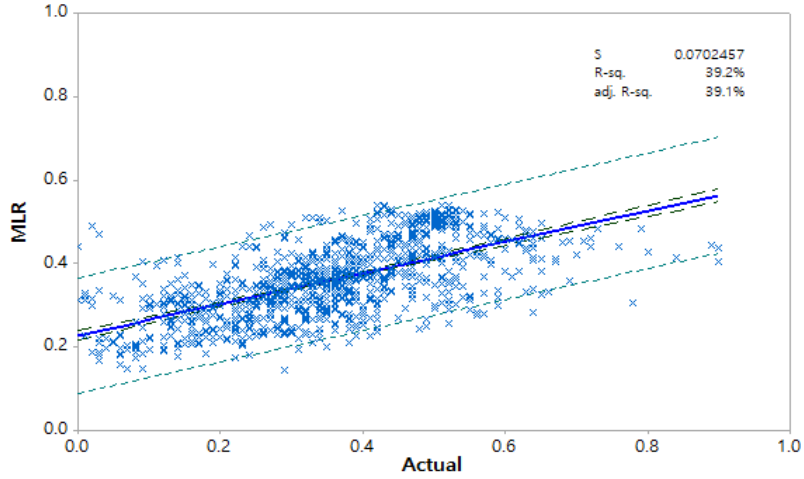


Fig. 8. Comparison between predicted residual chlorine by multiple regression and actual data with confidence and prediction limit

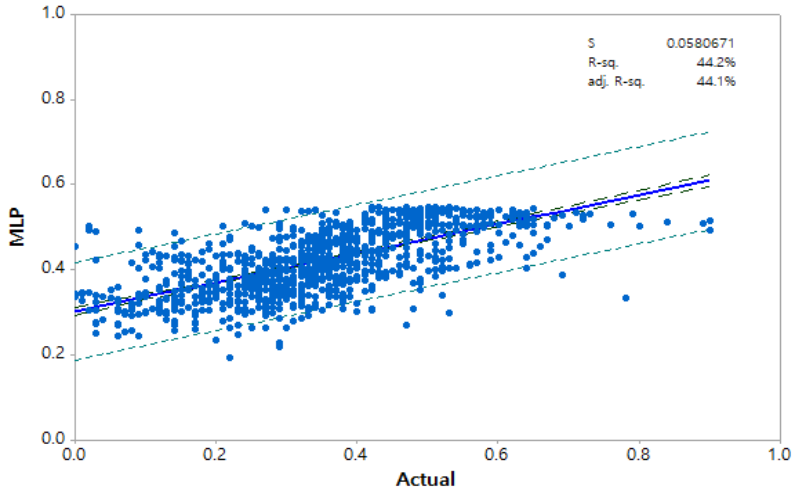


Fig. 9. Comparison between predicted residual chlorine by multi-layer perceptron and actual data with confidence and prediction limit

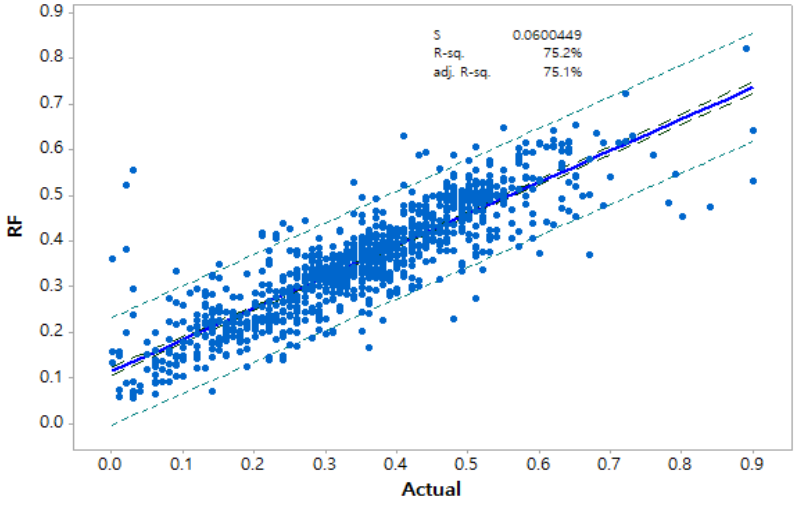


Fig. 10. Comparison between predicted residual chlorine by random forest and actual data with confidence and prediction limit

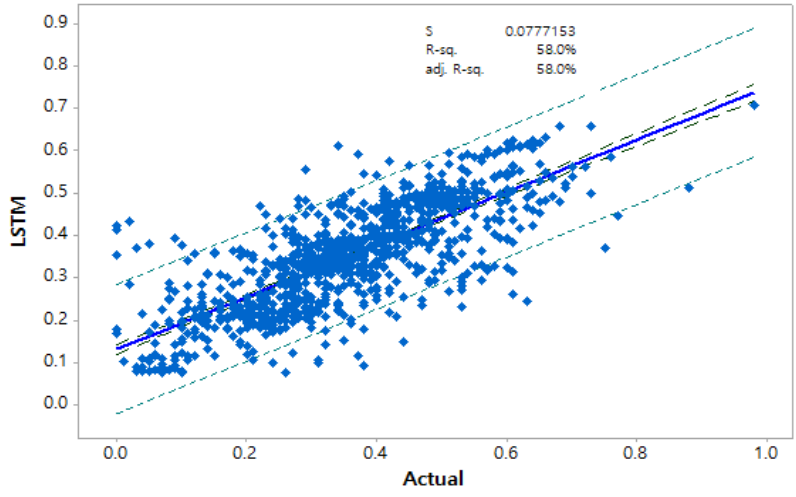


Fig. 11. Comparison between predicted residual chlorine by LSTM and actual data with confidence and prediction limit

가 가장 나은 결과를 보여 수질데이터를 대상으로 하는 지능화 모형은 의사결정 트리 구조가 가장 적합한 것으로 판단할 수 있다. 그러나 각 모형 수행에 필요한 조건이 다양하고 모델 매개변수 등 데이터 처리에 대한 경험과 시행착오 등에 따라 결과 또한 다양하게 도출될 수 있어 이에 대한 심사숙고가 필요하다. 다만 본 연구에서 선정한 모델의 비교 결과에서는 다양한 인공지능 알고리즘 중에서 수질데이터와 같이 분산이 크고 데이터의 차원이나 규모가 큰 차이가 있는 경우 랜덤포레스트와 같이 의사결정 트리구조의 도입과 적용이 타당한 것으로 사료된다.

5. 결 론

발암물질 생성 억제, 생물 멸균 등을 위해서 침전지까지 잔류염소농도를 일정하게 유지하는 것은 중요하다. 그러나 잔류염소농도는 원수의 수질 및 수온 등의 환경변화와 유량의 변화에 따라서 변화하기 때문에 이를 정확하게 예측하는 것은 어렵다. 이에 본 연구에서는 여러 분야에서 높은 정확도를 보이고 있는 인공지능 모형을 활용하여 잔류염소농도 산정 모형을 개발하고자 하였다. 정수처리장에서 수집 가능한 자료를 입력자료로 활용하였으며, 일반적으로 사용되는 중회귀모형과 인공지능 알고리즘 중 다층퍼셉트론 신경망, 랜덤포레스트 및 LSTM 모형을 적용하여 침전지 유출수의 잔류염소농도를 예측하였고 그 결과를 비교, 평가하였다. 검증결과에서는 랜덤포레스트 모형이 가장 양호한 결과를 보여 주었으며 다음으로 LSTM, 다층퍼셉트론 신경망으로 나타났으며 수학적 모형인 중회귀모형이 가장 낮은 결과를 보여 적용상의 한계성을 알 수 있었는데 이는 수량과 수질자료의 수치적인 규모나 차원의 차이뿐만 아니라 계절별 수질특성에 따라 염소소비특성이 매우 다양하게 반응하기 때문으로 판단된다. 인공지능기반의 모형은 블랙박스 모형으로 알려져 있어 수질자료와 같이 물리적, 화학적으로 설명될 수 없는 경우에 적용이 가능하고 내용의 설명보다는 입력 및 출력값이 주어졌을 때 잘 맞추는 장점을 가지고 있는 것은 분명하나, 모형수립 과정에서 학습에 필요한 적정 자료기간이나 계측시간의 적정 간격, 그리고 염소 소비특성을 대변할 수 있는 계측자료 항목의 추가적인 발굴 등이 요구되며 모형 수행에 필요한 조건이 다양하고 모형 매개변수에 따라 달라질 수 있어 정수장 수량 및 수질자료 처리에 대한 경험과 통찰력을 필요로 한다. 다만, 본 연구에서 선정한 모델의 비교 결과에서는 다양한 인공지능 알고리즘 중에서 수질자료와 같이 분산이 크고 자료의 차원이나 규모가 큰 차이가 있는 경우, 랜덤포레스트와 같이 의사결정 트리구조의 도입과 적용이 유리한 것으로 나타났다는데 의의가 있으며 이를 보다 구체적으로 절차화하기 위한 추가적인 연구가 요구된다. 본 연구에서 제시된 결과를 토대로 정수장의 환경 변화 및 수질 변화에 대응하기 위한 인공지능 정수장 구축에 필요한 정보로 활용될 수 있기를 기대한다.

감사의 글

본 결과물은 환경부의 재원으로 한국환경산업기술원의 가뭄대응 물관리 혁신 기술개발사업의 지원을 받아 연구되었습니다(2022003610002).

Conflicts of Interest

The authors declare no conflict of interest.

References

Breiman, L. (1996). “Bagging predictors.” *Machine Learning*, Vol. 24, pp. 124-140.

Breiman, L. (2001). “Random forests.” *Machine Learning*, Vol. 45, No. 1, pp. 5-32.

Jeon, H.B., Lee, Y.J., and Lee, J.D. (2001). “Effects of Prechlorination 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 water level prediction of river water level using deep-lenaring open library.” *Journal of Korean Society of Hazard Mitigation*, Vol. 18, No.1, pp.1-11.

Kang, K.W., Park, C.Y., and Kim, J.H. (1992). “Nonlinear prediction of streamflow by applying pattern recognition method.” *Journal of Korean Association of Hydrological Sciences*, Vol. 25. No.3, pp.105-113.

Kim, D., Kim, J., Kwak, J., Necesito, I.V., Kim, J., and Kim, H.S. (2020). “Development of water level prediction models using deep neural network in mountain wetlands.” *Journal of Wetland Research*, Vol. 22, No. 2, pp. 106-112.

Kim, J.H. (1993). *A study on hydrologic forecasting of stream flow by using artificial neural network*. Ph.D. Dissertation, Inha University.

Kumar, A.P.S., Sudheer, K.P., Jain, S.K., and Agarwal, P.K. (2005). “Rainfall runoff modeling using artificial neural networks: Comparison of network types.” *Hydrological Process* Vol. 19,

pp. 1277-1291.

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 the Korean Society of Water and Wastewater*, Vol. 21, No. 5, pp. 601-607.

Lisboa, P.G.J. (1992). *Neural networks: Current application*. Chapman & Hall, London, pp. 5-6.

Maneual, J.R., and Jean, B.S. (1999). "Assessing empirical linear and non-linear modeling of residual chlorine in urban drinking water systems." *Environmental Modeling & Software*, Vol. 14, No. 1, pp. 93-102.

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

Tiwari, M.K., and Chatterjee, C. (2010). "Development of an accurate and reliable hourly flood forecasting model using wavelet-bootstrap-ANN (WBANN) hybrid approach." *Journal of Hydrology*, Vol. 394, No. 3, pp. 458-470.

Uber, J.G. (2003). *Maintaining distribution system residuals through booster chlorination*, IWA Publishing, London, UK, pp. 42-47.

Yoon, J.Y., Byoun, S.J., and Choi, Y.S. (2001). "Importance of prechlorination 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.