Korean Chem. Eng. Res., 59(4), 532-541 (2021)
https://doi.org/10.9713/kcer.2021.59.4.532
PISSN 0304-128X, EISSN 2233-9558

시계열 교차검증을 적용한 2,3-BDO 분리공정 온도예측 모델의 초매개변수 최적화

안나현*,** · 최영렬*,** · 조형태* · 김정환*,†

*한국생산기술연구원 친환경재료공정연구그룹
44413 울산광역시 중구 종가로 55
**연세대학교 화공생명공학과
03722 서울특별시 서대문구 연세로 50
(2021년 5월 31일 접수, 2021년 7월 20일 수정본 접수, 2021년 8월 5일 채택)

Application of Time-series Cross Validation in Hyperparameter Tuning of a Predictive Model for 2,3-BDO Distillation Process

Nahyeon An*,, Yeongryeol Choi*,**, Hyungtae Cho* and Junghwan Kim*,†**

**Green Materials and Processes R&D Group, Korea Institute of Industrial Technology, 55, Jongga-ro, Ulsan, 44413, Korea*
***Department of Chemical and Biomolecular Engineering, Yonsei University, 50, Yonsei-ro, Seoul, 03722, Korea*
(Received 31 May 2021; Received in revised from 20 July 2021; Accepted 5 August 2021)

요 약

최근 인공지능에 대한 관심이 높아짐에 따라 화학공정분야에서도 인공지능을 활용한 연구가 많아지고 있다. 그러나 인공지능 기반 모델이 충분히 일반화되지 않아 학습에 이용되지 않은 새로운 데이터에 대한 예측률이 떨어지는 과적합 현상이 빈번하게 일어나고 있으며, 교차검증은 과적합을 해결하는 방법 중 하나이다. 본 연구에서는 2,3-BDO 분리공정 온도 예측 모델의 초매개변수 중에서 배치 개수와 반복횟수를 조정하기 위해 시계열 교차검증을 적용하고 일반적으로 사용되는 K 겹 교차검증과 비교하였다. 결과적으로 K 겹 교차검증을 사용했을 때 보다 시계열 교차검증 방식을 사용했을 때 MAPE는 0.61% 증가한 반면 RMSE는 9.06% 감소하였고 학습 시간은 198.29초 적게 소요되었다.

Abstract – Recently, research on the application of artificial intelligence in the chemical process has been increasing rapidly. However, overfitting is a significant problem that prevents the model from being generalized well to predict unseen data on test data, as well as observed training data. Cross validation is one of the ways to solve the overfitting problem. In this study, the time-series cross validation method was applied to optimize the number of batch and epoch in the hyperparameters of the prediction model for the 2,3-BDO distillation process, and it compared with K-fold cross validation generally used. As a result, the RMSE of the model with time-series cross validation was lower by 9.06%, and the MAPE was higher by 0.61% than the model with K-fold cross validation. Also, the calculation time was 198.29 sec less than the K-fold cross validation method.

Key words: Cross validation, Distillation process, Predictive model, Hyperparameter tuning, Time-series cross validation

1. 서 론

인공지능에 대한 관심이 높아짐에 따라 많은 분야에서 인공지능을 접목한 연구가 수행되고 있다. 화학공정분야에서도 인공지능을 활용한 공정 예측[1-4], 이상 감지[5-9], 공정 제어[11,12]에 대한 연구가 활발히 진행되고 있다. 인공지능 개발과정에서의 과적합은 개발한 모델이 충분히 일반화되지 않아 학습된 데이터는 잘 예측하지만 학습에 이용되지 않은 새로운 데이터에 대한 예측률이 떨어지는 현상으로 많은 연구자들이 이 문제를 극복하기 위해 노력해왔다[13].

과적합을 해결하기 위한 방법에는 교차검증[14], 드롭아웃[14], 가중치규제[15], 조기종료[15] 등이 있으며 그 중에서 교차검증은 모델 개발에 사용하는 훈련 데이터와 검증 데이터를 한 번 혹은 여러 번 분할하여 학습과 검증을 진행하는 방식이다[16]. K 겹 교차검증은 전체 훈련 데이터를 K개의 폴드(Fold)로 나누어 한 개의 폴드를 검증 데이터로 사용하고 나머지 데이터를 모델 훈련에 사용하여 모든 폴드가 한 번씩 검증에 쓰일 때까지 K번의 훈련과 검증을 반복하는 방식으로 많은 연구에서 사용되어 왔다[16].

예를 들어 이승훈 등[2]은 광업데이터의 시계열 분석을 통해 머

†To whom correspondence should be addressed.
E-mail: kjh31@kitech.re.kr

신러닝 모델을 개발하였으며 이를 K 겹 교차검증을 통해 검증하였다. Naiju Zha 등[3]은 철강산업에서의 용광로 내의 온도를 XGBoost와 GRU를 이용해 예측했으며 K 겹 교차검증을 통해 그 모델을 검증하였다. Xiaoxia Chen 등[17]은 철광석 소결 공정에서 발생하는 포괄적인 탄소 비(Comprehensive carbon ratio)를 예측하는 하이브리드 시계열 예측 모델 개발 과정에 K 겹 교차검증을 적용하여 모델을 검증하였다. Zhi-Jun Lu 등[5]은 K 겹 교차검증을 활용하여 복잡한 산업 공정에서의 예측률을 높이기 위해 서포트 벡터 회귀를 기반으로 한 예측모델을 제안하였다.

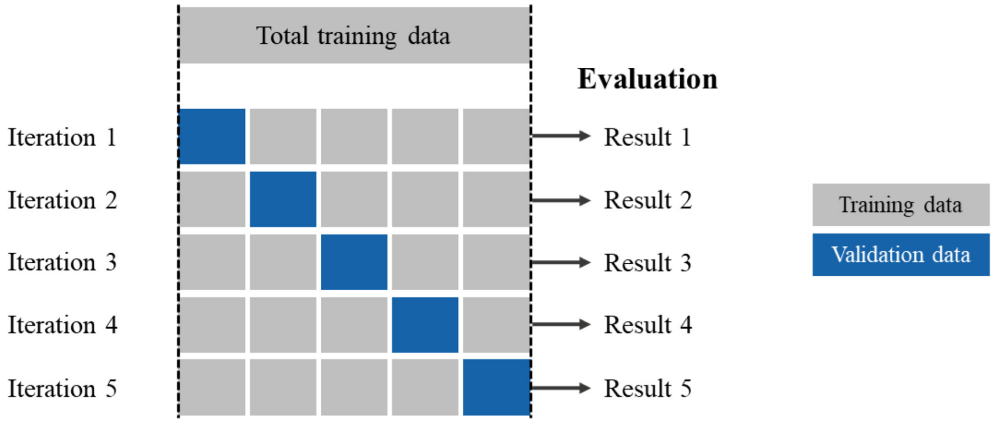
K 겹 교차검증은 방법이 간단하여 많은 연구에 활용되었지만 모델 학습과 검증을 K번 반복하는 점에서 학습과 검증에 많은 시간이 소요되는 단점이 있다[13]. 또한 일반적으로 K 겹 교차검증은 데이터 순서를 무작위로 섞은 후 교차 검증을 수행하기 때문에 시간에 따라 특성이 달라지는 시계열 데이터에 적용할 수 없다. 또한 전체 데이터를 K 개의 폴드로 나누고 학습하는 과정에서 다른 특성을 지니는 일부 폴드를 검증에 사용하면 학습 신뢰도가 감소할 수 있다. 유지 및 보수를 위해 공정을 중단하고 재시작 하는 것이 빈번하게 발생하는 화학공정에서 공정의 시작 - 안정된 운전 상태 - 공정 중단에 이르는 폭 넓은 범위를 반영하는 모델을 만들기 위해서는 일반적인 K 겹 교차검증을 적용할 수 없다. 따라서 화학공정의 시계열 데이터를 이용한 예측모델을 검증하기 위해서는 시계열 교차검증을 적용하여 모델을 개발하여야 한다.

본 연구에서는 증류공정에서 시계열 교차검증 적용의 타당성을 입증하기 위해 K 겹 교차검증과 비교하였다. 교차검증을 적용하여 2,3-부탄디올(2,3-BDO) 증류탑의 온도 예측 모델의 초매개변수인 배치 개수와 반복횟수를 최적화하여 두 가지 교차검증 방식을 비교하고자 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구에서 사용되는 교차검증 방법 두 가지에 대해 서술하였다. 3장에서는 연구에서 사용한 2,3-BDO 증류탑에 대한 설명과 예측 모델 설정에 대해 서술하였다. 4장에서는 두 가지 교차검증의 결과와 모델 예측 성능을 비교하였다. 마지막으로 5장에서는 본 논문의 결론을 요약하고 향후 연구에 대해서 논의하였다.

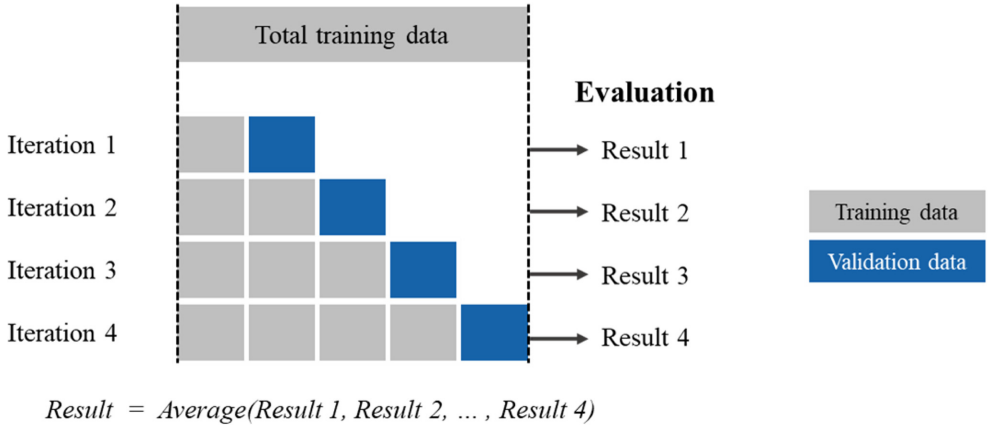
2. 교차검증

K 겹 교차검증은 Fig. 1과 같이 훈련 데이터를 총 K 개의 중첩되지 않는 폴드(Fold)로 나눈 후 K 개의 폴드 중 K-1 개의 폴드로 학습을 하고 나머지 한 개의 폴드로 검증을 진행하며, 모든 폴드가 검증에 사용될 때까지 K 번의 학습과 검증을 반복(Iteration)하는 방식이다 [13]. 각 반복에서의 평가 결과를 평균 내어 모델의 검증결과를 도출한다.



$$Result = Average(Result\ 1, Result\ 2, \ldots, Result\ 5)$$

Fig. 1. K-fold cross validation schematic diagram.



$$Result = Average(Result\ 1, Result\ 2, \ldots, Result\ 4)$$

Fig. 2. Time-series cross validation schematic diagram.

시계열 교차검증(Time series cross validation or rolling origin recalibration)은 시계열 데이터를 교차검증 하기 위해 고안된 방법으로 훈련 데이터를 N 개의 그룹으로 나눈 후에 마지막 그룹을 검증 데이터로 두고, 나머지 데이터를 훈련 데이터로 사용하는 방식이다[13]. 예를 들어 Fig. 2와 같이 5개의 그룹으로 훈련 데이터를 나누었다면 첫 번째 반복(Iteration 1)에서 1번 그룹으로 학습하고 2번 그룹으로 검증한다. 다음 반복(Iteration 2)에서는 1, 2번 그룹으로 학습하고 3번 그룹으로 검증하며 앞으로 나아가 N번째 그룹이 검증에 사용될 때까지 학습과 검증을 N-1 번 반복한다.

3. 연구방법

Fig. 3은 본 논문의 연구 과정의 개략도이며 순서는 다음과 같다.

(1) 전처리 한 데이터를 전체 학습 데이터(Training data)와 테스트 데이터(Test data)로 나눈다.

(2) 전체 학습 데이터를 교차검증 방식에 따라 학습 데이터(Training data)와 검증 데이터(Validation data)로 나눈다.

(3) 배치 개수와 반복횟수 사례에 따라 학습 데이터로 모델을 학습하고 검증 데이터로 모델을 검증한다.

(4) 사례연구 결과 검증 성능이 가장 좋은 배치 개수와 반복횟수 사례를 선택하여 전체 학습 데이터로 모델을 개발하고 테스트 데이터로 모델을 평가한다.

연구 순서에 따라 각 부분에 대한 설명을 다음 절부터 자세히 설명하였다. 본 연구에서 개발된 모델은 Intel(R) Core(TM) i7-9700 CPU@3.00GHz, 16.0 GB RAM 환경에서 개발되었다.

3-1. 공정 개요 및 대상 데이터

본 연구의 대상 공정은 군산에 위치하며 미생물의 발효를 통해

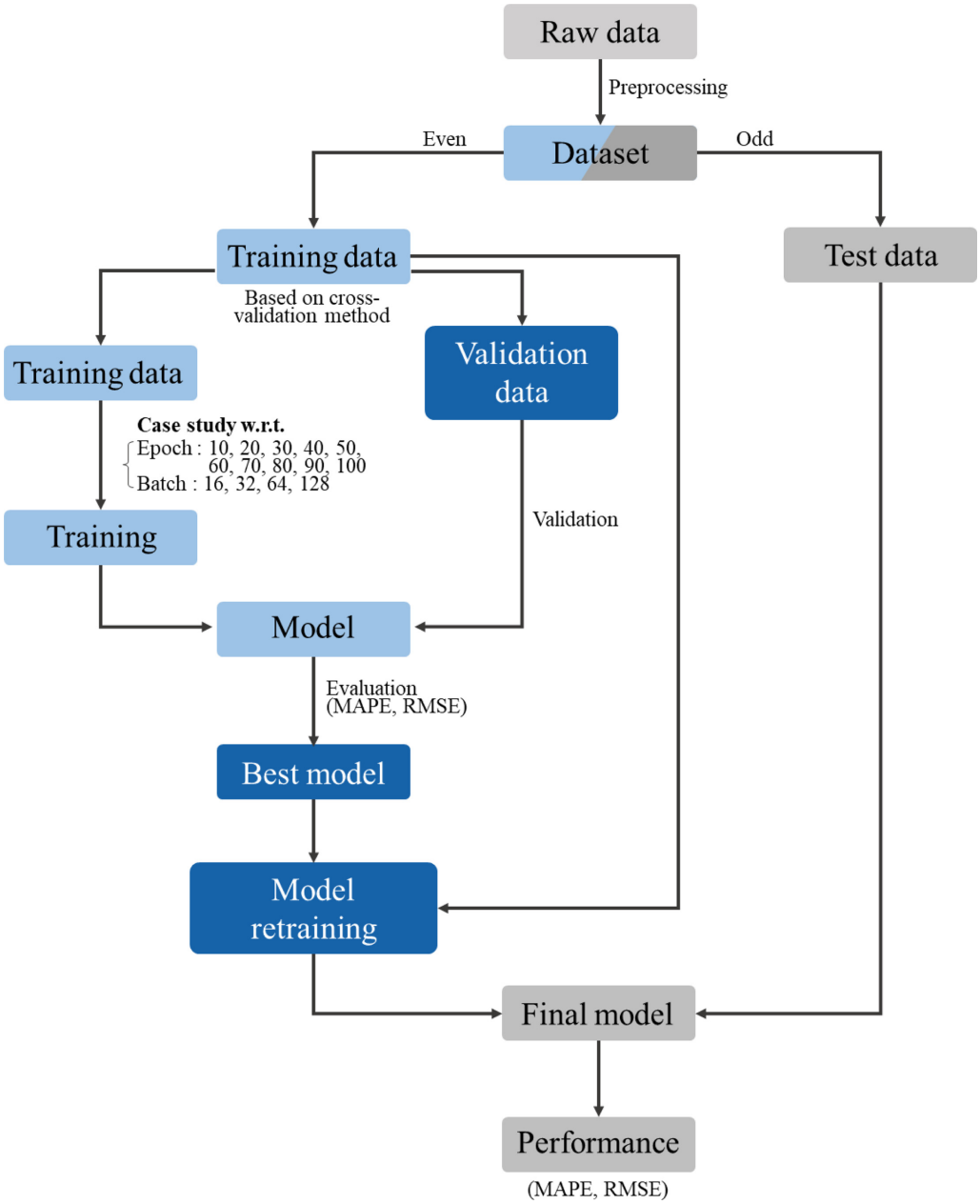


Fig. 3. Research procedure.

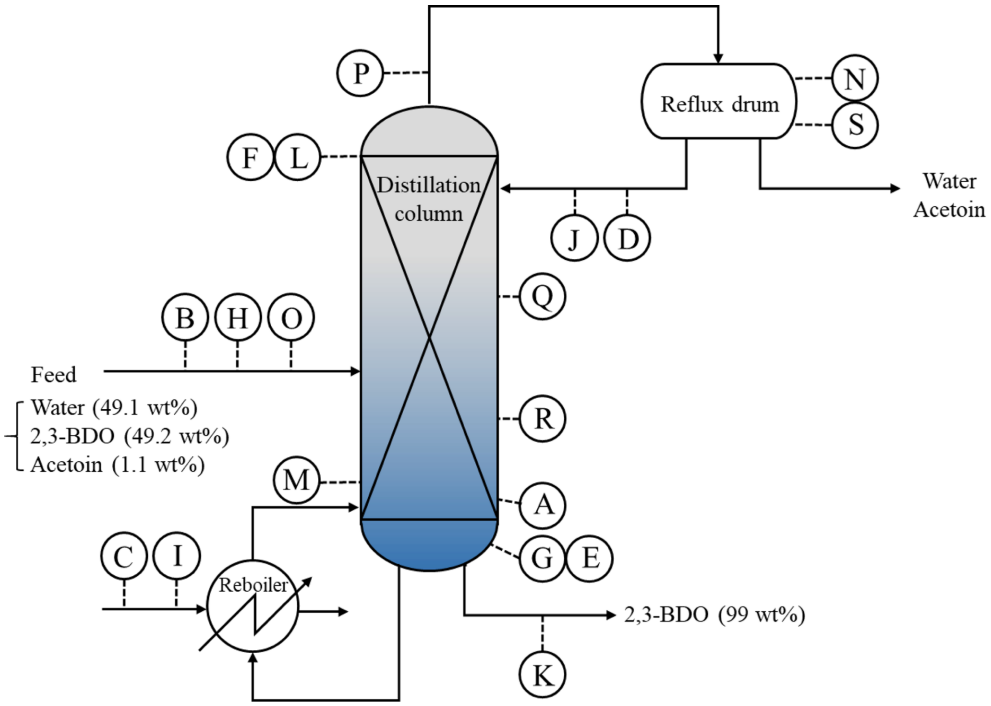


Fig. 4. Schematic diagram of 2,3-BDO distillation column.

Table 1. Process variables description

Notation	Variable description	Unit
A	Bottom product temperature controller	°C
B	Feed flow rate indicator	L/hr
C	Reboiler steam inlet flow rate indicator	kg/hr
D	Reflux flow rate controller	L/hr
E	Column bottom level	%
F	Top pressure controller	kg/cm^{2}g
G	Bottom liquid level controller	%
H	Feed flow rate controller	L/hr
I	Reboiler steam inlet flow rate controller	kg/hr
J	Reflux flow rate indicator	L/hr
K	Bottom product flow rate indicator	L/hr
L	Top pressure indicator	kg/cm^{2}g
M	Middle 1 temperature controller	°C
N	Reflux drum pressure indicator	kg/cm^{2}g
O	Feed temperature	°C
P	Distillate temperature indicator	°C
Q	Middle 2 temperature controller	°C
R	Middle 3 temperature controller	°C
S	Reflux drum temperature	°C

생산된 약 50 wt%의 2,3-부탄디올(2,3-Butanediol, 2,3-BDO) 원료를 증류시켜 순도 99 wt%의 2,3-BDO를 생산하는 파일럿 플랜트이고 Fig. 4는 대상 공정의 개략도이다. 본 연구에서는 2,3-BDO 제품이 생산되는 생산단 온도 예측 모델을 개발하기 위해 실제 공정 데이터 중 증류탑의 생산단 온도와 관련이 있는 19개의 변수를 수집하였으며 각 변수가 수집되는 계기를 Table 1에 나타내었다. 실제 공정에 저장되는 데이터 중에서 2020년 7월 10일부터 13일까지 1분 단위로 기록된 5,760개의 데이터를 사용하였다.

3-2. 데이터 전처리 및 특성선택

공정 데이터는 다양한 외란으로 인해 오차를 포함하고 있어 공정 데이터를 기반으로 예측모델을 개발할 때 데이터 전처리, 특성 선택 과정이 필요하다. 데이터 전처리 방법 중에서 잡음 제거는 데이터의 중요한 정보를 추출하기 위해 데이터의 잡음을 제거하는 기법이다[18]. 본 연구에서는 잡음 제거 기법의 하나인 저역 통과 필터(Low pass filter, LPF)를 이용하여 공정 데이터의 잡음을 제거하고 모델 개발에 이용하였으며 일부 변수에 대해 저역 통과 필터 적용 전후를 Fig. 5에 나타내었다.

출력변수와 무관한 변수들을 입력변수로 사용하면 모델의 성능을 악화시킬 수 있기 때문에 모델의 성능을 높이기 위해 특성 선택이 필요하다[18]. 피어슨 상관관계 계수(Pearson correlation coefficient)는 두 변수 사이의 선형적인 관계를 측정하는 지표이다[19]. 상관관계 계수(r)가 0에 가까우면 상관관계가 없고, -1에 가까우면 높은 음의 상관관계, 1에 가까우면 높은 양의 상관관계를 나타낸다. 일반적으로 상관계수의 절댓값이 0.3보다 크면 두 특성 간의 상관관계가 있다고 판단한다. 따라서 본 연구에서는 식 (1)과 같이 특성 선택을 위해 입력변수와 출력변수 간의 상관관계 계수 절댓값이 0.3을 넘는 변수들을 선택하였다. 하지만 선택된 입력 변수 간의 상관관계가 높으면 입력 변수의 독립성이 사라지는 다중 공선성을 갖게 되므로 이를 고려하여 입력변수 간의 상관관계가 높은 경우에는 높은 상관관계를 가지는 두 입력 변수 중에서 한 가지를 선택하였다.

$$\begin{cases} |r| < 0.3 & : \text{No correlation} \\ 0.3 \le |r| < 0.7 & : \text{Weak correlation} \\ |r| \ge 0.7 & : \text{Strong correlation} \end{cases} \quad (1)$$

3-3. 예측 모델 설정

3-3-1. 데이터 분할

예측 모델을 개발하기 위해서는 독립적으로 존재하는 훈련 데이

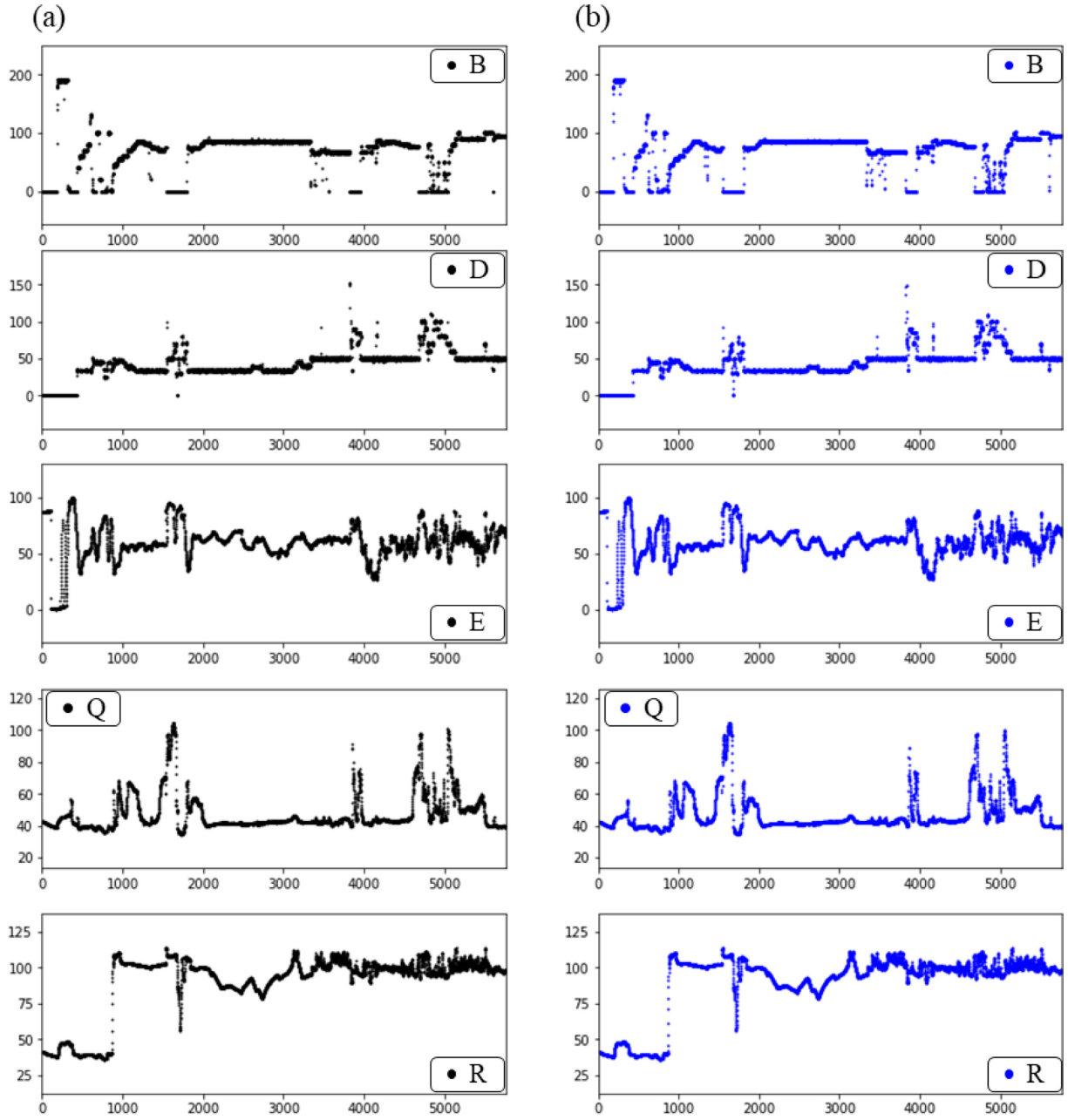


Fig. 5. Operation variables of (a) before LPF noise filtering, and (b) after.

터와 테스트 데이터가 필요하다. 일반적으로 전체 데이터를 무작위로 섞은 후 훈련 데이터와 테스트 데이터를 7:3 혹은 8:2의 비율로 나눈다. 그러나 연속 분리공정의 데이터는 시계열의 특성을 나타내므로 본 연구에서는 훈련 데이터와 테스트 데이터를 독립적으로 분할하면서 시계열 특성을 유지하기 위해 시간 순서로 배열된 데이터를 사용하였다. 또한 본 대상 공정의 데이터는 공정의 시작점(Strat-up)부터 안정화되는 시점을 포함하는 다이나믹한 특성을 가지는 데이터이므로 Fig. 6과 같이 짝수 번은 훈련 데이터로, 홀수 번은 테스트 데이터로 나누었다.

데이터를 전체 훈련 데이터와 테스트 데이터로 분할한 다음 전체 훈련 데이터를 교차검증 방식에 따라 훈련 데이터와 검증 데이터로 나눈다. 일반적으로 K 겹 교차검증에서 폴드 개수 K는 5개 또는 10개로 설정된다[20]. 본 연구에서는 K 겹 교차검증과 시계열 교차검증의 비교를 위해 두 방식에서 전체 훈련 데이터를 동일하게 5개로 나누었고 그 결과 개별 폴드에서의 데이터 개수는 576개로 설정되었으며 출력 변수인 2,3-BDO 생산단의 온도를 5개의 폴드로 나눈 결과는 Fig. 7에 나타내었다.

3-3-2. 초매개변수 선택

본 연구에서는 시계열 데이터를 예측하기 위해 개발된 순환 신경망(Recurrent neural network, RNN)의 종류 중 하나인 장단기 메모리(Long short-term memory, LSTM) 알고리즘을 사용하여 모델을 개발하였다. LSTM 알고리즘은 RNN 알고리즘에서 발생하는 기울기 소실 문제를 극복하기 위해 제안되었으며 다양한 분야에서 시계열 데이터 예측 및 분석을 위해 사용하는 모델이다[21]. LSTM 기반 예측 모델에서 사용자가 직접 설정해주어야 하는 초매개변수들은

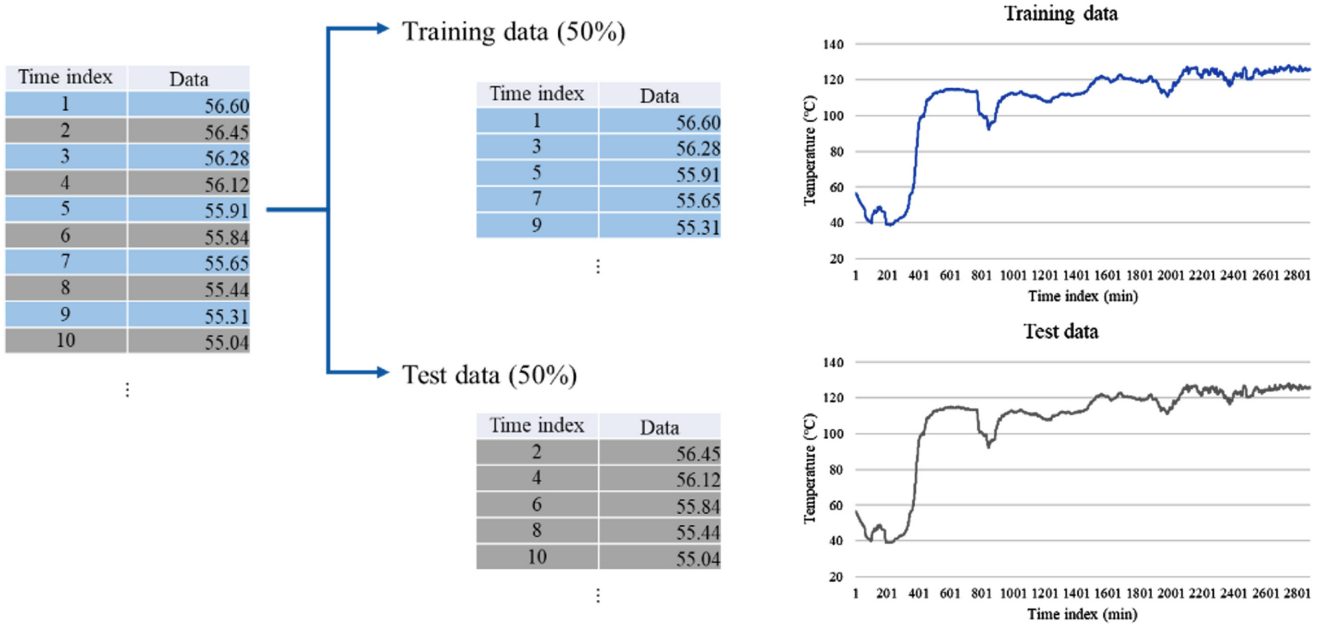


Fig. 6. Training and test data splitting method and results.

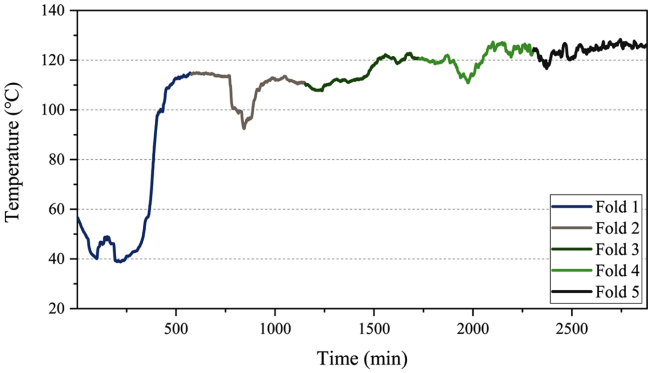


Fig. 7. Output data separation with 5-folds.

반복횟수(Epoch), 배치 개수(Batch number), 최적화 방법(Optimizer), 손실함수(Loss function), 학습률(Learning rate), 드롭아웃 비율(Dropout ratio), 활성화 함수(Activation function) 등이 있다.

배치는 모델의 가중치를 한 번 업데이트 하는데 사용되는 데이터의 묶음을 의미하고, 반복횟수는 모델이 학습 데이터를 학습하는 횟수를 의미한다[22]. 배치 개수는 안정적인 학습에 큰 영향을 미치며, 배치 개수가 적어진다면 모델의 최적화와 일반화가 어려워지고 배치 개수가 많아지면 모델의 가중치를 반복해서 업데이트하여 모델의 성능이 불안정해진다. 반복횟수는 모델의 정확도와 학습시간에 영향을 미치며 값이 커지면 모델의 정확도가 증가하지만 학습시간이 오래 소요되고 값이 작아지면 학습시간은 짧게 소요되지만 모델의 정확도가 떨어질 수 있다. 배치 개수와 반복횟수는 머신러닝 모델을 개발하는데 있어 가장 중요한 초매개변수이지만 두 값은 사전에 정해진 것이 없으므로 사례연구를 통해 최적의 값들을 찾아야 한다[22]. 따라서 본 연구에서는 최적의 배치 개수와 반복횟수를 조정하기 위해 배치 개수는 16, 32, 64, 128개, 반복횟수는 10에서 100까지 10의 간격으로 10개의 사례로 나누어 총 40개의 사례에 대해 교차검증을 수행하고 가장 좋은 검증 성능을 가지는 배치 개수와 반복횟수를 선택하였다.

Table 2. Hyperparameter setting

Item	Value
Loss function	MSE
Optimizer	Adam
Learning rate	0.01
Dropout	0.3
Activation function	Elu

본 연구에서 배치 개수와 반복횟수를 제외한 다른 초매개변수들은 Table 2와 같이 동일하게 사용하였다. 실제 결과와 모델이 예측한 결과값 사이의 오차를 손실 함수라고 하며 본 연구에서는 손실함수로 평균 제곱 오차(Mean squared error, MSE)를 사용하였고 최적화 방법은 Adam[23] 방법을 이용하였다. 학습률은 훈련 기간 가중치가 업데이트되는 스텝의 크기이며 보통 0.0001에서 0.1 사이의 값으로 지정한다. 본 연구에서는 학습률을 0.01로 지정하였다. 드롭아웃 또한 초기 가중치에 따라 예측모델의 결과가 달라지기 때문에 동일 한 조건에서의 결과를 비교하기 위해 초기가중치(seed)를 고정하였다.

3-3-3. 모델 평가 방법

모델의 검증 결과와 예측 결과를 평가하기 위해서 평균절대비오차(Mean absolute percentage error, MAPE)와 평균 제곱근 편차(Root mean square error, RMSE)를 비교하였으며 식은 다음과 같다.

$$\mathrm{MAPE} = \frac{100}{N}\sum\frac{|y_i - x_i|}{x_i} \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2} \tag{3}$$

N은 전체 데이터 개수이며, x_i는 실제 값, y_i는 인공신경망의 예

측값을 나타낸다. 두 평가지표 모두 오차를 나타내므로 작을수록 높은 예측 성능을 의미한다.

4. 연구 결과

4-1. 데이터 특성선택

Fig. 8은 공정 데이터에 대한 피어슨 상관관계 계수를 나타낸 그래프로 색이 진할수록 높은 상관관계를 나타낸다. 출력변수(A)와 상관관계가 없는 6개의 변수(B, D, F, G, H, Q)를 제외하였다. 나머지 변수들간의 다중공선성을 고려하여 최종적으로 출력 변수(A)와 6개의 입력 변수(C, E, J, O, R, S) 가 사용되었다.

4-2. 교차검증

Fig. 9는 두 교차검증 방법을 사용하여 배치 개수와 반복횟수의 사례에 따른 예측 모델의 검증 결과를 시각화한 그림이다. 색이 진할수록 낮은 정확도를 나타내며 색이 옅을수록 높은 정확도를 나타낸다. K 겹 교차검증을 사용했을 때는 배치 개수와 반복횟수가 증가함에 따라 정확도가 증가하여 사례연구 범위에서 최적점을 확인할 수 없었으나 시계열 교차검증을 사용했을 때는 배치 개수가 128개, 반복횟수가 30 일때 최적점을 확인할 수 있었다.

두 검증 방식을 비교하기 위해 각 방법에서 가장 좋은 성능을 나타낸 사례를 Table 3에 나타내었다. K 겹 교차검증을 사용했을 때 배치 개수가 128개, 반복횟수가 60 ($KF_{128,60}$) 일 때 평균 MAPE가 18.64, 평균 RMSE는 10.35 였으며 시계열 교차검증을 사용했을 때 배치 개수가 128개, 반복횟수가 30 ($TS_{128,30}$) 일 때 평균 MAPE가 7.41, 평균 RMSE가 8.13으로 도출되었다.

K 겹 교차검증에서 전반적으로 더 낮은 검증 성능을 보인 이유는 공정의 시작 지점(Strat-up)으로 나머지 폴드와 다른 양상을 보이는 Fold 1이 검증에 사용되어 평균 RMSE와 MAPE가 높아지는 현상이 발생했기 때문이다.

또한 시계열 교차검증에서 사용되는 학습 및 검증 시간은 132.41초이지만 K 겹 교차검증에서는 330.70초가 소요되었다. 이 결과는

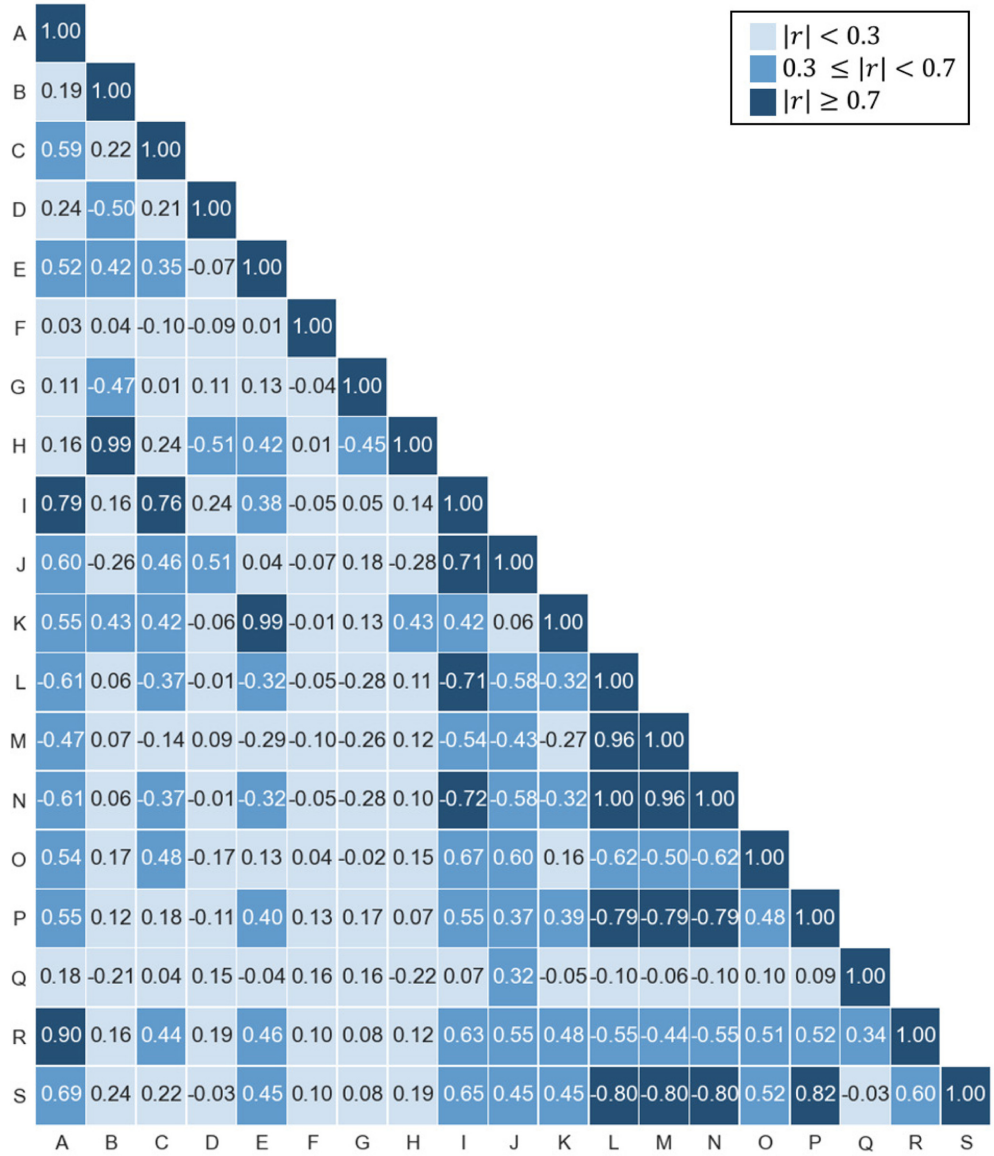


Fig. 8. Pearson correlation coefficient of process variables.

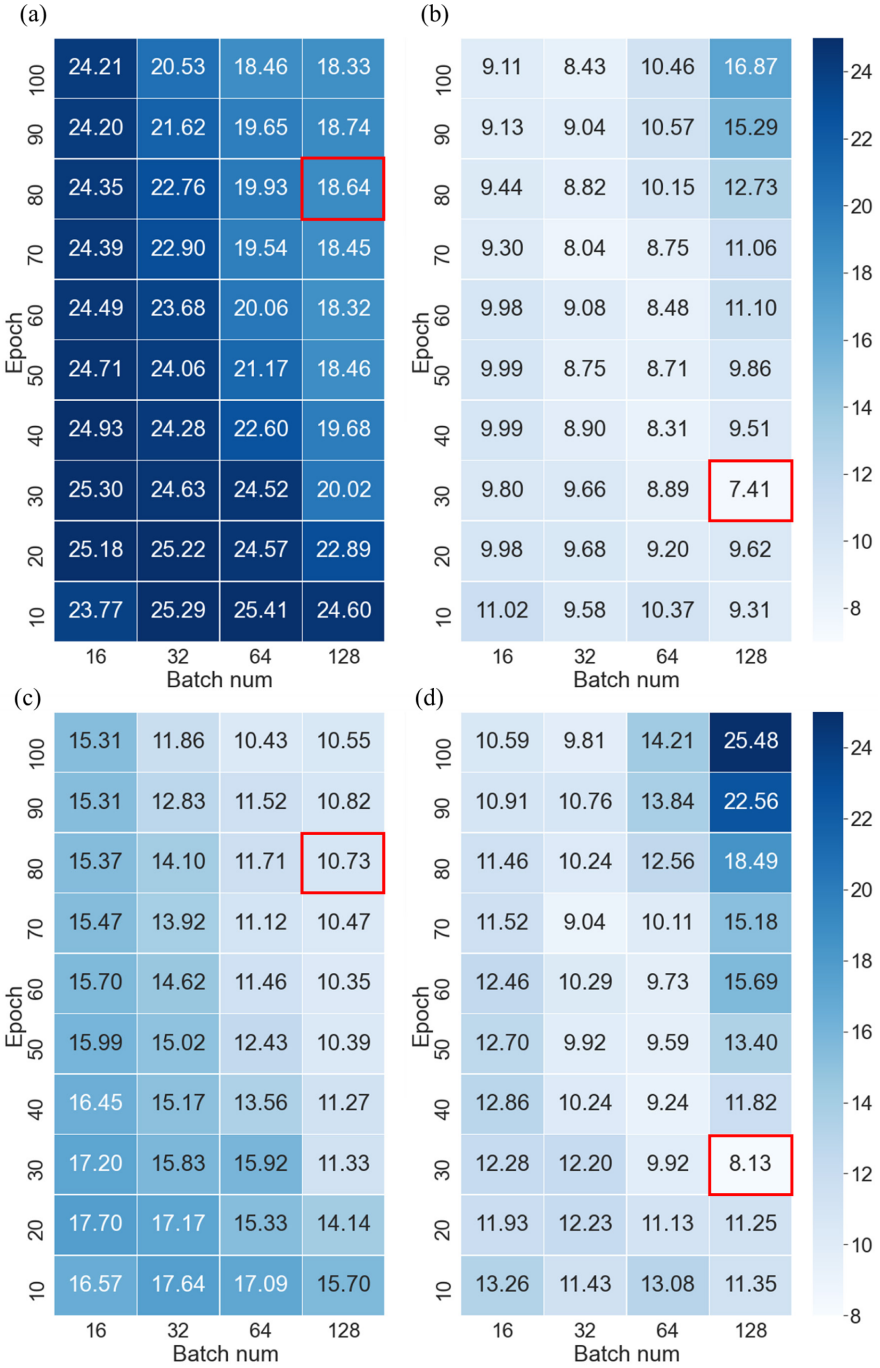


Fig. 9. (a) Average MAPE of K-fold cross validation, and (b) time-series cross validation. (c) Average RMSE of K-fold cross validation, and (d) time-series cross validation.

Table 3. Model validation results of best case

Model	Evaluation	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average	Time
$TS_{128,30}$	MAPE		8.31	6.01	12.01	3.32	7.41	132.41
	RMSE		8.88	7.08	12.69	3.87	8.13	
$KF_{128,60}$	MAPE	71.44	5.63	4.06	4.64	5.82	18.32	330.70
	RMSE	31.53	5.60	3.66	3.41	7.53	10.35	

시계열 교차검증이 K 겹 교차검증보다 훈련에 사용되는 폴드와 반복(iteration)이 적어 더 빠르게 훈련과 검증을 수행할 수 있기 때문이다.

4-3. 모델 성능 평가

두 가지 교차검증방법을 통해 선택된 최적의 초매개변수 ($KF_{128,60}$, $TS_{128,30}$)로 전체 훈련 데이터를 사용해 모델을 개발한 뒤

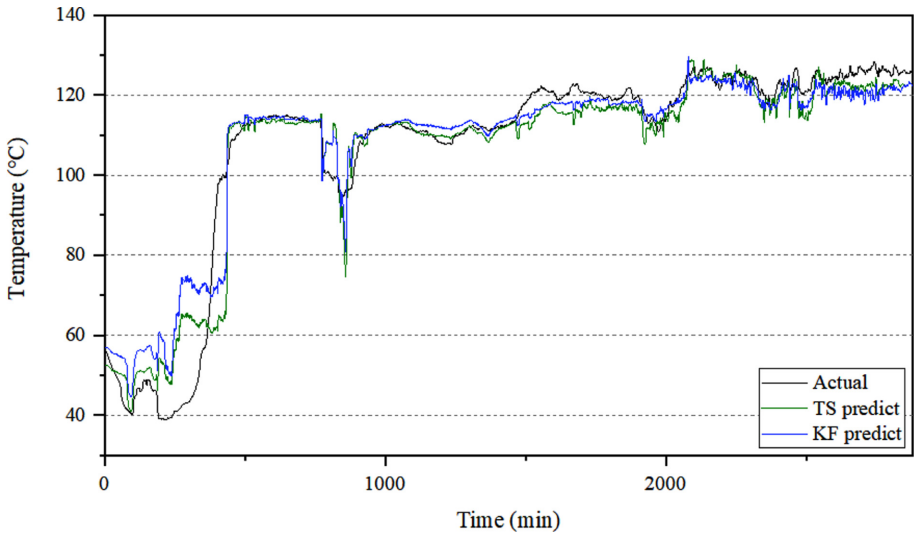


Fig. 10. Model prediction of $KF_{128,60}$, $TS_{128,30}$ and actual data.

Table 4. Model test results of $KF_{128,60}$ and $TS_{64,30}$

	MAPE	RMSE
$TS_{128,30}$ (i)	29.50	6.73
$KF_{128,60}$ (ii)	28.89	7.34
Difference (%) (i)-(ii)	0.61	-9.06

예측성능을 비교하였으며 그 결과는 Table 4와 Fig. 10에 작성하였다. K 겹 교차검증 방식($KF_{128,60}$)을 사용할 때보다 시계열 교차검증 방식($TS_{128,30}$)을 사용할 때 MAPE는 0.61% 높아 그 차이가 유의미 하지 않지만 RMSE는 약 9.06% 낮아 시계열 교차검증을 통해 개발된 모델이 더 높은 정확도를 보임을 확인하였다.

5. 결 론

본 연구에서는 시계열 교차검증을 적용하여 2,3-BDO 생산 공정의 온도 예측 모델의 배치 개수와 반복횟수를 조정하고 K 겹 교차검증과 비교하였다. 최종적으로 선택된 모델에 대해 MAPE는 두 방식에서 차이가 1%미만 이었으나 RMSE는 시계열 교차검증에서 더 높은 정확도를 보였다. 또한 K 겹 교차검증을 사용할 때보다 시계열 교차검증을 사용할 때 더 적은 학습 및 검증시간이 소요되었다. 결론적으로 시간에 따라 데이터가 다이나믹하게 변하는 화학공정에서는 시간이 많이 소요되는 K 겹 교차검증보다 시계열 교차검증을 이용하여 초매개변수를 조정한다면 모델의 신뢰도를 높이면서도 효율적으로 모델을 개발할 수 있음을 확인하였다. 본 연구에서 다양한 초매개변수 중 배치 개수와 반복횟수에 대한 사례연구를 진행하여 테스트 결과에서 실제 데이터와의 오차가 발생하였는데, 예측 모델의 정확도를 높이기 위해 다양한 초매개변수를 조정하는 후속 연구가 필요하다.

감 사

본 논문은 한국생산기술연구원 “기업체 에너지공정 최적화 지원사업(EM-21-0022)” 및 “화학산업 고도화를 위한 스마트 제조공정 AI 플랫폼 기술 개발(JH-21-0005)”의 지원으로 수행한 연구입니다.

References

1. Oh, K. C., Kwon, H., Roh, J., Choi, Y., Park, H., Cho, H. and Kim, J., “Development of Machine Learning-Based Platform for Distillation Column,” *Korean Chem. Eng. Res.*, **58**(4), 565-572 (2020).
2. Hoon, S., Ah, Y. and Hyeong, J., “A Machine Learning Model for Predicting Silica Concentrations through Time Series Analysis of Mining Data,” *J. Korean Soc. Qual. Manag.*, **48**(3), 499-508(2020).
3. Zhai, N., Yao, P. and Zhou, X., “Multivariate Time Series Forecast in Industrial Process Based on XGBoost and GRU,” in, IEEE, ITAIC 2020 - IEEE 9th Joint International Information Technology and Artificial Intelligence Conferencepp. 1397-1400.
4. Lee, Y., Choi, Y., Cho, H. and Kim, J., “Prediction of Distillation Column Temperature Using Machine Learning and Data Preprocessing,” *Korean Chem. Eng. Res.*, **59**(2), 191-199(2021).
5. Lu, Z. J., Xiang, Q., Wu, Y. M. and Gu, J., “Application of Support Vector Machine and Genetic Algorithm Optimization for Quality Prediction Within Complex Industrial Process,” *Proceeding - 2015 IEEE Int. Conf. Ind. Informatics, INDIN 2015*, 98-103(2015).
6. Wu, H. and Zhao, J., “Deep Convolutional Neural Network Model Based Chemical Process Fault Diagnosis,” *Comput. Chem. Eng.*, **115**, 185-197(2018).
7. Eslamloueyan, R., “Designing a Hierarchical Neural Network Based on Fuzzy Clustering for Fault Diagnosis of the Tennessee-Eastman Process,” *Appl. Soft Comput. J.*, **11**(1), 1407-1415(2011).
8. Wei, Y. and Weng, Z., “Research on TE Process Fault Diagnosis Method Based on DBN and Dropout,” *Can. J. Chem. Eng.*, **98**(6), 1293-1306(2020).
9. Jing, C. and Hou, J., “SVM and PCA Based Fault Classification Approaches for Complicated Industrial Process,” *Neurocomputing*, **167**, 636-642(2015).
10. Wang, T., Gao, H. and Qiu, J., “A Combined Adaptive Neural Network and Nonlinear Model Predictive Control for Multirate Networked Industrial Process Control,” *IEEE Trans. Neural Networks Learn. Syst.*, **27**(2), 416-425(2016).

11. Mazinan, A. H., "A New Algorithm to AI-based Predictive Control Scheme for a Distillation Column System," *Int. J. Adv. Manuf. Technol.*, **66**(9-12), 1379-1388(2013).
12. Mahdi, M. and Mehdi, B., A systematic review on overfitting control in shallow and deep neural networks, Springer Netherlands(2021).
13. Arlot, S. and Celisse, A., "A Survey of Cross-validation Procedures for Model Selection," *Stat. Surv.*, **4**, 40-79(2010).
14. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*,(2014).
15. Ying, X., "An Overview of Overfitting and its Solutions," *J. Phys. Conf. Ser.*, **1168**(2), (2019).
16. Bergmeir, C. and Benítez, J. M., "On the Use of Cross-validation for Time Series Predictor Evaluation," *Inf. Sci. (Ny).*, **191**, 192-213(2012).
17. Chen, X., Chen, X., She, J. and Wu, M., "A Hybrid Time Series Prediction Model Based on Recurrent Neural Network and Double Joint Linear-nonlinear Extreme Learning Network for Prediction of Carbon Efficiency in Iron Ore Sintering Process," *Neurocomputing*, (2017).
18. Zhao, J., Wang, W. and Sheng, C., Data-driven prediction for industrial processes and their applications, (2018).
19. Benesty, J., Chen, J., Huang, Y. and Cohen, I., Pearson Correlation Coefficient, (2009).
20. Andreas C. M. and Sarah, G., thirdIntroduction to machine learning with python, O'Reilly(2020).
21. Hochreiter, S. and Urgen Schmidhuber, J., "Long Shortterm Memory," *Neural Comput.*, (1997).
22. Brownlee, J., "What is the Difference Between a Batch and an Epoch in a Neural Network?," *Mach. Learn. Mastery*, (2018).
23. Kingma, D. P. and Ba, J. L., "Adam: A Method for Stochastic Optimization," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.