Korean Chem. Eng. Res., 59(2), 191-199 (2021)
https://doi.org/10.9713/kcer.2021.59.2.191
PISSN 0304-128X, EISSN 2233-9558

머신 러닝과 데이터 전처리를 활용한 증류탑 온도 예측

이예찬*,** · 최영렬*,*** · 조형태* · 김정환*,†

*한국생산기술연구원 친환경재료공정연구그룹
44413 울산광역시 중구 종가로 55
**서울과학기술대학교 화공생명공학과
01811 서울특별시 노원구 공릉로 232
***연세대학교 화공생명공학과
03722 서울특별시 서대문구 연세로 50
(2020년 10월 14일 접수, 2020년 12월 14일 수정본 접수, 2021년 1월 4일 채택)

Prediction of Distillation Column Temperature Using Machine Learning and Data Preprocessing

Yechan Lee*,, 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, Seoul National University of Science and Technology, 232, Gongneung-ro, Seoul, 01811, Korea*
****Department of Chemical and Biomolecular Engineering, Yonsei University, 50, Yensei-ro, Seoul, 03722, Korea*
Received 14 October 2020; Received in revised from 14 December 2020; Accepted 4 January 2021)

요 약

화학 공정의 주요 설비 중 하나인 증류탑은 물질들의 끓는점 차이를 이용하여 혼합물에서 원하는 생산물을 분리하는 설비이며 증류 공정은 많은 에너지가 소비되기 때문에 최적화 및 운전 예측이 필요하다. 본 연구의 대상 공정은 공급처에 따라 원료의 조성이 일정하지 않아 정상 상태로 운전이 어려워 효율적인 운전이 어렵다. 이를 해결하기 위해 데이터 기반 예측 모델을 이용하여 운전 조건을 예측 할 수 있다. 하지만 미가공 공정 데이터에는 이상치 및 노이즈가 포함되어 있어 예측 성능을 향상시키기 위해 데이터 전처리가 필요하다. 본 연구에서는 인공 신경망 모델인 Long short-term memory (LSTM)과 Random forest (RF)를 사용하여 모델을 최적화한 후, 데이터 전처리 방법으로 Low-pass filter와 One-class support vector machine을 사용하여 데이터 전처리 방법 및 범위에 따른 예측 성능을 비교하였다. 각 모델의 예측 성능과 데이터 전처리의 영향은 R^2과 RMSE를 사용하여 비교하였다. 본 연구의 결과, 전처리를 통해 LSTM의 경우 R^2은 0.791에서 0.977으로 RMSE는 0.132에서 0.029로 각각 23.5%, 78.0% 향상되었고, RF의 경우 R^2은 0.767에서 0.938으로 RMSE는 0.140에서 0.050으로 각각 22.3%, 64.3% 향상되었다.

Abstract – A distillation column, which is a main facility of the chemical process, separates the desired product from a mixture by using the difference of boiling points. The distillation process requires the optimization and the prediction of operation because it consumes much energy. The target process of this study is difficult to operate efficiently because the composition of feed flow is not steady according to the supplier. To deal with this problem, we could develop a data-driven model to predict operating conditions. However, data preprocessing is essential to improve the predictive performance of the model because the raw data contains outlier and noise. In this study, after optimizing the predictive model based long-short term memory (LSTM) and Random forest (RF), we used a low-pass filter and one-class support vector machine for data preprocessing and compared predictive performance according to the method and range of the preprocessing. The performance of the predictive model and the effect of the preprocessing is compared by using R^2 and RMSE. In the case of LSTM, R^2 increased from 0.791 to 0.977 by 23.5%, and RMSE decreased from 0.132 to 0.029 by 78.0%. In the case of RF, R^2 increased from 0.767 to 0.938 by 22.3%, and RMSE decreased from 0.140 to 0.050 by 64.3%.

Key words: Data preprocessing, low-pass filter, one-class support vector machine, distillation column, machine learning, random forests, Long-short term memory

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

1. 서 론

화학 산업의 주요 설비 중 하나인 증류탑은 혼합물의 끓는 점 차이를 이용하여 목표로 하는 물질을 분리하기 위해 많은 에너지가 사용되므로 공정 최적화 및 운전 예측이 필요하다. 이를 위해 많은 연구자들이 다양한 연구를 수행하고 있으며 크게 열역학 방정식을 바탕으로 한 이론 기반 모델과 데이터 간의 상관관계를 이용한 데이터 기반 모델로 구분할 수 있다.

이론 기반 모델은 열역학 법칙을 기반으로 공정의 운전 조건을 계산하는 방법으로 Aspen plus[1], PRO/II[2] 등 공정 모사 프로그램을 이용한다. 이론 기반 모델은 물질 및 에너지 수지 등 열역학 법칙을 만족하지 않은 경우 적용이 어렵기 때문에 비정상 상태로 운전되는 실제 공정의 운전 조건을 예측하기 위한 방법으로 적합하지 않다. 데이터 기반 모델은 데이터들 간의 상관관계를 바탕으로 공정의 운전조건을 예측하는 방법으로 실제 공정 데이터를 사용하기 때문에 비정상 상태의 조건에서도 예측할 수 있는 장점이 있어 많은 연구가 이루어지고 있다.

높은 성능의 예측 모델을 개발하기 위해서는 예측하고자 하는 공정과 데이터 유형에 적합한 예측 모델을 선택하고 내부 매개변수들을 최적화하는 것이 중요하기 때문에 여러 가지 데이터 기반 모델이 연구되고 있다. 일반적으로 데이터 기반 모델은 서포트 벡터 머신 (SVM)[3] 등의 회귀 모델, 랜덤 포레스트 (RF)[4] 등의 앙상블 모델, 순환신경망 (RNN)[5], 장단기 기억 (LSTM)[4] 등의 인공신경망 모델로 분류된다. 회귀 모델은 입력값과 출력값 간의 일반적인 관계 특성을 도출하는 방식으로 작동하며 수치 형태로 존재하는 데이터를 다룰 때 효과적이지만 비선형 데이터나 이상치의 영향을 크게 받는 단점이 있다. 앙상블 모델은 회귀 모델을 여러 개의 샘플 모델로 만든 다음 결합함으로써 회귀 모델보다 이상치에 강건하고 계산속도가 빠르다는 장점이 있다. 인공신경망 모델은 입력 값과 출력값의 상관관계를 은닉층에서 분석하여 가장 적합한 추론값을 도출하는 모델로 상관관계가 변하는 데이터들을 분석할 수 있는 장점이 있다. 증류탑은 비정상 상태로 운전되며 시간에 따라 데이터의 상관관계와 특성이 달라지는 시계열 데이터이기 때문에 이러한 데이터를 다룰 수 있는 앙상블 모델과 인공신경망 모델이 적합하다[4,5].

Lee와 Lee[4]는 시계열 데이터인 초미세먼지 농도를 예측하기 위해 LSTM 및 RF 기반 예측 모델을 설계하였다. 예측 모델의 성능을 비교한 결과, RF 기반 모델이 LSTM보다 높은 예측 성능과 안정성을 보였다. Vijaya Raghavan[5] 등은 증류탑 생성물의 조성을 예측하기 위해 RNN과 FNN을 사용하여 소프트 센서를 개발하고 이 센서가 비정상 상태에 안정적으로 적용 가능함을 확인하였다.

하지만 실제 공정 데이터에는 여러 변수에 의해 이상치와 노이즈 같은 예측 성능을 저하시키는 데이터들이 포함되어 있다. 따라서 모델의 예측 성능을 향상시키기 위해 이상치와 노이즈를 적절히 제거할 수 있는 전처리 방법을 선정하고 최적화하는 것도 중요하다. 이상치 제거 방법으로 사분범위 (Interquartile range)[6], 표준점수 (z-score)[7] 등의 모수적 방법과 OCSVM[8], Isolation forest[9] 등과 같은 비모수적 방법이 사용된다. Howsalya Devi [6] 등은 사분범위를 사용하여 질병 진단 데이터의 이상치를 제거하고 예측 모델의 정확성을 개선하였다. Erkuş와 Purutcuoğlu[7]는 임의로 발생시킨 시계열 데이터에 표준점수을 비롯한 이상치 감지 방법을 적용하고 감지 성능을 비교하여 시계열 데이터에 적용 가능한 최적의 이상치 제거 방법을 제안하였다.

하지만 모수적 방법은 데이터가 정규분포를 가진다는 가정을 사용하기 때문에 정규분포를 따르지 않는 데이터에서는 효율적으로 이상치를 제거하지 못하는 단점이 있다. 비모수적 방법은 모집단을 모수 분포로 가정하지 않고 알고리즘을 통해 데이터를 예측하는 방법으로 최소한의 가정만을 사용하기 때문에 오류의 가능성이 낮으며 비선형 데이터에 효과적으로 적용할 수 있다. 따라서 증류탑의 데이터는 정규 분포를 따르지 않기 때문에 이상치 제거를 위해 모수적 방법보다 비모수적 방법을 적용하는 것이 적합하다. Zhang[8] 등은 교통 데이터에 포함된 이상치를 OCSVM을 통해 제거하였다. OCSVM을 이용한 이상치 제거를 통해 인공지능을 이용한 시스템의 성능이 기존보다 향상됨을 확인하였다. Kim[9] 등은 수질관리 분야 데이터의 이상치 제거를 위해 비모수 방법인 Isoloation forest를 적용하여 통계적 분포가 명확하지 않은 데이터의 경우 모수적인 방법보다는 비모수적 방법이 더 높은 이상치 제거 성능을 보인다는 것을 검증하였다.

노이즈 제거를 위한 방법으로는 이동평균법[10], 최소자승법[10], 지수평활법[11] 등이 있다. 이동평균법은 일정 기간 데이터에 동일한 가중치를 부여하여 예측하는 방법으로 수식이 간단하고 예측 성능이 높은 장점이 있지만 모든 데이터에 같은 가중치를 부여하여 데이터 변동이 큰 시계열 데이터에 대해 적절히 대응하지 못하는 한계가 있다. 최소자승법은 데이터의 한 지점에서 주변 데이터 간 잔차가 최소인 지점을 예측하는 방법으로 데이터의 개형을 잘 유지한다는 장점이 있지만 노이즈의 제거 효과는 다른 방법에 비해 떨어진다. 지수평활법은 이동평균법을 기반으로 작동하지만 이동평균법과 다르게 모든 시계열 자료를 사용하며 최근 시계열에 더 많은 가중치를 부여한다. 따라서 가중치를 조절하여 노이즈를 효과적으로 제거 가능하다는 장점이 있다. 증류탑 공정은 데이터의 수가 많고 변동 규모가 크기 때문에 데이터에 동일한 가중치를 주는 이동평균법이나 노이즈 제거 효과가 적은 최소자승법보다 가중치를 임의로 조절하여 노이즈를 크게 줄일 수 있는 지수평활법이 적합하다. Guinon[10] 등은 광화학 및 전기화학 반응기 데이터에 이동평균법인 Moving average filter와 최소자승법인 Savitzky-golay filter를 적용하여 노이즈를 제거하였다. 필터 적용 결과, 두 필터 모두 노이즈를 효과적으로 제거하였으며 Savitzky-golay filter가 더 좋은 성능을 가짐을 확인하였다. Hang [11] 등은 관성 항법 데이터에 포함된 노이즈를 제거하기 위해 지수평활법인 LPF를 사용하고 노이즈 제거로 인한 데이터 정확도 상승을 확인하였다.

본 연구의 대상 공정은 혼합부탄 증류공정으로 원료의 조성이 공급처에 따라 크게 변하는 문제가 있어 공정 운전조건이 불안정하여 불필요한 에너지가 소비되고 생산 수율이 낮은 문제가 있다. 이를 해결하기 위해 운전 예측 및 제어 방법으로 데이터 기반 예측 모델을 개발하였지만 공정 데이터에 많은 이상치와 노이즈가 포함되어 있기 때문에 예측 성능이 높지 못한 문제가 있었다. 따라서 본 연구에서는 예측 모델의 성능을 향상시키기 위한 노이즈 제거 방법으로 LPF와 이상치 제거 방법으로 OCSVM을 이용하여 학습데이터의 전처리를 진행하고 이를 인공신경망 기법의 LSTM과 앙상블 기법의 RF에 적용하여 최적의 성능을 보이는 예측 모델을 개발하고자 하였다.

본 연구의 순서는 아래와 같다. 제 2장에서 공정 개요 및 연구 개발 과정에 대해 설명하고 본 연구에서 사용한 전처리 및 학습모델을 작성하였다. 제 3장에서는 LSTM 및 RF에 대한 모델 최적화를

진행한 후, 각 모델에 대해 전처리 방법 및 범위에 따른 예측 성능을 사례연구를 통해 비교하여 본 공정에 대한 최적의 예측 성능을 보이는 예측 모델 및 전처리 방법을 도출하였다. 마지막으로 제 4장은 본 연구의 결론을 작성하였다.

2. 연구 방법

2-1. 공정 개요

Fig. 1은 대상 공정인 증류 공정의 개략도이다. 대상 공정은 실제 가동 중인 혼합부탄 분리공정으로 총 78단으로 이루어져 있다. 노말부탄과 아이소부탄이 포함된 혼합부탄 원료는 35단으로 유입되고 주 생산품인 노말부탄은 64단으로 분리되며 노말부탄의 목표 순도는 99%이다. 현재 원료의 노말부탄 함량은 공급처에 따라 60 ~ 98%로 크게 변하기 때문에 공정 운전이 불안정하여 Fig. 2와 같이 원료 유량과 64단의 온도가 일정하지 않은 문제가 있다. 이를 예측하기 위해 본 연구에서는 30초 간격으로 측정된 20,073개의 실제 데이터를 사용하여 예측 모델에 활용하였다.

2-2. 데이터 기반 모델 설계

Fig. 3은 본 연구의 연구 개발 흐름으로 가장 먼저 목표로 하는 64단 온도 예측을 위한 변수 선정, 이상치 및 노이즈 제거를 위한 데이터 전처리, 공정에 적합한 예측 모델 개발을 위한 예측 모델 최적화, 데이터 전처리에 따른 예측 성능 비교를 위한 예측 성능 평가로 이루어져 있으며 각 부분에 대한 자세한 내용은 다음과 같다.

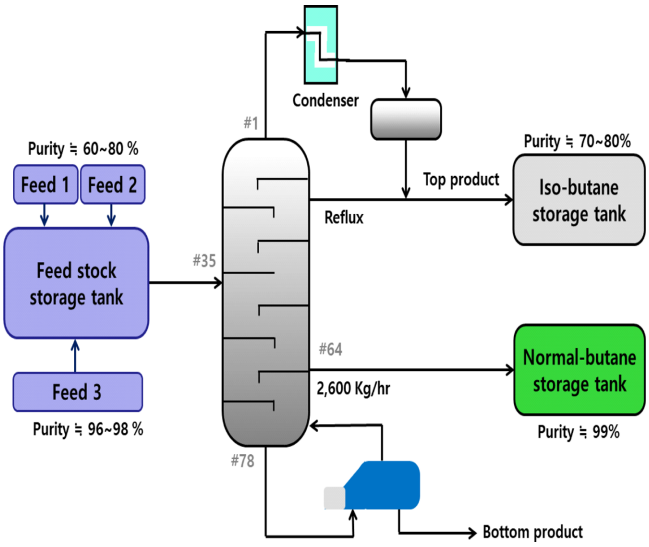

Fig. 1. Diagram of distillation column.

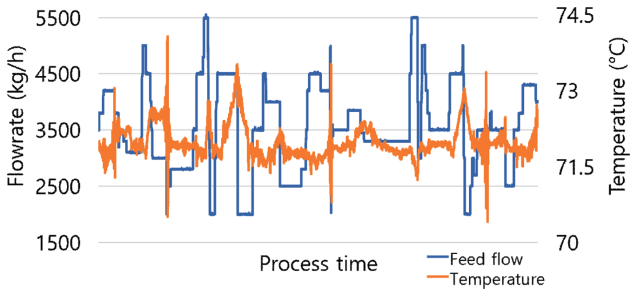

Fig. 2. Trend of #64 stage temperature and feedflow.

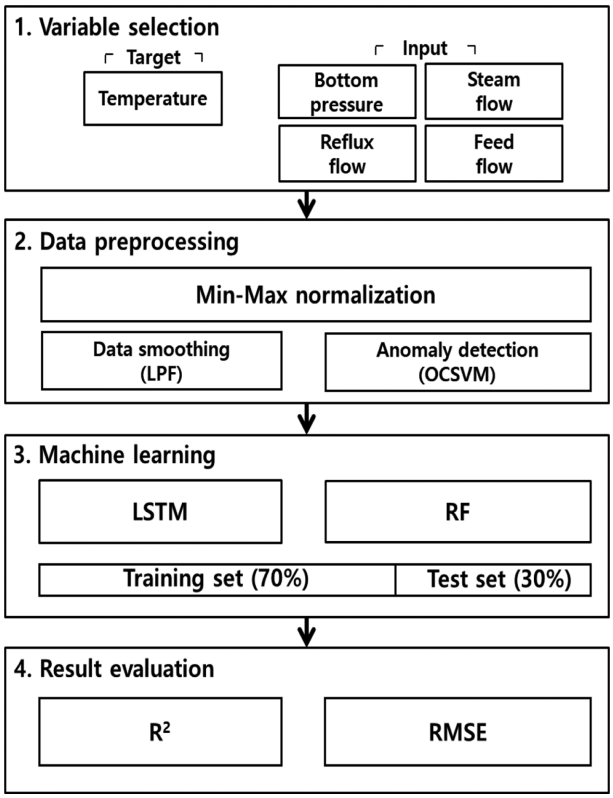

Fig. 3. Flow chart of model development.

2-2-1. 단계 1: 변수 설정

예측 모델이 높은 성능을 갖기 위해서는 입력변수와 예측하고자 하는 64단 온도가 높은 상관관계를 가져야 한다. 따라서 본 연구에서는 입력 변수를 설정하기 위해 피어슨 상관계수를 사용하였다. 피어슨 상관계수는 변수 간 선형 상관관계를 계량화한 수치로 –1에서 +1까지의 범위를 가지며 절댓값이 클수록 높은 상관관계를 의미한다. 일반적으로 상관계수가 0.3 이상일 경우 데이터들의 상관관계가 있다고 판단하며 상관계수를 분석한 결과 예측 변수인 64단 온도와 높은 상관관계를 가지는 스팀 유량, 환류 유량, 증류탑 하단 압력과 상관계수는 낮지만 증류탑 운전에 중요 인자인 원료 유량을 입력 변수로 선정하였으며 그 구성은 Table 1과 같다.

2-2-2. 단계 2: 데이터 전처리

선정된 입력변수는 각각 다른 크기를 가지기 때문에 예측 모델에 사용하려면 하나의 단위로 통일하는 데이터 정규화가 필요하다. 본 연구에서는 최소-최대 정규화를 통해 데이터를 정규화 하였다. 최소-최대 정규화는 식 (1)로 나타내며 데이터를 모델 학습에 적합한 형태인 [0,1] 사이의 값으로 변환한다[12].

Table 1. Pearson's correlation coefficient of the input variables

Type	Variable	Pearson's correlation coefficient
Input variable	Feed flow	-0.141
	Steam flow	0.630
	Reflow flow	0.649
	Bottom pressure	0.349
Target variable	#64 temperature	1.000

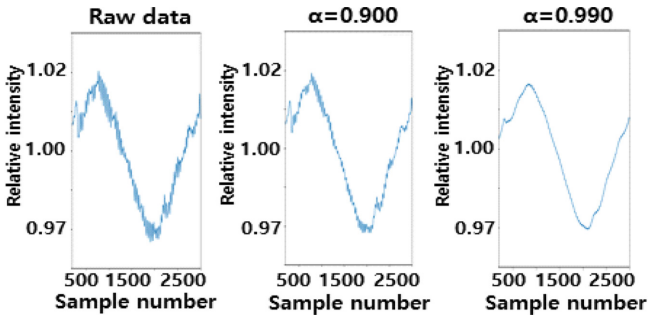

Fig. 4. Description of noise reduction with LPF.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

데이터 정규화 후 노이즈 제거를 위해 LPF와 이상치 제거를 위해 OCSVM을 적용하여 데이터 전처리를 수행하였으며 각 방법에 대한 설명은 아래와 같다.

먼저 LPF는 Fig. 4와 같이 가중치 (α)를 조절하여 특정 주파수보다 낮은 주파수 대역은 통과시키고, 이보다 높은 주파수 대역은 차단하여 노이즈를 제거하는 디지털 필터링 방법으로, 공정 데이터에 포함된 노이즈는 주파수가 높은 고주파 형태로 판단할 수 있기 때문에 LPF를 사용하면 얻고자 하는 데이터의 개형인 저주파 형태의 데이터만 통과시켜 노이즈를 제거할 수 있다. 본 연구에서 사용한 1차 LPF 식은 (2)와 같으며 재귀적으로 계산할 수 있는 무한 임펄스 응답 필터링 유형으로 사용된다[13].

$$\bar{x}_k = \alpha\bar{x}_{k-1} - (1-\alpha)x_k \quad (2)$$

α는 LPF의 가중치이며 0과 1사이의 값을 가진다. $\bar{x}_k$ 와 x_k는 각각 노이즈가 제거된 데이터의 추정치와 실제 데이터가 측정된 측정치를 의미하며 아래첨자 k는 데이터의 순서를 의미한다. 노이즈가 제거된 추정치 ($\bar{x}_k$)는 이전 추정치 ($\bar{x}_{k-1}$)와 측정치 (x_k)에 의해 업데이트되며 가중치가 클수록 측정치의 반영률이 작아져서 노이즈를 효과적으로 제거할 수 있다. 하지만 너무 높은 가중치를 설정하게 되면 원래 그래프와 다른 개형이 되기 때문에 사례연구를 통해 사용하는 데이터에 대해 적합한 가중치 값을 찾아야 한다.

Fig. 5는 OCSVM의 원리를 나타낸 개요이며, OCSVM는 비모수적 기반의 학습 알고리즘으로 커널 함수를 통해 데이터에 결정 경계를 생성하고 이를 벗어나는 데이터를 이상치로 분류하여 제거하는 방법이다[8]. OCSVM의 커널 함수는 선형 함수 (linear function), 동차 다항식 함수 (Polynomial function), 방사 기저 함수 (Radial basis function) 등이 있으며 본 연구에서는 방사 기저 함수를 사용하였다.

OCSVM의 수식은 LPF와 다르게 복잡한 수식으로 본 연구의 범위에 벗어난 주제라 판단하여 자세한 설명은 생략하였다. LPF의 가중치와 마찬가지로 OCSVM은 ν와 γ라는 두 가지 변수를 이용하여 이상치를 탐지하고 제거한다. ν는 데이터에서 이상치의 비율을 조절하는 매개변수로 ν가 클수록 많은 데이터를 이상치로 판단하며 γ는 결정 경계의 곡률을 정하는 파라미터로 γ가 높아지면 각각의 데이터 포인터가 영향력을 끼치는 거리가 짧아진다. 따라서 γ가 낮아지면 결정 경계에 영향을 끼치는 데이터가 많아지므로 결정 경계가 변형되어 세밀한 분류 성능을 갖는다. OCSVM의 이상치 분류 성능은 ν와 γ의 값에 따라 달라지므로 이를 적절하게 설정해야 한다.

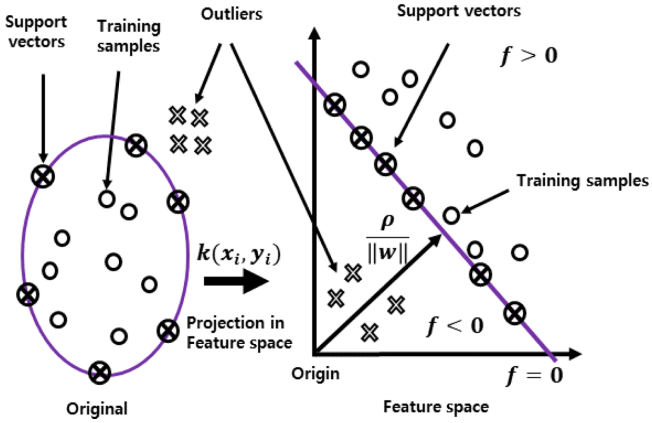

Fig. 5. Description of anomaly detection with OCSVM.

2-2-3. 단계 3: 예측 모델 최적화

학습 모델은 시계열 데이터 분석에 사용되는 인공신경망 기반 LSTM과 회귀 분석에 사용되는 의사결정 트리 기반 RF로 선정하였다[5].

LSTM은 데이터를 순차적 학습하고 예측하는 순환 신경망에 메모리 셀을 도입한 학습 모델이며 Fig. 6은 LSTM의 알고리즘을 나타낸다. LSTM의 알고리즘에는 입력 게이트, 출력 게이트 및 망각 게이트의 세 가지 게이트가 있으며 이를 통해 정보를 추가 또는 삭제할 수 있다. LSTM의 코어는 게이트를 통해 들어오는 연속 셀 (C_t)로 구성되며 이는 컨베이어 벨트라고도 불린다. 게이트는 정보를 연속 셀로 전송하여 데이터를 제거하거나 학습을 지속한다. f_t는 망각 게이트 벡터로 이전 세포 상태를 기억하기 위한 게이트 가중치다. i_t는 입력 게이트 벡터로 새로운 정보를 얻기 위한 게이트 가중치다. o_t는 출력 게이트 벡터로 출력 후보를 선택하는 역할을 한다. X_t는 입력 벡터, h_t는 출력 벡터다. c_t는 셀의 상태 벡터이고 W는 훈련 중 학습되는 파라미터 행렬 및 벡터로서 각 게이트 및 셀의 가중치와 관련된다. LSTM 장치의 이전 단계와 업데이트는 다음과 같이 공식화된다.

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f\boldsymbol{X}_t + \boldsymbol{W}_f\boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \quad (7)$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i\boldsymbol{X}_t + \boldsymbol{W}_i\boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \quad (8)$$

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_c\boldsymbol{X}_t + \boldsymbol{W}_c\boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \quad (9)$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t\boldsymbol{c}_{t-1} + \boldsymbol{i}_t\tilde{\boldsymbol{c}}_t \quad (10)$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o\boldsymbol{X}_t + \boldsymbol{W}_o\boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \quad (11)$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t\tanh(\boldsymbol{c}_t) \quad (12)$$

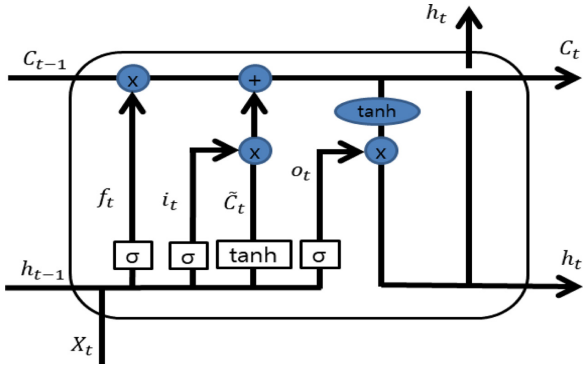

Fig. 6. Description of machine learning based LSTM.

LSTM의 예측 성능을 결정하는 초매개변수는 배치 사이즈 (batch size), 에포크 (Epoch), 셀 개수 (cell number), 학습률 (Learning rate), 활성화 함수 (Activation function) 및 최적화 방법 (Optimizer) 등이 있다. 본 연구에서는 LSTM 모델을 확인하기 위해 배치 사이즈를 제외한 모든 초매개변수는 이전 연구의 값을 사용하였으며 설정한 값은 Table 2에 요약하였다[14-16]. 최적화 기법과 활성화 함수으로 Adam과 Elu을 사용하였으며 학습률은 0.001, 셀의 개수는 20으로 설정하였고 에포크 수는 최대 200으로 설정하였다. 학습의 과적합 방지를 위해 조기 종료 (Early Stopping) 기법을 설정하였으며 여러 작동 기준 중 최소의 예측 오차 계산 후 얼마나 더 작동할지를 정하는 Patience를 이용하였고 기준을 5로 설정하였다.

Fig. 7은 RF를 사용한 예측 모델의 대한 개요이며, 의사결정 트리 (decision tree)를 기반으로 기존 데이터에서 중복을 허용하여 원 데이터셋과 같은 크기의 데이터셋을 만드는 부트스트랩 (bootstrap)을 적용한 앙상블 기법이다. RF는 부트스트랩을 통해 N개의 트리를 샘플링하여 학습한 후 예측 결과치의 평균을 계산한다. 이 과정을 배깅 (bagging)이라 하며 이를 수식으로 표현하면 다음과 같다.

$$\hat{f}_{avg}(x) = \frac{1}{N}\sum_{n=1}^{N}\hat{f}^{*n}(x) \tag{13}$$

$\hat{f}^{*n}(x)$ 은 부트스트랩을 통해 얻은 N 번째 의사결정 트리의 예측치이며 이 값의 평균인 $\hat{f}_{avg}$ 가 RF의 결과이다. 의사결정 트리 기반 알고리즘은 기법 특성상 데이터 변동에 따른 분산의 변화가 크다는 단점이 있는데 RF는 배깅을 이용하여 개별 트리의 높은 분산을 여러 개의 트리가 나누어 이를 해결한다. 따라서 RF는 이상치에 강건하며 연산이 빠르다는 장점을 갖는다.

또한 RF는 변수의 임의 추출 기법을 사용한다. 각 트리의 분할 기준이 동일하다면 목표 변수와 상관관계가 높은 변수가 집중 학습된다. 따라서 예측치의 다양성을 보장하기 위해 노드를 분할할 때 총 P개의 입력 변수 중 M개의 설명 변수를 중복을 허용하여 새로 선택한다. 이외에도 랜덤 포레스트 안에 있는 각 의사결정 트리의 깊이를 설정하고 분류 속성의 기준을 정해야한다. 분류 속성의 기준은 엔트로피 (Entropy)와 지니계수 (Gini index)가 있다.

RF의 성능은 의사결정 트리 수 (N)과 설명변수의 개수 (M)에 따라 달라지므로 우수한 예측 성능을 갖도록 적절하게 조정해야 한다. 본 논문에서는 RF의 의사결정 트리 수 (N)을 1~300 까지 사례연구를 진행하였으며 설명변수 (M)은 일반적으로 입력 변수 개수의 제곱근으로 설정하므로 2로 설정하였다[17]. 노드 분할 기준을 의미하는 정보량 지수는 Gini로 설정하고 분류 정밀도를 나타내는 의사결정 트리 깊이는 제한을 두지 않았으며, RF 초매개변수 설정값은 Table 3에 요약하였다.

Table 2. Hyperparameters for LSTM

Name	Value
Batch size	1 (2^0) ~ 512 (2^9)
Activation function	Elu
Optimizer function	Adam
Number of cell	20
Learning rate	0.001
Epochs	200
Early stopping (patience)	5

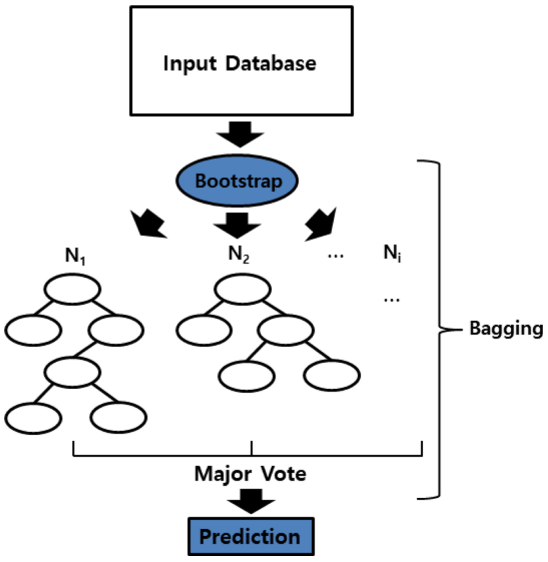

Fig. 7. Description of machine learning based RF.

2-2-4. 단계 4: 예측 성능 평가

예측 모델의 성능을 검증하기 위해 20,072 개의 공정 데이터를 모델의 학습을 위한 훈련 데이터 (Training data)와 모델이 잘 학습되었는지 평가하기 위한 테스트 데이터 (Test data)로 사용하였다. 일반적으로 훈련 데이터와 테스트 데이터는 7:3의 비율을 갖는 구간으로 나누며, 구간을 변경하여 예측 성능을 검증하는 교차검증 (Cross validation)을 사용한다. 하지만 본 연구의 공정 데이터는 Fig. 2와 같이 기간에 따라 데이터 변화폭이 크기 때문에 일정 구간으로 나누어서 학습 및 테스트를 진행하면 일부 구간에서 예측 성능이 떨어질 수 있다. 따라서 본 연구에서는 학습 데이터를 전체 데이터 중 임의로 70%를 사용하고, 나머지 30%를 학습 데이터로 사용하였다.

예측 성능을 검증하기 위해 예측 결과가 실제 데이터와 얼마나 일치하는 경향을 보이는지 판단하는 정확도와 그 차이가 얼마나 되는지 판단하는 정밀도를 확인하였다[18,19]. 정확도와 정밀도를 판단하는 지표로는 결정 계수 (R^2)와 평균 제곱근 오차 (RMSE)를 사용하였으며 식은 아래와 같다.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2} \tag{14}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{15}$$

Table 3. Hyperparameters for RF

Name	Value
Tree number (N)	1 ~ 300
Number of attributes (M)	2
Max. depth	None
Criterion	Gini
Bootstrap	True

3. 사례 연구 결과

3-1. 연구 방법

본 연구에서는 전처리 방법인 LPF 및 OCSVM과 예측 모델인 LSTM 및 RF을 적용한 사례연구를 통해 예측 성능이 높은 최적 매개변수를 결정하고자 하였다. 예측 모델은 초기 가중치에 따라 예측 성능이 근소하게 변화하므로 개발한 예측 모델의 견고성을 확인하기 위해 각각 10회씩 예측을 실시하였고 그 중간값을 비교하였다.

3-2. 결과

3-2-1. 학습 모델 최적화

Table 4은 배치 사이즈에 따른 LSTM 모델의 예측 성능과 계산 시간을 나타낸 것이다. 연구 결과 배치 사이즈가 작아질수록 모델의 학습 횟수가 많아지기 때문에 시간이 증가하는 것을 확인할 수 있었으며, 사례연구 결과 가장 적합한 배치 사이즈로 128일 때 가장 높은 예측 성능을 보였다.

Table 5 및 Fig. 9는 RF에서 사용할 의사결정 트리수를 정하기 위해 초매개변수의 변화에 따른 모델 성능을 나타낸 것이다. 사례연구 결과 의사결정 트리 수가 50보다 클 경우 성능의 개선 폭이 작아지고 훈련 시간이 크게 증가하므로 최적의 의사결정 트리 수를 50으로 설정하였다.

Table 4. The performance according to batch size

Batch size	Predictive performance		
	R^2	RMSE	Training time (s)
1 (2^0)	0.785	0.134	118.434
2 (2^1)	0.787	0.134	111.247
4 (2^2)	0.760	0.141	51.968
8 (2^3)	0.787	0.133	40.596
16 (2^4)	0.767	0.139	25.283
32 (2^5)	0.789	0.133	22.700
64 (2^6)	0.785	0.134	14.660
128 (2^7)	0.791	0.132	9.342
256 (2^8)	0.783	0.135	5.943
512 (2^9)	0.778	0.136	5.598

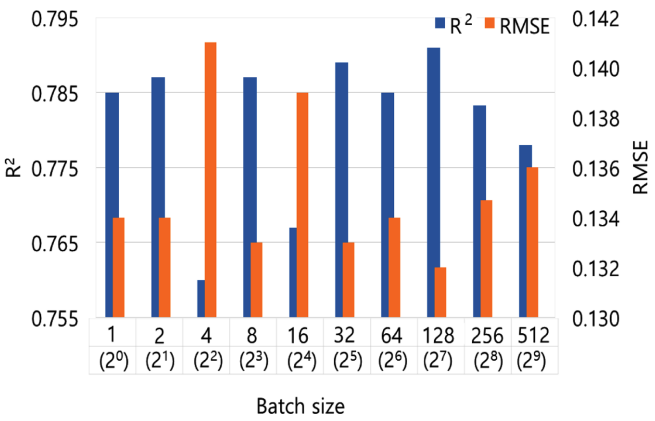

Fig. 8. The performance according to batch size.

Table 5. The performance according to tree number

Tree number	Predictive performance		
	R^2	RMSE	Training time (s)
5	0.705	0.157	0.280
7	0.722	0.153	0.360
10	0.730	0.151	0.531
20	0.744	0.146	1.008
30	0.758	0.143	1.375
50	0.767	0.140	2.508
100	0.766	0.140	10.265
200	0.766	0.140	25.609
300	0.767	0.140	43.541

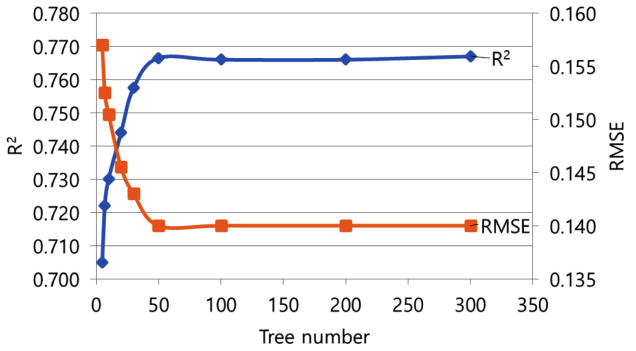

Fig. 9. The performance according to tree number.

3-2-2. 데이터 전처리 최적화를 위한 사례연구

Table 6와 Fig. 10은 LPF 적용에 따른 예측 모델의 성능 결과를 나타낸 것이다. 가중치 α가 증가함에 따라 RMSE는 감소하는 경향을 보이며 R^2은 증가하다가 최대에 도달한 이후부터 감소한다. 따라서 본 연구에서는 LSTM 및 RF의 성능이 최대치가 될 때의 가중치인 0.997과 0.996일 때를 최적의 가중치로 설정하였다.

Table 7과 Fig. 11은 OCSVM의 매개변수에 따른 모델의 예측 결과이다. 사례연구를 통해 각 ν값에서 예측 모델이 가장 높은 성능을 갖게 하는 γ값을 최적의 변수로 선정하고자 하였으며 LSTM의 경우 ν 및 γ가 5% 및 0.01으로, RF의 경우 ν 및 γ가 5% 및 0.10으로 선정되었다.

Table 6. Effect of LPF on the predictive performance

α	LSTM		RF	
	R^2	RMSE	R^2	RMSE
0.000	0.791	0.132	0.767	0.140
0.700	0.807	0.125	0.785	0.132
0.800	0.820	0.119	0.802	0.125
0.900	0.845	0.107	0.810	0.119
0.990	0.946	0.055	0.908	0.072
0.991	0.948	0.053	0.908	0.071
0.992	0.952	0.051	0.908	0.070
0.993	0.955	0.048	0.919	0.065
0.994	0.961	0.044	0.920	0.064
0.995	0.967	0.039	0.920	0.061
0.996	0.972	0.034	0.938	0.051
0.997	0.973	0.031	0.929	0.051
0.998	0.932	0.041	0.889	0.056
0.999	0.879	0.041	0.884	0.040

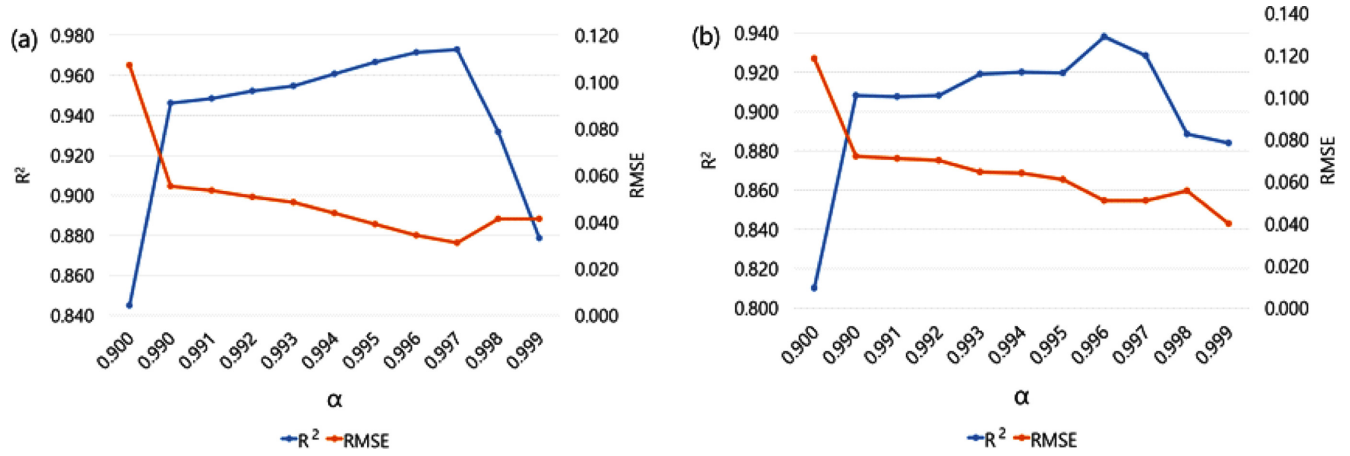

Fig. 10. (a) Effect of LPF on the LSTM performance (b) effect of LPF on the RF performance.

Table 7. Effect of OCSVM on the predictive performance

v	LSTM			RF		
	γ	R^2	RMSE	γ	R^2	RMSE
1%	0.05	0.847	0.110	0.10	0.794	0.128
2%	0.03	0.852	0.106	0.10	0.793	0.126
3%	0.03	0.866	0.100	0.05	0.795	0.124
4%	0.03	0.864	0.101	0.07	0.800	0.121
5%	0.01	0.866	0.098	0.10	0.801	0.119
6%	0.03	0.863	0.098	0.03	0.789	0.122
7%	0.05	0.853	0.098	0.10	0.775	0.130
8%	0.10	0.856	0.097	0.05	0.757	0.135
9%	0.10	0.860	0.101	0.01	0.724	0.120
10%	0.10	0.843	0.097	0.10	0.749	0.122

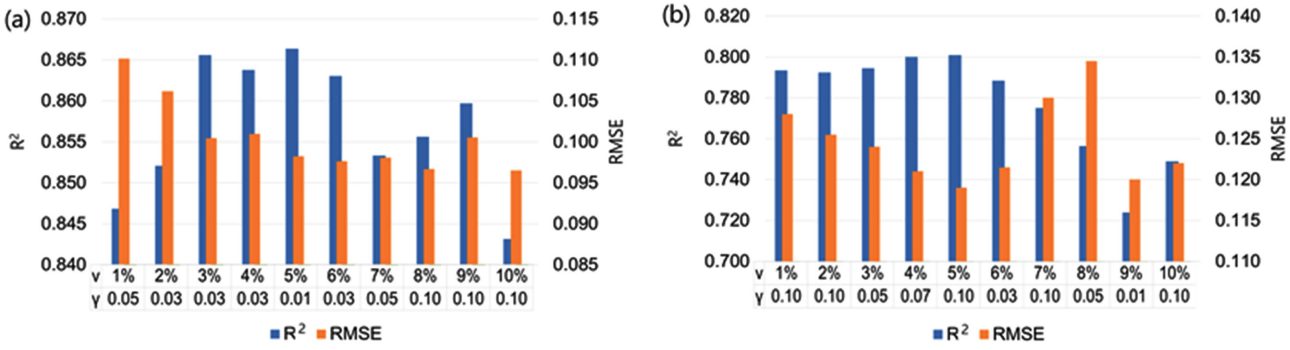

Fig. 11. (a) Effect of OCSVM on the LSTM performance (b) effect of OCSVM on the RF performance.

3-2-3. 전처리 순서에 따른 성능 결과

Table 8는 전처리 순서에 따른 LSTM 및 RF의 성능 결과이다. LPF와 OCSVM의 최적 매개변수를 적용하여 전처리를 실시하였으며 노이즈를 제거한 후 이상치를 제거한 경우와 이상치를 제거한 후 노이즈를 제거한 경우의 예측 성능을 비교하였다. LSTM과 RF 모두 노이즈를 제거 후 이상치 제거한 경우가 이상치를 제거 후 노이즈를 제거한 경우보다 높은 성능을 보였다.

Fig. 12는 모델의 예측 성능에 대한 전처리의 영향을 비교하기 위해 각 예측 모델을 10회 수행한 결과를 나타낸 그래프이다. LSTM은 RF보다 뛰어난 예측 성능을 보이지만 이상치에 의한 성능 저하가 나타난다. 반대로 RF는 이상치에 강건한 모습이 관찰되었다. 그래프를 통해 두 모델 모두 데이터 전처리에 의한 성능 개선을 보이며 특히 LSTM의 경우 성능 저하 현상이 해결된 것을 확인할 수 있다.

Table 8. Effect of data preprocessing on the performance

	Model	Predictive performance	
		R^2	RMSE
(1)	LSTM	0.791	0.132
(2)	LSTM (LPF after OCSVM)	0.968	0.033
(3)	LSTM (OCSVM after LPF)	0.977	0.029
(4)	RF	0.767	0.140
(5)	RF (LPF after OCSVM)	0.922	0.055
(6)	RF (OCSVM after LPF)	0.938	0.050

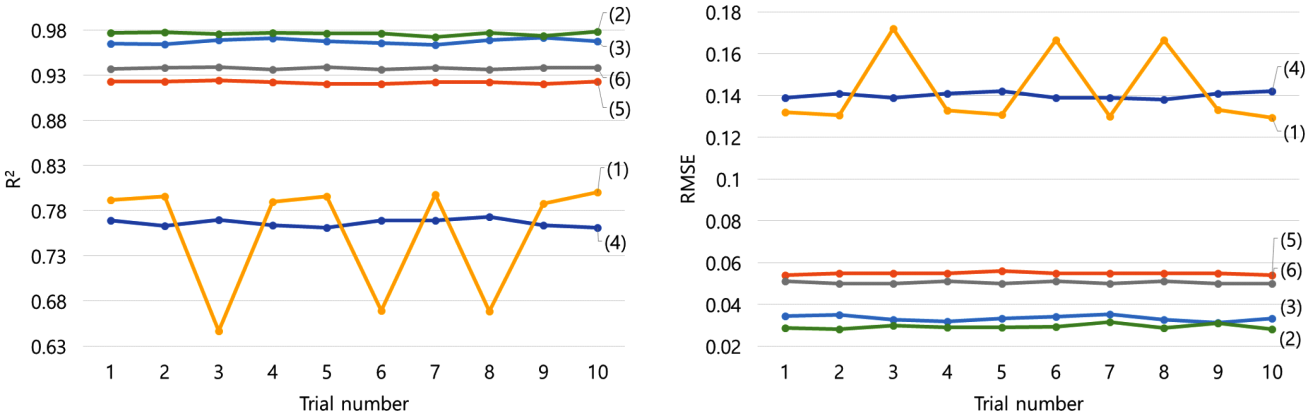

Fig. 12. Effect of data preprocessing on the predictive performance.

4. 결론 및 토의

본 연구는 혼합부탄 증류탑의 데이터 기반 온도 예측 모델 개발을 위하여 LSTM 및 RF 기반의 예측 모델을 설계하고 최적화한 후 데이터 전처리에 의한 모델의 성능 개선을 확인하였다.

모델의 최적화를 위해 LSTM과 RF의 배치 사이즈와 의사결정 트리 수를 사례연구 한 결과 LSTM은 배치 사이즈가 128일 때, RF는 의사결정 트리 수가 50일 때 가장 높은 예측 성능을 보였다.

데이터 전처리에 따른 성능 향상을 확인하기 위해 LPF의 가중치 α와 OCSVM의 ν 및 γ를 조절하여 사례 연구를 진행하였고, 각 모델에 최적화된 전처리 방법 및 범위를 제시하였다. LPF 및 OCSVM 모두 데이터 전처리 범위에 상관없이 RMSE가 개선되었고, LSTM의 경우 α가 0.997이면서 ν 및 γ가 5% 및 0.01일 때 각각 가장 높은 R^2을 보였으며, RF의 경우 α가 0.996이면서 ν 및 γ가 5% 및 0.10일 때 각각 가장 높은 R^2을 보였다.

최종적으로 이상치 및 노이즈 제거 순서에 따른 성능 변화를 확인한 결과 LSTM와 RF 모두 노이즈 제거 후 이상치 제거가 높은 성능을 보였다. 이를 통해 예측 모델 성능 향상을 위한 적합한 전처리 순서는 노이즈 제거 후 이상치 제거임을 알 수 있었으며 본 연구 결과 최종적으로 LSTM의 경우 R^2은 0.791에서 0.977으로 RMSE는 0.132에서 0.029로 각각 23.5%, 78.0% 향상되었고, RF의 경우 R^2은 0.767에서 0.938로 RMSE는 0.140에서 0.050으로 각각 22.3%, 64.3% 향상되었다.

본 연구에서 제시된 성능 수치는 실제 공정은 예측 데이터의 외란 및 이상치 특성에 따라서 달라질 수 있고, 데이터에 따라서 최적의 성능을 나타내지 않을 수 있다. 본 연구의 결과를 바탕으로 데이터 특성 변화를 감지하며 일정한 성능을 유지할 수 있도록 주기적인 업데이트가 가능한 실시간 예측 모델을 개발을 위한 추후 연구를 진행할 예정이다.

감 사

본 논문은 한국생산기술연구원 민간수탁활성화지원사업 “기업체 에너지공정 최적화 지원 사업(KITECH EE-20-0019)” 및 기획재정부 제조혁신지원사업 “화학산업 고도화를 위한 스마트 제조공정 AI 플랫폼 기술 개발(KITECH 21-0005)”의 지원으로 수행한 연구입니다.

Reference

1. Kartal, F. and Özveren, U., “A Deep Learning Approach for Prediction of Syngas Lower Heating Value from CFB Gasifier in Aspen Plus®,” *Energy*, **209**, 118457(2020).
2. Sneesby, M. G., Tadé, M. O., Datta, R. and Smith, T. N., “ETBE Synthesis via Reactive Distillation. 1. Steady-State Simulation and Design Aspects,” *Ind. Eng. Chem. Res.*, **36**(5), 1855-1869 (1997).
3. Sharma, N. and Singh, K., “Neural Network and Support Vector Machine Predictive Control of Tert-amyl Methyl Ether Reactive Distillation Column,” *Syst. Sci. Control Eng.*, **2**(1), 512-526(2014).
4. Lee, D. W. and Lee, S. W., “Hourly Prediction of Particulate Matter (PM2.5) Concentration Using Time Series Data and Random Forest,” *Trans. Softw. Data Eng.*, **9**(4), 129-136(2020).
5. Vijaya Raghavan, S. R., Radhakrishnan, T. K. and Srinivasan, K., “Soft Sensor Based Composition Estimation and Controller Design for an Ideal Reactive Distillation Column,” *ISA Trans.*, **50**(1), 61-70(2011).
6. Howsalya Devi, R. D., Bai, A. and Nagarajan, N., “A Novel Hybrid Approach for Diagnosing Diabetes Mellitus Using Farthest First and Support Vector Machine Algorithms,” *Obes. Med.*, **17**, 100152(2020)
7. Erkuş, E. C. and Purutçuoğlu, V., “Outlier Detection and Quasi-periodicity Optimization Algorithm: Frequency Domain Based Outlier Detection (FOD),” *Eur. J. Oper. Res*, **291**(2), 560-574(2020).
8. Zhang, R., Zhang, S. and Muthuraman, S, J. J. “One Class Support Vector Machine for Anomaly Detection in the Communication Network Performance Data,” *5th WSEAS Int. Conf. Appl. Electromagn. Wirel. Opt. Commun. Tenerife, Spain, December 14-16* (2007).
9. Kim, J., Park, N. S., Yun, S., Chae, S. H. and Yoon, S., “Application of Isolation Forest Technique for Outlier Detection in Water Quality Data,” *J. Korean Soc. Environ. Eng.*, **40**(12), 473-480(2018).
10. Guiñón, J. L., Ortega, E., García-Antón, J. and Pérez-herranz, V., “Moving Average and Savitzki-Golay Smoothing Filters Using Mathcad,” *International Conference on Engineering Education, July, Coimbra*, 1-4(2007).
11. Guo, H., Yu, M., Liu, J. and Ning, J., “Butterworth Low-pass Filter for Processing Inertial Navigation System Raw Data,” *J. Surv. Eng.*, **130**(4), 175-178(2004).

12. Panigrahi, S., Karali, Y. S. and Behera, H., "Time Series Forecasting Using Evolutionary Neural Network," *Int. J. Comput. Appl.*, **75**(10), 13-17(2013).
13. Zhang, Z., Wu, Z., Rincon, D. and Christofides, P. D., "Real-time Optimization and Control of Nonlinear Processes Using Machine Learning," *Mathematics*, **7**(10), 1-25(2019).
14. Oh, K., 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*., **23**(4), 565-572(2020).
15. Kwon, H., Oh, K., Chung, Y. G., Cho, H. and Kim, J., "Development of Machine Learning Model for Prediction Distillation Column Temperature," *Appl. Chem. Eng*., **31**(5), 520-525(2020).
16. Kwon, H., Oh, K., Choi, Y., CHung, Y. G, Kim, J., "Development and Application of Machine Learning-based Prediction Model for Distillation Column, *Int. J. Intell. Syst*., **36**, 1970-1997(2021).
17. Kim, T. J. and Hong, J. S., "Classification of Parent Company's Downward Business Clients Using Random Forest: Focused on Value Chain at the Industry of Automobile Parts," *J. Soc. E-bus. Stud*., **23**(1), 1-22(2018).
18. Kim, D. W., Lee, S. C., Kim, M. J., Lee, E. J. and Yoo, C. K., "Development of QSAR Model Based on the Key Molecular Descriptors Selection and Computational Toxicology for Prediction of Toxicity of PCBs," *Korean Chem. Eng. Res*., **54**(5), 621-629(2016).
19. Giap, V., Pineda, I. T., Lee, J. Y., Lee, D. K., Kim, Y. S., Ahn, K. Y. and Lee, Y. D., "Performance Prediction Model of Solid Oxide Fuel Cell Stack Using Deep Neural Network Technique, Trans," *Korean Hydrog. Energy Soc*., **31**(5), 436-443(2020).