Korean Chem. Eng. Res., 58(4), 588-595 (2020)
https://doi.org/10.9713/kcer.2020.58.4.588
PISSN 0304-128X, EISSN 2233-9558

인공지능 기반 질소산화물 배출량 예측을 위한 연구모형 개발

조하늬 · 박지수 · 윤용주†

포항공과대학교 화학공학과
37673 경상북도 포항시 남구 청암로 77
(2020년 4월 29일 접수, 2020년 5월 28일 수정본 접수, 2020년 6월 11일 채택)

Development of Prediction Model for Nitrogen Oxides Emission Using Artificial Intelligence

Ha-Nui Jo, Jisu Park and Yongju Yun†

Department of Chemical Engineering, Pohang University of Science and Technology,
77, Cheongam-ro, Nam-gu, Pohang-si, Gyeongsangbuk-do, 37673, Korea
(Received 29 April 2020; Received in revised from 28 May 2020; accepted 11 June 2020)

요 약

지속적으로 강화되는 환경오염 물질 배출 규제로 인해, 질소 산화물(NO_x)의 배출량 예측 및 관리는 산업 현장에서 많은 관심을 받고 있다. 본 연구에서는 인공지능 기반 질소산화물 배출량 예측모델 개발을 위한 연구모형을 제안하였다. 제안된 연구모형은 데이터의 전처리 과정부터 인공지능 모델의 학습 및 평가까지 모두 포함하고 있으며, 시계열 특성을 가지는 NO_x 배출량을 예측하기 위하여 순환 신경망 중 하나인 Long Short-Term Memory (LSTM) 모델을 활용하였다. 또한 의사결정나무 기법을 활용하여 LSTM의 time window를 모델 학습 이전에 선정하는 방법을 채택하였다. 본 연구에서 제안된 연구모형의 NO_x 배출량 예측 모델은 가열로에서 확보한 조업 데이터로 학습되었으며, 최적 모델은 hyper-parameter를 조절하여 개발되었다. 개발된 LSTM 모델은 학습 데이터 및 평가 데이터에 대하여 모두 93% 이상의 NO_x 배출량 예측 정확도를 나타내었다. 본 연구에 제안된 연구모형은 시계열 특성을 가지는 다양한 대기오염 물질의 배출량 예측모델 개발에 응용될 수 있을 것으로 기대된다.

Abstract – Prediction and control of nitrogen oxides (NO_x) emission is of great interest in industry due to stricter environmental regulations. Herein, we propose an artificial intelligence (AI)-based framework for prediction of NO_x emission. The framework includes pre-processing of data for training of neural networks and evaluation of the AI-based models. In this work, Long-Short-Term Memory (LSTM), one of the recurrent neural networks, was adopted to reflect the time series characteristics of NO_x emissions. A decision tree was used to determine a time window of LSTM prior to training of the network. The neural network was trained with operational data from a heating furnace. The optimal model was obtained by optimizing hyper-parameters. The LSTM model provided a reliable prediction of NO_x emission for both training and test data, showing an accuracy of 93% or more. The application of the proposed AI-based framework will provide new opportunities for predicting the emission of various air pollutants with time series characteristics.

Key words: NO_x, Artificial intelligence, Long short-term memory, Decision tree, Time series

1. 서 론

질소 산화물(NO_x)은 일산화질소(NO), 이산화질소(NO_2) 및 오산화이질소(N_2O_5)와 같은 기타 질소 산화물을 통칭하여 일컫는 용어로, 주로 자동차 혹은 산업 현장에서 연료를 연소하는 과정에서 발생하게 된다. NO_x는 생성 유형에 따라 고온에서 연소 공기 중 질소의 산화에 의해 생성된 Thermal NO_x, 질소 원자를 포함하는 연료의 연소 반응에 의해 생성된 Fuel NO_x, 탄화수소를 포함하는 연료의 연소 중 생성된 탄화수소 라디칼과 질소의 반응에 의해 생성된 Prompt NO_x로 나누어지며, 대부분의 NO_x는 Thermal NO_x에 의해 생성되고 있다[1]. NO_x는 공기 중으로 배출되었을 때 햇빛의 광화학 반응을 통해 미세먼지와 오존 등을 생성하므로 국내외적으로 환경 규제의 대상으로 지정되었으며, 유럽의 EURO-7과 한국의 대기환경보전법 및 사업장 대기오염 총량관리제 등을 통해 배출이 엄격히 규제되고 있다[2,3]. 특히, 한국은 최근 강화된 대기환경보전법에 따라, 2020년도부터

†To whom correspondence should be addressed.
E-mail: yjyun@postech.ac.kr
‡이 논문은 POSTECH 이인범 교수님의 정년을 기념하여 투고되었습니다.

NO_x 배출 사업장에 대해서 기준치를 초과하여 배출된 질소산화물에 대하여 kg당 최대 2130원의 배출 부과금[2]을 부과하고 있어 사업장 내의 NO_x 배출량에 대한 모니터링 강화가 요구되고 있다.

연소 장치 내부에서 생성되는 NO_x량을 계산하기 위해서는 연소실로 주입되는 연료 및 공기의 유체 거동과 NO_x 생성과 관련된 화학반응을 동시에 고려한 모델을 필요로 한다[4,5]. 그러나 실제 산업 현장에서는 조업 조건에 따라 실시간으로 연료 및 공기 투입량이 변화하기 때문에, NO_x의 시간적 변화뿐만 아니라 연소실 내부에서 공간적 구배를 동시에 고려해야 하므로 물리적 모델을 기반으로 한 NO_x 배출량 산출 모델은 많은 계산 시간을 소모하게 된다[6]. 또한 NO_x 배출량은 온도에 영향을 받기 때문에, 실시간으로 변화하는 연료 및 공기 투입량으로 인해 연소실 내부 임의의 지점에 발생되는 hot spot을 모델에 반영하기 어렵다는 단점도 존재한다[7]. 따라서 공기 중으로 배출되기 전 초과된 NO_x 배출량에 대해 즉각적인 대응을 필요로 하는 산업 현장에서는 위와 같은 물리적 모델을 기반으로 한 실시간 NO_x 배출량 모니터링은 한계를 지닌다. 이에 대한 대응으로 인공지능을 활용한 NO_x 배출량 예측에 대한 연구가 최근에 많은 관심을 받고 있다[8-10]. 이는 연소실 내부의 NO_x 생성에 따른 복잡한 패턴을 대량의 조업 데이터로부터 인공지능을 기법을 통해 학습시켜 모델을 개발하는 접근법으로, 최종 학습된 모델을 활용하여 새로운 조업 데이터에 대해 빠르게 NO_x 배출량을 산출할 수 있어 실시간 모니터링에 활용할 수 있다는 장점을 지닌다. 그러나 기존의 문헌에서는 인공지능 모델로 인공 신경망(Artificial Neural Network, ANN)을 주로 채택하고 있어, NO_x 배출에 따른 시간 의존성을 고려하지 않고 시간별 측정 데이터를 하나의 데이터 sample로 ANN의 학습에 사용하고 있다. 하지만, 실제 공정 운전에서 현재 시점의 조업이 NO_x 배출량에 영향을 미치기까지는 얼마간의 시간 지연이 존재하므로 현재 시점에서 측정된 NO_x 배출량은 과거 시점부터 현재까지의 조업에 영향을 받게 된다. 이와 같이 데이터는 통계 기법을 기반으로 한 고전 시계열 모델과 데이터 간의 시간 의존성을 고려한 순환 신경망(Recurrent Neural Network, RNN)을 통해 예측할 수 있다. 고전 시계열 모델에는 Autoregressive (AR), Moving Average (MA), Autoregressive Integrated Moving Average (ARIMA) 등이 있으며, 특히 ARIMA 모델은 예측의 정확성 측면에서 AR 및 MA와 같은 타 시계열 모델보다 뛰어난 성능을 보인다[11,12]. 그러나, 이러한 모델은 시계열 데이터의 선형 관계를 가정하고 있으므로 비선형성을 가지는 실제 산업 데이터에 적용 시 성능이 떨어지는 단점이 있다[13,14]. 반면에, RNN 모델의 경우 비선형성을 가지는 데이터에 대해 높은 예측 성능을 가지며, 기존의 시계열 모델에 비하여 경쟁력 있는 결과를 제공한다[15-17].

그러므로, 본 연구에서는 NO_x 배출량과 조업 데이터 간의 시간 의존성을 고려하기 위하여 순환 신경망(Recurrent Neural Network, RNN)을 활용한 NO_x 배출량 예측 연구모형을 개발하였다. NO_x 배출량 예측을 위해 RNN을 도입한 연구는 본 연구에서 처음 제안된 것으로, 제안된 연구모형은 공정 데이터의 전처리부터 RNN 모델 학습 및 평가까지 포괄한다. 먼저, 데이터 전처리 과정에서 조업 데이터를 표준화하고, 상관관계 분석을 통해 비슷한 변수 거동을 보이는 변수를 제거하여 실제 모델에 사용될 조업 변수를 선정하였다. 또한, 의사결정나무(Decision tree) 기법을 활용하여 조업 변수들의 시간에 따른 중요도를 파악하고 RNN의 입력 크기를 결정하여 RNN 구조 선정에 대한 타당성을 확보하였다. 학습된 RNN 모델은 실시간 NO_x 배출량 모니터링에 활용 가능하며, 위의 제안된 연구모형은 NO_x 이외의 시간 의존성을 고려한 대기 오염 물질 예측에 대하여 동일하게 적용 가능하다.

2. 연구 방법

2-1. 데이터 전처리

대량의 데이터를 활용하여 인공지능 모델을 학습하기에 앞서, 수집된 데이터의 전처리 과정이 필요하다. 데이터 전처리 과정은 이상치 제거 및 보완, 데이터 표준화, 변수들 간의 상관관계 분석을 통한 인공지능 모델의 입력 변수 선정 과정을 모두 포함한다.

먼저, 수집된 데이터의 분석을 통해 이상치 유무를 파악하고, 정상 거동을 벗어난 경우 제외한다. 데이터의 결측치는 아래의 식 (1)과 같이 선형 내삽법을 통해 결측치를 보완하여 데이터를 전처리한다[18].

$$y = y_1 + k(t + t_1) \tag{1}$$

여기서, k는 $k = (y_2 - y_1)/(t_2 - t_1)$이며, t_1, t_2는 선형 내삽이 필요한 시점의 처음과 끝을, y_1, y_2은 t_1, t_2 시점의 데이터 값을 의미한다.

다음으로, 데이터 표준화를 통해 모든 조업 변수들의 크기를 동일하게 조절한다. 실제 공정의 조업 데이터는 각각의 변수마다 측정 단위 및 조업 범위가 상이하므로 변수 값의 크기가 다르게 존재한다. 이러한 변수들을 그대로 인공지능 모델의 입력 변수로 활용을 하게 될 경우, 각 변수 별로 NO_x 배출량에 미치는 중요도를 제대로 반영을 할 수 없게 된다. 예를 들어, 10 bar에서 측정된 압력 변수의 1 단위 변화는 측정 변수의 10% 변화를 나타내지만, 1,000,000 Pa에서 측정된 압력 변수의 1 단위 변화는 측정 변수의 0.0001% 변화를 나타내므로 같은 1 단위만큼의 변화지만 변수 값의 크기 또는 단위에 따라 변화의 영향이 달라지게 된다. 따라서 모든 조업 변수들의 변화에 따른 영향도를 동일하게 조정하기 위해 데이터 표준화를 하였다. 이를 위한 방법에는 Min-Max normalization, Z-score normalization, Median normalization과 같은 방법들이 있으며[19], 본 연구에서는 데이터가[0, 1] 범위의 값을 가지도록 Min-Max normalization을 통해 표준화하였고 그 식은 아래와 같다.

$$a_i = \frac{y_i - \min(y)}{\max(y) - \min(y)} \tag{2}$$

마지막으로, 표준화된 데이터를 바탕으로 상관관계를 분석하였다. 상관관계는 다변수 데이터에서 변수들 간의 선형 종속성을 -1에서 1 사이의 값으로 수치화한 것으로, 아래와 같이 정의된다[20].

$$\rho(\mathbf{x},\mathbf{y}) = \frac{Cov(\mathbf{x},\mathbf{y})}{\sigma_x \sigma_y} = \frac{E[(\mathbf{x}-\mu_x)(\mathbf{y}-\mu_y)]}{\sigma_x \sigma_y} \tag{3}$$

여기서, Cov(**x**, **y**)는 변수 **x**와 **y**의 공분산을, σ는 분산을 나타낸다. 식 (3)을 통해 계산된 상관관계 ρ는 그 값의 범위에 따라 $0<\rho\leq1$의 경우 양의 상관관계, $-1\leq\rho<0$의 경우 음의 상관관계, $\rho=0$의 경우 상관관계가 없는 것으로 분석하며, 0.7 이상의 절대값을 가질 경우 두 변수 간에 강한 상관관계가 있는 것으로 판단한다. Fig. 1에서 상관관계 값에 따른 변수의 관계를 확인할 수 있는데, 상관관계가 높지 않은 변수들은 독립적으로 움직이는 반면, 상관관계가 높은 변수들은 움직임이 동일한 것을 알 수 있다. 이러한 변수를 중복하여 인공

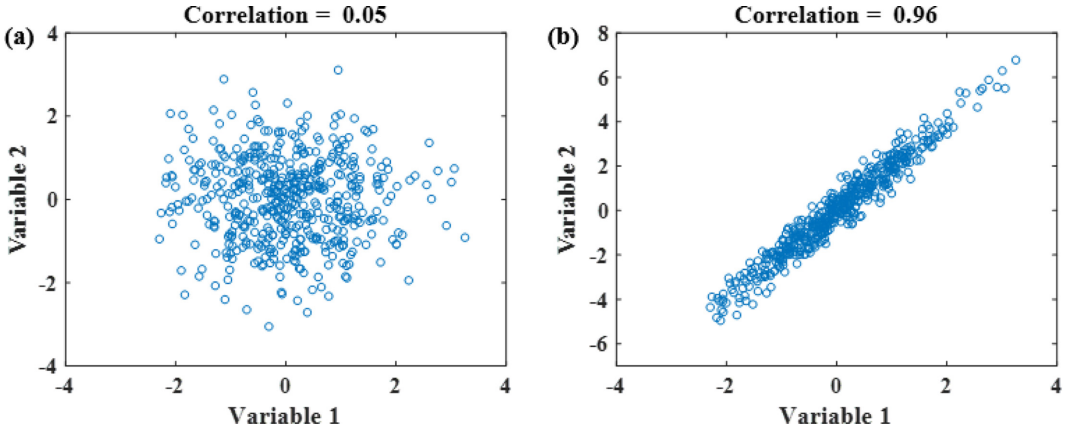

Fig. 1. Example of correlation analysis. (a) Low correlation case, (b) High correlation case.

지능의 모델의 입력 변수로 사용할 경우, 학습된 모델이 중복되는 거동을 가지는 변수들에 치우치게 되며 모델의 성능이 저하된다 [14,15]. 따라서, 상관관계를 바탕으로 변수의 움직임이 중복되는 변수를 제외함으로써, 인공지능 모델에 입력 변수를 축소시키는 동시에 모델 강건성을 확보하였다.

2-2. 순환 신경망을 이용한 인공지능 모델 개발

가장 일반적으로 사용되는 인공지능 모델 중 하나인 Feed-Forward Neural Network (FFNN)은 고형 폐기물의 발열량[23], 엔진의 그을음, CO_2 및 NO_x 발생량[24], 평균 강우량[25] 등의 예측 모델 연구에 적용되어 왔다. FFNN의 기본적인 구조는 p개의 입력 노드를 가지는 입력층, m개의 노드를 가지는 은닉층, q개의 출력 노드를 가지는 출력층으로 구성되며, 이는 Fig. 2(a)에서 확인할 수 있다. 이러한 FFNN은 구조상 연속적으로 측정된 데이터 간의 영향도를 고려할 수 없다는 한계점이 존재하는데, 이를 개선한 방법이 RNN이다. RNN의 구조는 s개의 연속적으로 측정된 데이터 각각에 대하여 p개의 입력 노드로 구성된 입력층, m 개의 노드로 구성된 은닉층, q개의 출력 노드로 구성된 출력층으로 구성되며, 이는 Fig. 2(b)에서 확인할 수 있다. 이러한 RNN의 구조는 연속적으로 측정된 데이터가 순차적으로 입력되고, 입력된 정보가 다음 시간 데이터에 정보를 넘겨주기 때문에 시간에 따른 데이터의 종속성을 모델에 함께 반영할 수 있게 된다[26]. 그러나 일반적인 RNN의 경우 가중치 소실 문제로 인해 장기간 데이터를 보존할 수 없다는 단점이 있어, 이를 보완하기 위한 방법으로 Long Short-Term Memory (LSTM)을 기반으로 한 RNN 기법이 제안되었다[27].

LSTM은 망각, 입력, 출력 게이트로 구성되어 있으며, 세부 구조는

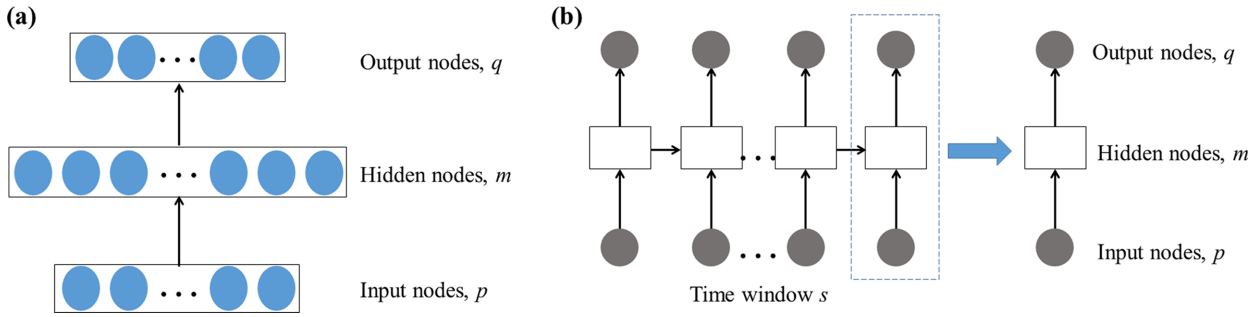

Fig. 2. (a) Structure of Feed-Forward Neural Network (FFNN), (b) Structure of Recurrent Neural Network (RNN).

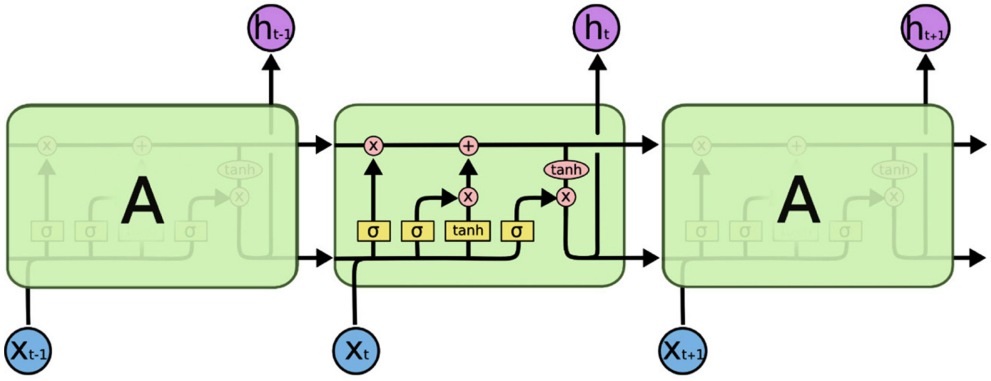

Fig. 3. Structure of long short-term memory (LSTM) network [28].

Fig. 3에 나타나 있다. 현재 시점(t)에서 LSTM 셀의 각 게이트는 이전 시점(t-1) LSTM 셀의 은닉 상태(h_{t-1})와 현재 입력 데이터(x_t)를 받게 되며 학습을 통해 각 게이트의 가중치와 편향을 결정하게 된다. 먼저, 망각 게이트(f_t)에서는 이전 시점의 기억(C_{t-1})으로부터 제거해야 할 정보를 삭제한다.

$$f_t = \sigma(x_t W_{x,f} + h_{t-1} W_{h,f} + b_f) \quad (4)$$

여기서, σ는 활성화 함수, $W_{x,f}$는 망각 게이트의 x_t에 대한 가중치, $W_{h,f}$는 망각 게이트의 h_{t-1}에 대한 가중치, b_f는 망각 게이트의 편향을 나타낸다.

두 번째로 입력 게이트(i_t)에서는 과거 시점의 기억(C_{t-1})에 추가하고자 하는 현재의 정보를 결정한다. 식 (5)을 통해 x_t와 h_{t-1}로부터 C_{t-1}에 추가할 정보를 생성하고, 입력 게이트에서 식 (6)과 같이 현재 시점에서 추가하고자 하는 정보의 가치를 결정한 뒤, 식 (7)을 통해 최종적으로 현재 시점의 기억(C_t)을 생성한다.

$$\tilde{C}_t = \tanh(x_t W_{x,g} + h_{t-1} W_{h,g} + b_g) \quad (5)$$

$$i_t = \sigma(x_t W_{x,i} + h_{t-1} W_{h,i} + b_i) \quad (6)$$

$$C_t = f_t \bullet C_{t-1} + i_t \bullet \tilde{C}_t \quad (7)$$

여기서, $W_{x,i}$는 입력 게이트의 x_t에 대한 가중치, $W_{h,i}$는 입력 게이트의 h_{t-1}에 대한 각각의 가중치를, b_i는 입력 게이트의 편향을 나타낸다.

마지막으로 출력 게이트(o_t)는 식 (8)-(9)를 통해 다음 시점(t+1)에 전달할 은닉 상태(h_t)를 결정한다.

$$o_t = \sigma(x_t W_{x,o} + h_{t-1} W_{h,o} + b_o) \quad (8)$$

$$h_t = o_t \bullet \tanh(C_t) \quad (9)$$

여기서, $W_{x,o}$는 출력 게이트의 x_t에 대한 가중치, $W_{h,o}$는 출력 게이트의 h_{t-1}에 대한 각각의 가중치를, b_o는 출력 게이트의 편향을 나타낸다.

LSTM은 위의 설명된 구조를 기본 cell로 하여 연속적으로 측정된 데이터에 대해 처리 후 순차적으로 다음 LSTM cell로 정보를 넘겨주기 때문에 데이터 간의 시간 종속성을 고려한 모델을 학습할 수 있다.

2-3. 의사결정나무 모델을 이용한 인공지능 모델 구조 확정

시계열 데이터에 대하여 LSTM을 기반으로 인공지능 모델을 구축하기 위해서는 한번에 처리할 시간 데이터의 범위를 결정 해야한다. LSTM 모델의 입력 시간을 look-back 혹은 time window라고 하며, 다수의 연구에서 LSTM의 time window를 특정 범위에서 변경하면서 그에 따른 모델을 학습한 뒤, 학습 에러를 분석하여 최소의 에러 값을 가지는 time window를 최적의 입력 구조로 선정한다[22,23]. 이러한 접근법은 time window 크기의 변화에 따라 개별 모델을 학습하고 비교를 통해 최종 모델을 취사선택할 수 있지만, 각 time window에 따라서 LSTM의 최적 hyper-parameter도 변하기 때문에 time window와 hyper-parameter를 동시에 고려한 최적의 모델 선정을 위해서 많은 시간을 소모하게 된다. 그러므로, 본 연구에서는 의사결정나무 모델을 이용하여 시간별 변수 중요도를 분석하여 time window를 모델 학습 이전에 결정하였다.

의사결정나무는 의사결정규칙을 나무 구조로 도식화하여 분류를 수행하는 머신러닝 기법 중 하나이다[31]. 의사결정나무의 구조는 Fig. 4(a)와 같이 뿌리 노드, 중간 노드, 말단 노드로 구성되며, 뿌리 노드에서 의사결정규칙에 의해 최초로 데이터가 분기되며, 분기된 데이터는 뿌리 노드와 말단 노드 사이인 중간 노드에서 의사결정규칙에 따라 추가적으로 분기가 일어나고, 최종적으로 말단 노드에서 데이터가 분류된다. 의사결정나무는 어떤 변수를 이용하는 것이 가장 데이터를 잘 분류하는지 파악하여 각 노드의 의사결정규칙을 학습하게 된다. 이때 각 노드의 분류 기준으로 Gini index 또는 Entropy를 사용할 수 있으며 이를 최소화하는 방향으로 학습된다.

의사결정나무 기법을 활용하여 LSTM의 time window를 결정하기 위해서 Fig. 4(b)와 같이 데이터를 확장한 뒤 목표값(Y)를 분류한다. 이는 분석하고자 하는 시간 범위(k)까지 데이터를 1분 간격으로 확장하여 각각의 의사결정나무 모델을 만드는 과정을 필요로 한다. 예를 들어, 1분 전 조업 데이터(T1)를 활용하여 현재 목표값(Y)을 분류하는 의사결정나무 모델을 만들고, 1-2분 전 조업 데이터(T1, T2)를 활용하여 Y를 분류하는 의사결정나무 모델을 만든다. 여기서, T1와 T2에서 측정된 변수는 동일할 지라도 측정된 시점이 다르므로 새로운 변수로 고려하여 Y값을 분류하게 된다. 이후, 총 k개의 의사결정나무 모델에 대하여 변수 중요도를 분석하여 어느 시점(s)에 측정된 변수들이 Y에 큰 영향도를 가지는지 파악하고, 이때 확인된 시점 s를 최종 time window로 선정하게 된다. 이렇게 time window를 사전에 결정함으로써, Y값에 유의미한 영향을 가지는 데이터만 LSTM의 모델에 활용할 수 있도록 하여 과도하게 모델의 크기가 커지는 것을 방지할 수 있다.

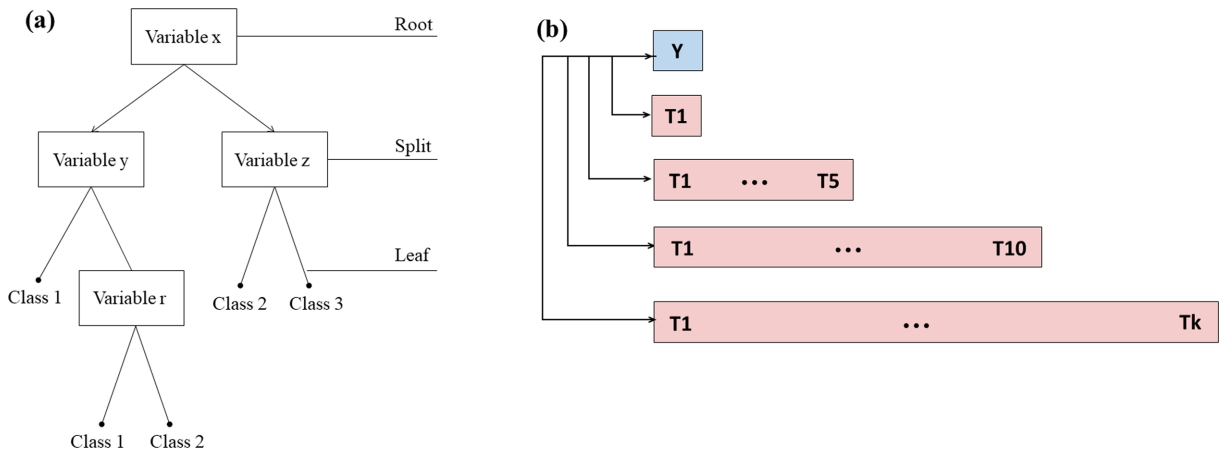

Fig. 4. (a) Structure of Decision tree, (b) Extension of time series data.

3. 결과 및 고찰

본 연구는 다량의 조업 데이터로부터 NO_x의 배출량을 예측하는 인공지능 모델을 개발하기 위한 연구모형을 제시하였고, 이를 산업 현장에서 확보한 조업 데이터에 적용하여 제안된 연구모형의 타당성을 확인하였다. 조업 데이터는 가열로에서 측정된 것으로, 개략적인 구조는 Fig. 5에 표현되어 있다. 가열로에 장입된 물질은 예열대(preheating zone), 가열대(heating zone), 균열대(soaking zone)를 거쳐 이동하게 되는데, 버너에 공급된 연소 가스와 공기의 고온 연소 반응을 통해 발생되는 열 에너지를 활용해 가열이 이루어진다. 이때 연소 반응을 위해 공급된 공기 중의 질소가 고온에서 분해되면서 산소와 반응을 통해 환경 오염 물질인 NO_x를 배출하게 된다. 본 연구를 위해 확보한 조업 변수는 가열로의 상, 하부에서 측정된 연소 가스 및 공기의 유량, 온도, 압력 등이 있으며, 이는 Table 1에 정리되어 있다. 연구 모형의 타당성 검증을 위해 활용한 조업 데이터는 Table 1의 조업 변수 48개와 NO_x 측정치로 구성되며, 1분 간격으로 측정된 총 4,407개의 데이터를 활용하였다.

먼저, 데이터 전처리 과정에서 이상치 및 결측치의 유무를 파악하였다. 측정된 4,407개의 데이터 중 5개의 NO_x 측정치에 대한 결측치가 확인하였고, 선형 내삽법을 통해 보완하였다. 또한, 데이터를 표준화하여 모든 변수들의 단위를 제거하였다. 이후 표준화된 데이터의 상관관계 분석을 통해 모델의 입력 변수를 선정하였다. 실제 가열로 조업 시 버너로 공급되는 공기 유량은 연소 가스 유량에 비례하여 결정되며, 상, 하부의 온도 또한 동일한 수준으로 제어되고 있어 유사한 변수 거동을 보인다. 이러한 변수들의 상관관계 분석을 통하여 절대값 0.9 이상의 상관관계 값을 가지는 것을 확인하였고, 이를 최종 변수 선정의 기준 값으로 설정하였다. 따라서, 상관관계가 절대값 0.9 이상인 각각의 변수 쌍에 대하여 하나의 변수를 제거하여, 초기 48개의 변수 중 24개의 변수를 인공지능 모델의 입력 변수로 최종 활용하였다.

전처리 된 데이터를 바탕으로 LSTM을 활용하여 NO_x 배출량 예측 모델을 개발하였으며, 의사결정나무 기법을 활용하여 LSTM의 time window의 크기를 선정하였다. 이를 위해, Fig. 4(b)와 같이 1분 전부터 15분 전까지의 데이터를 1분 단위로 확장하여 총 15개의 case를 만들었다. 이때, 각 case에서 확장된 변수는 측정 변수는 동일하더라도 다른 시간대에서 측정된 변수이므로 새로운 변수 번호를 부여하였다. 따라서, 1번 case에서는 1분 전 데이터 총 24개, 2번 case에서는 1~2분 전 데이터 총 48개, 15번 case에서는 1~15분 전 데이터 총 360개의 변수를 가지게 되며, 각 case별 의사결정나무를 만들고 그에 따른 변수의 중요도를 확인하였다. Fig. 6는 각 case별 상위 4개의 변수 중요도 결과를 보여주며, 고려하는 time window에 따른 변수 중요도는 5분 이후로는 크게 변화하지 않는 것을 확인할 수 있다. 이러한 결과를 바탕으로 본 연구에서는 계산에 소요되는 시간을 최소화하는 동시에 높은 예측 정확도를 유지하기 위해 LSTM의 time window를 5개로 결정하였다.

전처리 과정을 거친 데이터는 인공지능 모델의 입력에 적합한 형태로 변환하였다. 의사결정나무를 활용하여 LSTM time window의

Table 1. Process variables of heating furnace

Variables	Description of variables
1 - 2	Air flowrates in the top and bottom of the preheating zone
3	Temperature in the preheating zone
4 - 13	Air flowrates in the top and bottom of the heating and soaking zones
14 - 23	Fuel gas flowrates in the top and bottom of the heating and soaking zones
24 - 33	Temperatures in the top and bottom of the heating and soaking zones
34	Air temperature
35	Total air flowrate
36	Preheater pressure
37	O_2 concentration in the preheating zone
38 - 39	O_2 concentrations in the heating and soaking zones
40	Air-fuel ratio
41 - 47	Fuel gas composition
48	Theoretical NO_X emission

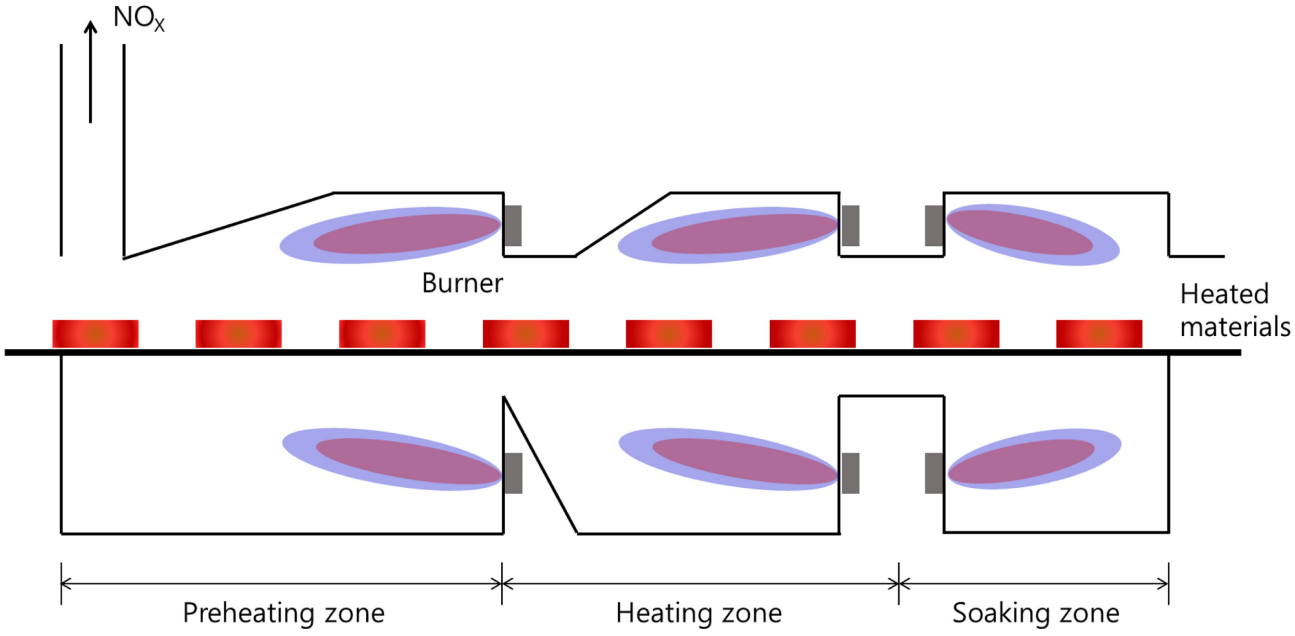

Fig. 5. Layout of heating furnace.

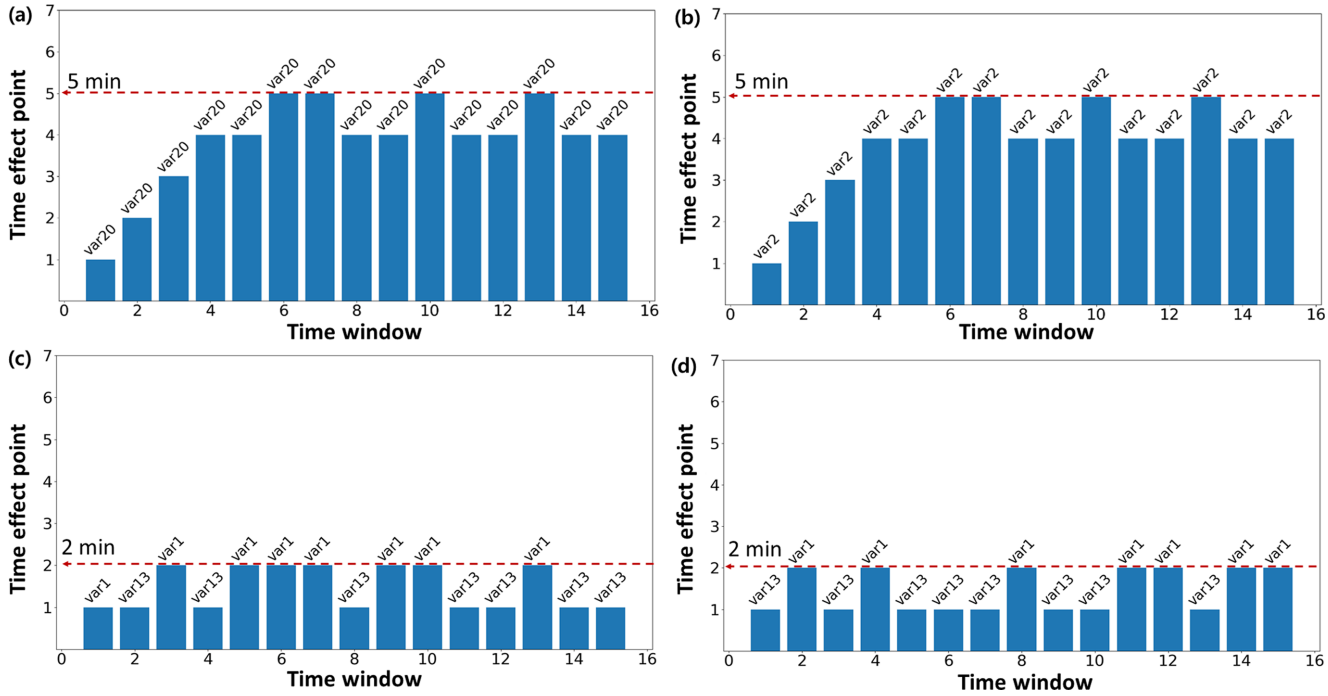

Fig. 6. Feature importance as a function of time window. (a) Top 1 variables, (b) Top 2 variables, (c) Top 3 variables, (d) Top 4 variables.

크기를 5로 결정하고, 상관관계 분석을 통해 24개의 변수를 입력 변수로 선정하였으므로, LSTM의 각 sample은 5개의 입력 데이터에 대하여 24개의 변수로 구성되었다. 따라서 총 4,407개의 연속 측정된 데이터에 대하여 크기 5인 time window를 1분 단위로 옮겨 4,397개의 sample을 생성하였다. 이렇게 생성된 데이터의 80%는 모델의 학습용 데이터로 활용하였으며, 나머지 20%는 모델의 평가용 데이터로 활용하였다.

NO_x 예측 모델은 1개의 LSTM layer로 구성하였으며, 은닉층의 크기 6, 12, 16, 20, 24, 48개의 모델에 대해 학습 속도 0.01로 반복 학습하였다. 모델의 평가를 위해 식 (10)의 Root mean square error (RMSE)를 사용하였고, 그 식은 아래와 같다.

$$RMSE = \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n} \tag{10}$$

여기서, y_i는 i번 sample의 실제 값, $\hat{y}_i$는 i번 sample의 예측값, n은 sample 개수이다. 학습 시행 횟수 100번, 300번, 500번, 1000번에 대한 학습 및 평가 RMSE 결과는 Table 2과 같다. 이때, 과도한 반복 시행은 학습 데이터에 대해서는 오차를 감소시키지만, 평가 데이터에 대해서는 오차를 증가시키는 과적합 문제를 발생시킨다. 따라서, 과적합이 발생하지 않는 지점의 모델을 얻기 위해 평가 데이터의 RMSE가 최소가 되는 최적 학습 횟수를 확인하였고, 이를 Table 2에서 함께 정리하였다. 또한 은닉층의 크기 변화에 따른 학습 및 평가 결과를 Table 2에서 확인 할 수 있는데, 평가 데이터의 최소 RMSE 값은 은닉층의 크기가 6에서 24로 변함에 따라 점차 증가하다가 감소하며, 은닉층의 크기가 48인 경우와 은닉층의 크기가 24인 경우의 RMSE 결과가 동일한 것을 알 수 있다. 은닉층의 크기가 클 경우 모델 학습 시 찾아야 하는 최적 weight 및 bias의 개수가 증가하며, 추후 모델의 활용 시에도 계산 비용을 증가 시키므로, 본 연구에서는 은닉층의 크기를 24로 선정하였으며 학습 횟수 705번째에서 학습

Table 2. Model performance result for NO_x prediction

Case #	Hidden Dimension	Iteration	Train RMSE	Test RMSE
1	6	100	0.202	0.163
2		300	0.115	0.187
3		500	0.090	0.159
4		1000	0.047	0.086
Optimal iteration		361	0.034	0.034
5	12	100	0.151	0.173
6		300	0.089	0.142
7		500	0.068	0.123
8		1000	0.028	0.077
Optimal iteration		751	0.043	0.043
9	16	100	0.060	0.105
10		300	0.027	0.070
11		500	0.022	0.075
12		1000	0.019	0.082
Optimal iteration		211	0.034	0.063
13	20	100	0.058	0.104
14		300	0.026	0.090
15		500	0.023	0.087
16		1000	0.019	0.088
Optimal iteration		718	0.046	0.068
17	24	100	0.069	0.106
18		300	0.030	0.062
19		500	0.023	0.065
20		1000	0.017	0.067
Optimal iteration		705	0.032	0.032
21	48	100	0.059	0.124
22		300	0.027	0.077
23		500	0.020	0.072
24		1000	0.017	0.062
Optimal iteration		755	0.032	0.032

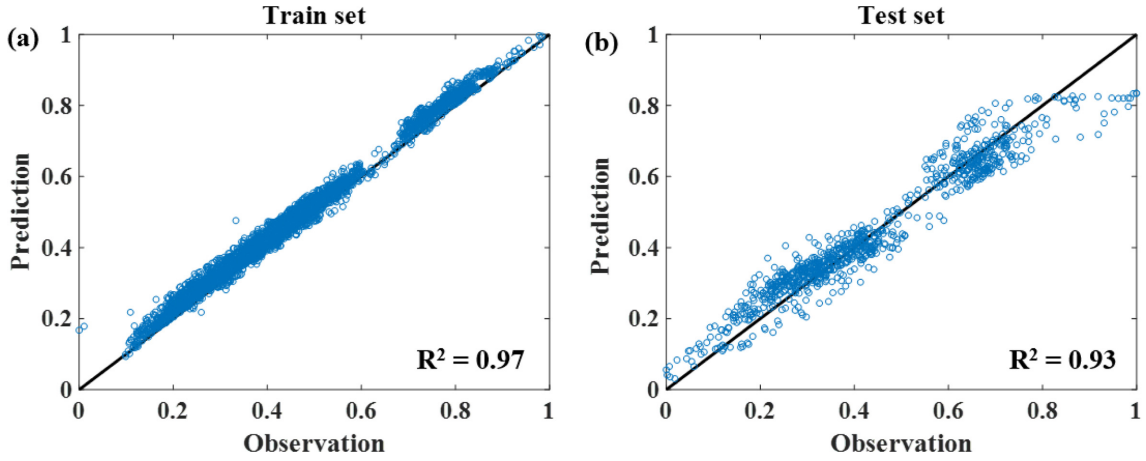

Fig. 7. Parity plot of train and test sets for the developed NO_x prediction model.

된 모델이 평가 데이터에 대하여 최소의 RMSE를 가지므로 이를 NO_x 배출량 예측의 최종 모델로 선정하였다. 선정된 NO_x 예측 모델의 정확도는 Fig. 7의 Parity plot을 통해 확인할 수 있다. Fig. 7(a)는 NO_x 배출량 예측 모델을 학습 데이터로 학습한 결과로 R^2 값 0.97의 정확도를 가지는 학습 모델을 확보하였음을 나타낸다. Fig. 7(b)는 학습된 NO_x 배출량 예측 모델을 평가 데이터로 평가한 것으로 R^2 값 0.93의 정확도를 가진다. 이러한 결과로부터, 학습 데이터뿐만 아니라 평가 데이터에 대해서도 높은 정확도를 가지는 NO_x 예측 모델이 개발되었음을 확인하였다.

4. 결 론

이번 연구에서는 인공지능 기법인 LSTM을 이용하여 NO_x 배출량 예측을 위한 모델 개발 연구모형을 제시하였고, 이를 조업 데이터에 적용하여 타당성을 검증하였다. 먼저, 데이터 전처리 과정에서 이상치를 제거하고 결측치를 선형 내삽법을 통해 보안하였으며, 상관관계 분석을 통해 비슷한 변수 거동을 가지는 변수를 제외하여 변수의 중복 사용을 제한하였다. 그 결과, 전체 48개의 변수 중 24개로 변수를 축소할 수 있었다. 또한, 의사결정나무 기법을 활용하여 LSTM의 time window를 선정하였다. 이를 위해 1분 ~ 15분 전 데이터까지 1분 간격으로 확장하여 개별 의사결정나무 모델을 만들고, 각 모델의 변수 중요도를 분석하여 최종적으로 time window 크기를 5로 선정하였다. 이후 학습 데이터를 활용하여 LSTM 기반 NO_x 예측 모델을 학습하였고, 평가 데이터를 활용하여 학습된 NO_x 예측 모델을 평가하였다. 그 결과 학습 데이터와 평가 데이터에서 93%이상의 예측 정확도를 확보할 수 있었고, 제안된 연구모형의 타당성을 확보하였다. 이 연구결과를 바탕으로, 해당 연구모형은 추후 NO_x 뿐만 아니라 시계열 특성을 가진 다른 대기오염 물질의 인공지능 기반 배출량 예측모델에도 응용될 수 있을 것으로 판단된다.

References

1. Studzinski, W., Liiva, P., Choate, P. and Acker, W., "A Computational and Experimental Study of Combustion Chamber Deposit Effects on NOx Emissions," *SAE Tech. Pap.*, 932815(1993).
2. Lee, H., "Status and Improvement Measures for Total Air Pollution Load Management," *National Assembly Research Service*, 36(2019).
3. Karim, Z. A. A., Khan, M. Y., Rashid, A., Aziz, A. and Hagos, F. Y., "Attaining Simultaneous Reduction in Nox and Smoke by Using Water-in-biodiesel Emulsion Fuels For Diesel Engine," *Platform : A Journal of Engineering*, **3**, 1-21(2019).
4. Xu, M., Azevedo, J. L. T. and Carvalho, M. G., "Modelling of the Combustion Process and NO_x Emission in a Utility Boiler," *Fuel*, **79**, 1611-1619(2000).
5. Khoshhal, A., Rahimi, M. and Alsairafi, A. A., "CFD Study on Influence of Fuel Temperature on NO_x Emission in a HiTAC Furnace," *Int. Commun. Heat Mass Transfer*, **38**, 1421-1427(2011).
6. Ferretti, G. and Piroddi, L., "Estimation of NO_x Emissions in Thermal Power Plants Using Neural Networks," *J. Eng. Gas Turbines Power*, **123**, 465-471(2001).
7. Reinbacher, F. and Regele, J. D., "Influence of Smooth Temperature Variation on Hotspot Ignition," *Combust. Theor. Model.*, **22**, 110-130(2018).
8. Joo, S., Yoon, J., Kim, J., Lee, M. and Yoon, Y., "NO_x Emissions Characteristics of the Partially Premixed Combustion of H_2/CO/CH_4 Syngas Using Artificial Neural Networks," *Appl. Therm. Eng.*, **80**, 436-444(2015).
9. Park, C. and Kim, Y., "A Study on NO_x Emission Control Methods in the Cement Firing Process Using Data Mining Techniques," *J. Korean Soc. Qual. Manag.*, **46**, 739-752(2018).
10. Taghavifar, H., Taghavifar, H., Mardani, A. and Mohebbi, A., "Exhaust Emissions Prognostication for DI Diesel Group-hole Injectors Using a Supervised Artificial Neural Network Approach," *Fuel*, **125**, 81-89(2014).
11. Tabachnick, J., Fidell, B. G. and Ullman, L. S., *Using Multivariate Statistics*, Pearson Education, London (2007).
12. Meyler, A., Kenny, G. and Quinn, T., "Forecasting Irish Inflation Using ARIMA Models," *Munich Personal RePEc Archive*, **11359** (1998).
13. Khashei, M., Bijari, M. and Raissi Ardali, G. A., "Improvement of Auto-Regressive Integrated Moving Average models using Fuzzy logic and Artificial Neural Networks (ANNs)," *Neurocomputing*, **72**(4-6), 956-967(2009).
14. Zhang, G., Patuwo, B. E. and Hu, M. Y., "Forecasting with Artificial Neural Networks-the State of Art," *Int. J. Forecast.*, **14**(1), 35-62(1998).

15. Song, X., Liu, Y., Xue, L., Wang, J., Zhang, J., Wang, J., Jiang, L. and Cheng, Z., "Time-series Well Performance Prediction Based on Long Short-Term Memory (LSTM) Neural Network Model," *J. Petrol. Sci. Eng*., **186**, 106682(2020).
16. Li, Y. and Cao, H., "Prediction for Tourism Flow based on LSTM Neural Network," *Procedia Comput. Sci*., **129**, 277-283(2018).
17. Karakoyun, E. Ş. and Çıbıkdiken, A. O., "Comparison of ARIMA Time Series Model and LSTM Deep Learning Algorithm for Bitcoin Price Forecasting," *Proceedings of the Multidisciplinary Academic Conference*, 171-179(2018).
18. Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M., "Methods for Imputation of Missing Values in Air Quality Data Sets," *Atmos. Environ.*, **38**, 2895-2907(2004).
19. Sola, J. and Sevilla, J., "Importance of Input Data Normalization for the Application of Neural Networks to Complex Industrial Problems," *IEEE Trans. Nucl. Sci.*, **44**, 1464-1468(1997).
20. Benesty, J., Chen, J., Huang, Y. and Cohen, I., *Noise Reduction in Speech Processing*, Springer-Verlag Berlin Heidelberg(2009).
21. Kuhn, M., "Building Predictive Models in R Using the Caret Package," *J. Stat. Softw.*, **28**, 1-26(2008).
22. Kwon, S. H., Lee, J. and Chung, G., "Snow Damages Estimation using Artificial Neural Network and Multiple Regression Analysis," *J. Korean Soc. Hazard Mitig.*, **17**, 315-325(2017).
23. Dong, C., Jin, B. and Li, D., "Predicting the Heating Value of MSW with a Feed Forward Neural Network," *Waste Manage.*, **23**, 103-106(2003).
24. Taghavifar, H., Taghavifar, H., Mardani, A., Mohebbi, A., Khalilarya, S. and Jafarmadar, S., "Appraisal of Artificial Neural Networks to the Emission Analysis and Prediction of CO_2, soot, and NO_x of n-heptane Fueled Engine," *J. Cleaner Prod.*, **112**, 1729-1739(2016).
25. Chattopadhyay, S., "Feed forward Artificial Neural Network model to predict the average summer-monsoon rainfall in India," *Acta Geophys.*, **55**, 369-382(2007).
26. Connor, J. T., Martin, R. D. and Atlas, L. E., "Recurrent Neural Networks and Robust Time Series Prediction," *IEEE Trans. Neural Networks*, **5**, 240-254(1994).
27. Hochreiter, S., "Long Short-Term Memory," *Neural Comput.*, **1780**, 1735-1780(1997).
28. Olah, C., "Understanding LSTM Networks," GITHUB blog (2015), Retrieved from http://colah.github.io/posts/2015-08-Understanding-LSTMs/.
29. Freeman, B. S., Taylor, G., Gharabaghi, B. and Thé, J., "Forecasting Air Quality Time Series Using Deep Learning," *J. Air Waste Manage. Assoc.*, **68**, 866-886(2018).
30. Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K. and Soman, K. P., "Stock Price Prediction Using Lstm, Rnn and CNN-sliding Window Model," *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1643-1647(2017).
31. Safavian, S. R. and Landgrebe, D., "A Survey of Decision Tree Classifier Methodology," *IEEE Trans. Syst. Man. Cybern.*, **21**, 660-674(1991).