JANT 한국항행학회논문지
Journal of Advanced Navigation Technology

응용 및 융합 기술

J. Adv. Navig. Technol. 25(6): 576-581, Dec. 2021

PM10 예측 성능 향상을 위한 농도별 예측 모델 설계

Prediction Model Design by Concentration Type for Improving PM10 Prediction Performance

조 경 우[1] · 정 용 진[2] · 오 창 헌[2*]
[1]한국정보통신기술협회 AI시험검증팀
[2]한국기술교육대학교 전기전자통신공학과

Kyoung-Woo Cho[1] · Yong-jin Jung[2] · Chang-Heon Oh[2*]

[1]AI Testing Team, Telecommunications Technology Association, Seongnam 13591, Korea
[2]Department of Electrical, Electronics and Communication Engineering, Korea University of Technology and Education(KOREATECH), Cheonan 31253, Korea

[요 약]

고농도의 경우 저농도와 비교하였을 때, 발생 빈도수의 차이와 발생 환경에 대한 차이로 예측 성능의 한계를 두드러지게 보이고 있다. 이러한 문제를 해결하기 위해 본 논문에서는 인공신경망 알고리즘을 이용하여 저농도와 고농도로 분류하고 구분된 농도별로 특성을 학습시킨 두 가지 예측 모델을 통해 예측을 수행하는 모델을 제안하였다. 저농도와 고농도를 분류하기 위해 DNN 기반의 분류 모델을 설계하고 분류모델을 통해 구분된 저농도와 고농도를 기준으로 농도별 특성을 반영하기 위한 저농도 예측 모델과 고농도 예측 모델을 설계하였다. 농도별 예측 모델의 성능 평가 결과, 저농도 예측 정확도가 90.38%, 고농도 예측 정확도는 96.37%의 예측 정확도를 보였다.

[Abstract]

Compared to a low concentration, a high concentration clearly entails limitations in terms of predictive performance owing to differences in its frequency and environment of occurrence. To resolve this problem, in this study, an artificial intelligence neural network algorithm was used to classify low and high concentrations; furthermore, two prediction models trained using the characteristics of the classified concentration types were used for prediction. To this end, we constructed training datasets using weather and air pollutant data collected over a decade in the Cheonan region. We designed a DNN-based classification model to classify low and high concentrations; further, we designed low- and high-concentration prediction models to reflect characteristics by concentration type based on the low and high concentrations classified through the classification model. According to the results of the performance assessment of the prediction model by concentration type, the low- and high-concentration prediction accuracies were 90.38% and 96.37%, respectively.

Key word : Artificial neural network, Classification, Deep learning, Deep neural network, Particulate matter.

https://doi.org/10.12673/jant.2021.25.6.576

Received 12 November 2021; **Revised** 1 December 2021
Accepted (Publication) 23 December 2021 (30 December 2021)

***Corresponding Author, Chang-Heon Oh**

Tel: +82-41-560-1187
E-mail:choh@koreatech.ac.kr

Ⅰ. 서 론

석탄, 석유 등 화석 연료 기반의 산업과 조리 과정, 자동차 등 일반 생활환경과 같이 다양한 환경에서 미세먼지가 발생되고 있다. 눈에 보이지 않을 정도의 미세한 크기로 인해 건강에 유해한 영향을 주며, 이에 따라 사회 활력 저하 등 다양한 사회적 문제의 원인으로 분석되고 있다[1-5]. 미세먼지에 대해 다양한 방법을 이용하여 대중들에게 미세먼지와 관련된 정보들을 제공하고 있으며, 많은 사람들은 미세먼지의 영향을 최대한 줄이기 위해 관련 정보를 요구하고 있다. 그러나 60% ~ 80%의 미세먼지 예보 정확도는 사람들의 요구를 충족시키지 못하고 있다[6]. 따라서 기존의 예보 시스템 외의 머신러닝, 딥러닝 등 다양한 방법을 이용하여 정확도를 높이기 위한 연구들이 활발히 진행되고 있다. 하지만 미세먼지 농도에 따라 불규칙한 발생 비율과 다양한 특성으로 인해 예측 모델의 학습에서 미세먼지 농도의 특성을 적절히 반영하지 못하는 문제가 있다.

K. H. Jeon의 연구에서는 미세먼지에 대한 예보 정확도가 88% 수준까지 향상되었으나, $80\mu g/m^3$ 이상의 고농도 예보 정확도의 경우 60% 수준으로 고농도 미세먼지에 대한 예보 정확도 개선이 필요하다고 보고하고 있다[7]. H. L. Kim의 연구에서는 미세먼지의 발생 요인 탐색 및 미세먼지 발생 예측 모델 설계를 제안하였다. 여러 알고리즘을 적용하여 예측 모델의 성능 평가 결과, 미세먼지 예측 모델의 성능은 훈련용 데이터의 비율과 비례하지 않았으며, 미세먼지의 실제 값이 높을 때, 예측 값이 떨어지는 것을 확인하였다[8]. 이러한 연구들을 바탕으로 고농도의 미세먼지의 경우, 예측 성능의 어려움이 있으며, 고농도에 대한 예측 성능 향상 시 전반적인 미세먼지 농도 예측의 성능 향상을 기대할 수 있다.

따라서 본 논문에서는 미세먼지의 발생 환경에 따라 구분될 수 있는 저농도와 고농도의 분류를 통해 각 특성을 개별적으로 반영한 농도별 예측 모델을 제안한다. PM_{10}값의 $80\mu g/m^3$를 기준으로 저농도와 고농도의 분류를 위해 인공신경망 알고리즘 중 DNN(deep neural network) 알고리즘을 이용하여 농도별 분류 모델의 설계를 진행한다. 분류모델을 통해 구분된 저농도와 고농도를 기준으로 농도별 특성을 반영하기 위한 저농도 예측 모델과 고농도 예측 모델을 설계한다. 농도별 분류 모델과 예측 모델의 성능 평가를 위해 분류 정확도, 예측 정확도, RMSE(root mean square error)를 이용하여 확인한다.

Ⅱ. 데이터셋 수집 및 구성

2-1 데이터 수집 및 전처리

미세먼지에 영향을 주는 환경 변수들 중 대표적으로 대기오염 물질과 기상 요소들이 있다. 미세먼지 농도의 예측을 위한 연구들은 이러한 환경 변수들을 사용하여 연구가 진행되고 있다 [9-15]. 본 논문에 사용하기 위해 수집한 데이터는 표 1과 같다.

표 1. 주요 데이터 수집
Table 1. Collected data.

variable		number of data
Meteorological elements	Temperature	87,622
	Wind speed	87,612
	Wind direction	87,601
Air pollutants	PM_{10}	249,268
	O_3	255,464
	CO	252,924
	NO_2	254,244
	SO_2	252,623

수집한 데이터는 2009년부터 2018년 동안 시간 단위로 천안 지역에서 측정한 데이터를 사용하였다. 측정소에서 제공하는 데이터 중 PM_{10}과의 상관분석을 통해 기상 요소 중 온도, 풍속, 풍향을 사용하였으며, 대기오염 물질 요소 중 O_3, CO, NO_2, SO_2, PM_{10}을 사용하였다. 수집한 데이터 중 측정소의 점검 및 기타 환경으로 인해 일부 결측 데이터가 존재함에 따라 모델의 오학습 방지를 위해 결측 데이터가 존재하는 시간대의 모든 데이터를 제거하여 구성하였다.

데이터들의 경우 값을 표현하는 형식이나 범위가 서로 다르기 때문에 전처리 과정이 필요하다. 풍향의 경우, 16방위의 범주 형으로 표현되는 데이터이며, 다른 데이터들과 동일하게 수치형 데이터로 표현하기 위해 one-hot encoding을 이용하여 16개의 벡터 형으로 변환하였다. 나머지 데이터들은 min max scaling을 통해 수치 표현의 범위를 동일하게 적용하였다.

2-2 데이터 구성

수집된 데이터는 모델의 학습에 사용되기 위한 training set과 학습 후 모델 평가를 위한 test set으로 구분된다. training set은 순수하게 학습을 위해 사용되는 train set과 학습이 잘 이루어 졌는지 검증하기 위한 validating set으로 구분된다. 본 논문에서는 그림 1과 같이 수집과 전처리가 완료된 데이터를 최종적으로 학습과 평가에 사용될 데이터로 구성하였다.

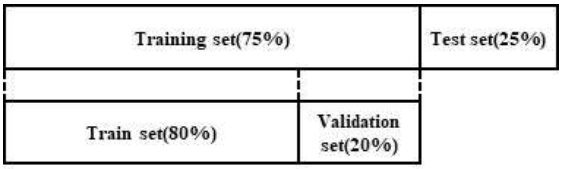


그림 1. 데이터셋 구조
Fig. 1. Structure of dataset

Ⅲ. 농도별 예측 모델 설계

3-1 농도별 분류 모델 설계

농도별 분류 모델은 PM_{10} 값의 $80\mu g/m^3$을 기준으로 저농도와 고농도로 분류하기 위한 모델이다. 분류를 위한 모델의 알고리즘은 여러 알고리즘을 이용하여 분류 성능 평가를 진행한 선행 연구 결과로 도출된 DNN 알고리즘을 사용하였다.

DNN는 신경망 알고리즘으로서 학습 과정에서의 예측을 통해 도출된 값과 실제 값을 활성화 함수를 이용하여 각각 반환된 값이 근사한 수치를 보일 때까지 가중치를 수정하여 모델을 학습하는 알고리즘이다 [16-18].

가중치를 수정하기 위한 활성화 함수는 sigmoid를 사용하였으며, 최적화 함수는 rmsprop, 손실 함수는 binary cross entropy를 사용하였다. 과대적합과 과소적합을 방지하기 위해 hidden layer, node, batch size, epoch, 가중치 규제에 대한 L2와 Dropout rate 값들의 최적화가 필요하다. 표 2는 하이퍼 파라미터 탐색을 통해 도출된 각 파라미터들의 최적의 값이며, 해당하는 값을 기반으로 분류 모델을 설계하였다.

표 2. 하이퍼 파라미터 탐색 결과(분류모델)
Table 2. Hyper parameter search result(classification).

parameter	value
hidden layer	2
node	20
L2	0.001
dropout rate	0.3
batch size	80
epoch	100

3-2 저농도와 고농도 예측 모델 설계

농도별 예측 모델의 경우 분류 모델과 동일하게 DNN 알고리즘을 이용하여 저농도와 고농도에 대한 모델로 설계하였다. 예측 모델의 경우 분류 모델에 사용된 활성화 함수, 최적화 함수, 손실 함수를 ReLU, adam, MSE(mean squared error)로 변경하여 적용하였다. 그 외의 파라미터는 분류 모델의 설계 과정과 동일하게 하이퍼 파라미터 탐색을 통해 최적의 값을 도출하였다.

예측 모델의 경우 저농도와 고농도로 분류된 데이터들을 통해 각각 농도별 예측을 수행하기 위한 모델이다. 따라서 농도에 따라 각 예측 모델이 최적화 되어야 하며, 이에 따라 구성된 dataset 중 PM_{10} 값의 $80\mu g/m^3$을 기준으로 저농도와 고농도를 구분하여 하이퍼 파라미터 탐색을 진행하였다. 표 3은 저농도와 고농도에 대한 예측 모델의 하이퍼 파라미터 탐색 결과이며, 해당하는 값을 적용하여 저농도 예측 모델과 고농도 예측 모델을 설계하였다.

표 3. 하이퍼 파라미터 탐색 결과(예측모델)
Table 3. Hyper parameter search result(prediction).

parameter	low concentration	high concentration
hidden layer	4	2
node	20	40
L2	0	0.001
dropout rate	0	0
batch size	40	80
epoch	100	100

Ⅳ. 성능 평가

저농도와 고농도를 분류하기 위한 분류 모델과 저농도 예측 모델, 고농도 예측 모델을 구축하였으며, test set을 이용하여 모델들의 성능 평가를 진행하였다.

분류 모델의 경우, 저농도와 고농도에 대한 분류 정확도도 평가를 진행하였다. 농도 분류 모델의 경우, 표 4와 같이 전체 데이터의 분류 정확도는 96.39%를 보였다. 저농도 데이터 분류 정확도 경우 97.54%의 정확도를 보였으며, 고농도 데이터 분류 정확도의 경우 85.51%를 보였다. 고농도 분류 결과는 저농도 분류보다 낮은 정확도를 보였다.

표 4. 분류 모델 성능
Table 4. Classification performance.

concentration	real frequency	classification frequency	true frequency	accuracy
low	19,713	19,532	19,229	97.54%
high	2,091	2,272	1,788	85.51%
total	21,804	-	21,017	96.39%

예측 모델들의 경우, 농도 분류 모델을 통해 구분된 저농도와 고농도를 이용하여 각각 학습하였다. 이에 따라 저농도 전용 예측 모델, 고농도 전용 예측 모델을 구축하였으며, 각 모델을 통한 미세먼지 농도 예측 진행 시 사용되는 데이터도 분류 모델을 통해 도출된 저농도와 고농도를 이용하였다. 예측 모델의 성능 평가를 위해 RMSE와 실제 값과 비교를 통한 정확도 확인을 진행하였다.

표 5의 모델별 예측 성능에서 저농도 예측 모델의 경우, 6.6107의 RMSE로 실제 값과 예측 값의 오차범위를 보인다. 고농도 예측 모델의 경우, RMSE가 13.7826으로 저농도 예측 모델보다 큰 오차범위를 보인다. 그러나 저농도와 고농도의 예측 정확도 경우, 저농도는 90.38%인 반면 고농도는 96.37%인 것

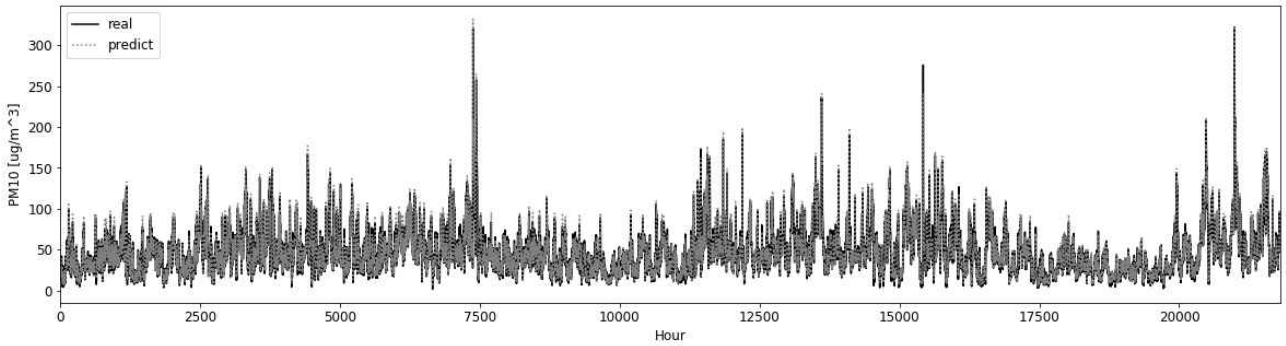


그림 2. 예측 결과
Fig. 2. Predicted results

을 확인할 수 있었다. 이러한 결과는 test set을 구성하고 있는 저농도와 고농도에 해당하는 데이터를 보았을 때, 저농도 데이터 수가 고농도 데이터 수보다 월등히 많은 것을 확인할 수 있으며, 이에 따라 발생한 결과로 볼 수 있다.

표 5. 예측 성능
Table 5. Prediction performance.

indicator	low concentration prediction model	high concentration prediction model
RMSE	6.6107	13.7826
low concentration accuracy	90.38%	-
high concentration accuracy	-	96.37%
false frequency	1 (high concentration)	0 (low concentration)

그림 2는 저농도 예측 모델과 고농도 예측 모델의 예측 결과를 통합하여 표현한 전체 예측 결과이다. 실제 미세먼지 농도 값과 예측 값을 비교하였을 때 근사한 모습을 보였다. 이는 농도별 예측 모델을 통해 저농도와 고농도에 대한 특성이 잘 반영되어 예측 값이 도출된 결과로 볼 수 있다.

Ⅴ. 결 론

미세먼지 농도 예측을 수행함에 있어 미세먼지 발생 환경에 따른 농도별 특성으로 인해 예측 모델 성능의 한계를 보이고 있다. 특히 고농도의 경우 저농도와 비교하였을 때, 발생 빈도수의 차이와 발생 환경에 대한 차이로 예측 성능의 한계를 두드러지게 보이고 있다.

이러한 문제를 해결하기 위해 본 논문에서는 인공신경망 알고리즘을 이용하여 저농도와 고농도로 분류하고 구분된 농도별로 특성을 학습시킨 두 가지 예측 모델을 통해 예측을 수행하는 모델을 설계하였다. 이를 위해 천안지역에서 10년간 수집된 기상 데이터와 대기오염 물질 데이터를 사용하여 학습 데이터를 구성하였다. 구성한 데이터를 PM_{10} 값의 $80\mu g/m^3$을 기준으로 저농도와 고농도로 분류하기 위해 DNN 알고리즘을 이용하여 농도별 분류 모델을 설계하였다. 이후 저농도와 고농도의 특성을 구분하여 학습하고 예측하기 위해 분류 모델을 통해 구분된 저농도와 고농도를 이용하여 DNN 알고리즘 기반의 저농도 예측 모델과 고농도 예측 모델의 설계를 진행하였다. 각 모델의 최적화를 위해 하이퍼 파라미터 탐색을 수행하였다.

평가 결과, 농도별 분류 모델의 경우, 96.39%의 전체 분류 정확도를 보였으며, 저농도와 고농도의 분류 정확도는 97.54%와 85.51%를 보였다. 고농도의 경우, 저농도의 분류 정확도에 비해 상대적으로 낮은 정확도를 보였다. 농도별 예측 모델의 경우 분류 모델을 통해 구분된 저농도와 고농도를 이용한 성능 평가에서 저농도 예측 정확도가 90.38%, 고농도 예측 정확도는 96.37%로 높은 미세먼지 예측 정확도를 보였다. 실제 값과 농도별 예측 값을 비교하였을 때 근사한 모습을 보였으며, 농도별 예측 모델을 통해 저농도와 고농도의 개별 특성 반영의 효과를 확인할 수 있었다. 그러나 농도별 예측 모델의 경우, 1차로 진행되는 농도별 분류 모델의 분류 성능이 농도별 예측 정확도에 영향을 줄 수 있다. 따라서 분류 모델의 성능 향상이 필요하며, 특히 데이터 불균형 문제로 인한 고농도 분류 정확도의 개선이 필요하다. 향후, 분류 모델의 성능 향상과 분류 및 예측 모델의 최적화에 대한 연구를 진행할 계획이다.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2019R1I1A3A01059038) and This paper was supported by the Education and Research Promotion Program of KOREATECH in 2021.

References

[1] C. A. Pope III, and D. W. Dockery, "Health effects of fine particulate air pollution: line that connect," *Journal of the Air & Waste Management Association*, Vol. 56, No. 6, pp. 709-742, Jun. 2006.

[2] A. Valavanidis, K. Fiotakis, and T. Vlachogianni, "Airborne particulate matter and human health: toxicological assessment and importance of size and composition of particles for oxidative damage and carcinogenic mechanisms," *Journal of Environmental Science and Health, Part C*, Vol. 26, No. 4, pp. 339-362, Sep. 2008.

[3] K. H. Kim, E. Kabir, and S. Kabir, "A review on the human health impact of airborne particulate matter," *Environment international*, Vol. 74, pp. 136-143, Jan. 2015.

[4] N. J. Hime, G. B. Marks, and C. T. Cowie, "A comparison of the health effects of ambient particulate matter air pollution from five emission sources," *International Journal of Environmental Research and Public Health*, Vol. 15, No. 6, Jun. 2018.

[5] World Health Organization (WHO), "Health effects of particulate matter: policy implications for countries in eastern europe, caucasus and central asia," *Regional Office for Europe*, 2013.

[6] Board of Adit and Inspection (BAI), "Weather forecast and earthquake notification system operation," *International THE Board of Audit and Inspection of KOREA*, 2017.

[7] H. L. Kim, and T. H. Moon, "Machine learning-based fine dust prediction model using meteorological data and fine dust data," *Journal of the Korean Association of Geographic Information Studies*, Vol. 24, No. 1, pp. 92-111, Mar. 2021.

[8] K. H. Jeon, J. H. Lee, J. H. Park, H. J. Park, Y. H. Lee, M. S. Jung, H. S. Lee, K. P. Nam, J. S. Myoung, K. C. Choi, and T. H. Kim, "A study of data accuracy improvement for national air quality forecasting(Ⅲ)," *National Institute of Environmental Research*, Dec. 2016.

[9] J. M. Han, J. G. Kim, and K. H. Cho, "Verify a causal relationship between fine dust and air condition-weather data in selected area by contamination factors," *The journal of Bigdata*, Vol. 2, No. 1, pp. 17-26, Feb. 2017.

[10] X. Zhao, R. Zhang, J. L. Wu, and P. C. Chang, "A deep recurrent neural network for air quality classification," *Journal of Information Hiding and Multimedia Signal Processing*, Vol. 9, No. 2, pp. 346-354, Mar. 2018.

[11] B. T. Ong, S. Komei, and Z. Koji, "Dynamic pre-training of deep recurrent neural networks for predicting environmental monitoring data," *in 2014 IEEE International Conference on Big Data (Big Data)*, Washington DC, pp. 760-765, 2014.

[12] X. Li, L. Peng, X. Yao, S. Cui, Y.Hu, C. You, and T. chi et al., "Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation," *Environmental Pollution*, Vol. 231, No. 1, pp. 997-1004, Dec. 2017.

[13] Y. B. Lim, I. Aliyu, and C. G. Lim, "Air pollution matter prediction using recurrent neural networks with sequential data," *Proceedings of the 2019 3rd International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, pp. 40-44, Mar. 2019.

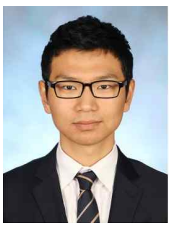
[14] S. W. Kang, N. G. Kim, and B. D. Lee, "Fine dust forecast based on recurrent neural networks," *2019 21st International Conference on Advanced Communication Technology (ICACT)*, pp. 456-459, Feb. 2019

[15] S. Y. Yoo, J. C. Lee, J. H. Lee, H. J. Hwang, and S. S. Lee, "A study on time series data filtering of spar platform using recurrent neural network," *Journal of the Korean Society of Marine Engineering*, Vol. 43, No. 1, pp. 8-17, Jan. 2019.

[16] M. M. Dedovic, S. Avdakovic, I. Turkovic, N. Dautbasic, and T. Konjic, "Forecasting PM10 concentrations using neural networks and system for improving air quality," *2016 XI International Symposium on Telecommunications(BIHTEL)*, pp. 1-6, Oct. 2016.

[17] J. B. Ahn, and Y. M. Cha, "A comparison study of corrections using artificial neural network and multiple linear regression for dynamically downscaled winter temperature over south korea," *Asia-Pacific Journal of Atmospheric Sciences*, Vol. 41, pp. 401-413, Jun. 2005.

[18] J. W. Oh, J. H. Song, K. H. Kim, and S. H. Jung, "Automatic composition using training capability of artificial neural networks and chord progression," *Journal of Korea Multimedia Society*, Vol. 18, No. 11, pp. 1358-1366, Nov. 2015.



조 경 우 (Kyoung-Woo Cho)

2020년 11월 ~ 현재 한국정보통신기술협회 책임연구원
2020년 8월 한국기술교육대학교 전기전자통신공학과 공학박사
2015년 2월 한국기술교육대학교 전기전자통신공학과 공학석사
2013년 2월 공주대학교 전기전자제어공학부 전자공학따노정보공학전공 전자공학트랙 공학사
※관심분야 : 미세먼지 예측, 인공신경망, 심층신경망, Industrial IoT, LPWA

정 용 진 (Yong-Jin Jung)

2018년 3월 ~ 현재 한국기술교육대학교 전기전자통신공학과 박사과정
2016년 2월 한국기술교육대학교 전기전자통신공학과 공학석사
2014년 2월 공주대학교 전기전자제어공학부 전자공학따노정보공학전공 전자공학트랙 공학사
※관심분야 : 미세먼지 예측, 기계 학습, 인공신경망, 심층신경망

오 창 헌 (Chang-Heon Oh)

1999년 3월 ~ 현재 한국기술교육대학교 전기전자통신공학부 교수
2006년 8월 ~ 2007년 7월 방문교수(University of Wisconsin-Madison)
1993년 10월 ~ 1999년 2월 삼성전자(주) CDMA 개발팀 선임연구원
1990년 2월 ~ 1993년 8월 한진전자(주) 기술연구소 전임연구원
1996년 2월 한국항공대학교 항공전자공학과 공학박사
1990년 2월 한국항공대학교 항공통신정보공학과 공학석사
1988년 2월 한국항공대학교 항공통신공학과 공학사
※관심분야 : 무선/이동통신, IoT, 기계학습 기반 통신시스템