https://doi.org/10.5392/JKCA.2020.20.07.674

딥러닝 알고리즘을 활용한 천식 환자 발생 예측에 대한 연구

A Study on Asthmatic Occurrence Using Deep Learning Algorithm

성태응
연세대학교 컴퓨터정보통신공학부

Tae-Eung Sung(tesung@yonsei.ac.kr)

요약

최근 산업화 및 인구과밀화로 인해 대기오염에 대한 문제가 세계적 관심사로 대두되고 있다. 대기 오염은 인간의 건강에 다양한 악영향을 초래할 수 있는데, 그 중 본 연구에서 관심을 둔 천식과 같은 호흡계 질환은 직접적 영향을 받을 수 있다. 기존의 연구에서는 임상 데이터를 활용하여 상대적으로 적은 표본을 기반으로 천식과 같은 질환에 대기 오염 인자가 어떠한 영향을 미치는지를 파악하였다. 이는 수집 표본 별 일관성이 없는 결과를 초래할 소지가 다분하며, 의료계 종사자 이외에는 연구의 시도가 어렵다는 점에서 큰 한계를 가지고 있다. 본 연구에서는 정부에서 공개하는 대기 환경 데이터와 천식 발병 빈도 수에 대한 데이터를 기반으로, 실제 천식 발병 빈도를 예측하는 것에 연구의 주안점을 두었다. 본 연구는 시차를 적용한 피어슨 상관계수를 통해 각 대기오염 인자가 천식 발병에 어느 정도의 시차를 가지고 유의한 영향을 주는지를 검증하였다. 검증 결과를 기반으로 구축된 학습데이터는 딥러닝 알고리즘에 활용되며, 천식 발병 빈도의 예측에 최적화 된 모델을 설계하였다. 모델의 평균 대비 오차율은 약 11.86%로 타 머신러닝 기반의 알고리즘 대비 우수한 성능을 나타냄을 확인하였다. 제안한 모델은 국가 보험 체계 및 보건 예산 관리에서의 효율화 및 병원에서의 의료 인력 배치 및 수급에의 효율성 또한 제공할 수 있다. 또한 만성 천식 질환자에 대한 대기 환경별 발병 위험에 대한 조기 경보를 통해 국민 건강 증진에 기여할 수 있다.

■ 중심어 : | 딥러닝 | 천식 | DNN | 질병 | 대기 | 대기오염 | 보건정책 |

Abstract

Recently, the problem of air pollution has become a global concern due to industrialization and overcrowding. Air pollution can cause various adverse effects on human health, among which respiratory diseases such as asthma, which have been of interest in this study, can be directly affected. Previous studies have used clinical data to identify how air pollutant affect diseases such as asthma based on relatively small samples. This is high likely to result in inconsistent results for each collection samples, and has significant limitations in that research is difficult for anyone other than the medical profession. In this study, the main focus was on predicting the actual asthmatic occurrence, based on data on the atmospheric environment data released by the government and the frequency of asthma outbreaks. First of all, this study verified the significant effects of each air pollutant with a time lag on the outbreak of asthma through the time-lag Pearson Correlation Coefficient. Second, train data built on the basis of verification results are utilized in Deep Learning algorithms, and models optimized for predicting the asthmatic occurrence are designed. The average error rate of the model was about 11.86%, indicating superior performance compared to other machine learning-based algorithms. The proposed model can be used for efficiency in the national insurance system and health budget management, and can also provide efficiency in the deployment and supply of medical personnel in hospitals. And it can also contribute to the promotion of national health through early warning of the risk of outbreak by atmospheric environment for chronic asthma patients.

■ keyword : | Deep Learning | Asthma | DNN | Disease | Atmosphere | Air Pollution | Public Health Policy |

접수일자 : 2020년 06월 04일
수정일자 : 2020년 07월 17일
심사완료일 : 2020년 07월 20일
교신저자 : 성태응, e-mail : tesung@yonsei.ac.kr

I. 서 론

1. 대기오염의 증가와 천식

산업의 고도화 및 도시화, 인구 증가로 인한 인류의 활동성 및 자원의 소비량 증대는 대기 오염 물질의 증가와 더불어 인류의 건강상태까지 위협하는 상황에 이르렀다. 인체에 만성적으로 영향을 미칠 수 있는 대기 오염 물질은 그 영향력이 대규모 인구집단을 대상으로 발현될 수 있다는 점에서 큰 위험성을 내포하고 있다[1]. 일례로, 1952년 발생한 런던의 스모그 현상은 대기의 정체와 대기 오염 물질 농도의 급증으로 인해 대규모의 사상자가 발생한 바 있으며, 이는 대기 오염 물질이 건강에 미치는 유해성에 대한 관심을 유발하기에 충분하였다[2].

대기 오염 물질은 일반적으로 다양한 오염 물질의 복합적 작용으로 인해 단계적으로 생성된다. 따라서 단일 물질의 인체에 대한 영향력 대비 다양한 질병을 유발할 수 있다. 그 중 호흡계 질환의 유병률은 특히 대기 오염 물질과 연관성이 높은 것으로 확인되었다[3-5]. 그 중 도시화와 더불어 전 세계적으로 유병률이 증가, 사회경제적 부담을 초래하고 있는 대표적 호흡계 질환인 천식은 이산화질소, 오존, 이산화황, 미세먼지 등 다양한 대기 오염 물질에 직접적 영향을 받는 것으로 알려져 있다[6-8].

2. 천식에 대한 대기오염인자의 영향력

일반적으로 인체에 유의한 영향을 미칠 수 있는 대기오염 인자는 SO_2, NO_2, O_3 및 CO와 같은 기체성 물질과 미세먼지와 같은 입자성 물질로 분류될 수 있다. 기체성 물질은 대기의 조성 변화에 영향을 미치며 주로 화석 연료의 연소 등 인간의 경제활동에 의한 부산물에 기인한다. 기체성 물질 중 SO_2는 석유를 연소할 때 원유에 함유되어 있는 황이 산화되면서 발생하는 물질이다. 다량의 SO_2에 대한 노출은 기도 수축을 유발, 타 대기 오염 물질과의 상호작용을 통해 천식의 유병률을 높일 수 있는 물질로 확인되었다[9][10].

NO_2는 자동차 배기가스가 주요 발생원으로, 이로 인해 도시화 및 산업화 된 지역에 발생 농도가 높은 경향이 있다. NO_2에 대한 노출은 기도 과민성을 증가시킬 수 있으며, 천식 환자의 폐 기능을 감소시키는 등 천식 발병에 직접적 연관이 있는 물질로 확인되었다[11][12].

O_3는 자동차의 배기가스에서 발생하는 NO_X와 C_XH_X가 태양 광선과 광화학산화반응을 일으킴으로써 생성된다. O_3는 기온이 높아질수록 대기 중 농도가 증가하는 특성이 있으며, 이로 인해 여름철에 인체에 대한 영향력이 증대된다. 대기 중 O_3의 농도가 높아지면 호흡기 환자의 폐 기능 감소 및 기도 과민성 증가를 유발할 수 있다[13][14].

CO는 C_XH_X의 불완전 연소에 의해 발생하는 기체로, 가정에서 사용되는 연소장치 및 자동차 배기가스가 주요 발생원으로 손꼽힌다. CO에 대한 인체의 장기간 노출은 호흡 대사의 방해 등 유해한 영향을 미칠 수 있다[15].

마지막으로 산업화가 유발한 최악의 부산물이자, 세계보건기구(WHO)로부터 1급 발암물질로 지정되기도 한 미세먼지(Particulate Matter, PM)는 고체 상태와 입자와 액체 상태 입자의 혼합물로 구성된 물질이다. 미세먼지는 입자의 직경에 따라 지름이 10μm 미만인 경우 PM_{10}으로, 지름이 2.5μm 이하인 경우 $PM_{2.5}$로 불리운다. 이들은 주로 산업 공정에서의 연소 과정 및 자동차 배기가스에서 발생하는 1차 오염 물질이 다른 물질과의 화학 반응을 통해 생성된다. PM_{10}은 상기도 혹은 기관지에 침작되는 경향이 있으며, $PM_{2.5}$는 상대적으로 작은 입자의 크기로 인해 소기도 혹은 폐포에 침착되는 등 호흡기 질환에 악영향을 초래할 수 있다[16][17].

3. 관련 선행연구

대기 환경 인자의 복합적 상호작용으로 인해 발생하는 천식은 앞서 언급한 바와 같이 전 세계적으로 유병률이 증가, 다양한 사회경제적 문제를 초래하고 있다. 천식의 발병은 개인에 대한 진료비 부담 및 국가적 차원에서의 생산성 감소를 유발할 수 있으며, 질환자에 대한 실직 위험 또한 높은 등 국가 및 개인에 사회적 비용을 발생시킬 수 있다[18].

이처럼 국가적 차원에서 사회적 비용을 야기할 수 있는 천식의 위험성에도 불구하고, 질병에 대한 관리체계

가 상대적으로 미흡한 것이 현실이다[19]. 또한 도시화 및 산업화의 진행이 가속화 된 현 시점에서, 앞서 기술한 바와 같이 대부분의 연구는 임상 사례 기반의 천식 유발 인자에 대한 연구를 기반으로 당장의 실현 가능성이 높지 않은 환경적 정책 제언에 주안점을 두고 있어 한계가 명확하다. 본 연구에서는 실제 조성된 대기 환경에 대해 발생할 수 있는 명확한 환자 수의 예측이 가능하다면, 국가적 차원에서의 보건 정책 전개 및 예산 분배, 병원 내 응급의료인력 양성 및 배치 효율화를 이룩할 수 있을 것이라 판단하였다. 따라서 임상 사례 데이터를 활용한 기존 연구와는 달리, 대기 환경에 대한 장기 데이터를 기반으로 천식 발병을 유발할 수 있는 주요 요인을 도출한 후 실제 천식 발병 환자의 수를 예측하는 모델을 구축하는 것에 연구의 주안점을 두었다.

현재까지 진행된 선행연구에서는 대기 환경 데이터를 질병 발병 빈도 예측에 접목한 사례는 다소 미비한 실정이나, 대기 환경 자체를 예측하는 연구는 다소 진행되었다. McKendry의 연구에서는 대기 환경 인자 및 기상 관측 정보를 바탕으로 미세먼지의 일별 농도를 예측하는 연구를 수행한 바 있으며, 타 모형 대비 DNN의 성능 우수성을 입증하였다[20]. Shahraiyni 외의 연구에서는 미세먼지 예측에 다중회귀분석 모델 대비 NN 모델의 예측 성능이 우수함을 확인하였다[21].

본 연구에서는 우선 학습에 활용된 데이터의 형태를 확인한 후, 천식 발병 환자 수와 대기환경인자 간 영향력을 시차를 구분하여 파악하였다. 이를 바탕으로 학습 데이터를 구성, 딥러닝 모형에 학습하여 천식 환자 수 예측에 최적화 된 모델을 구축하였다. 이후 타 예측 모형과 성능 비교를 통해 구축된 모델의 우수성을 입증하였다.

II. 데이터 수집 및 분석

1. 활용된 데이터

본 연구에 사용된 데이터는 크게 대기 환경 데이터와 천식 발병 빈도 데이터로 구분된다.

그 중, 모형의 독립변수로 활용된 대기 환경 데이터는 환경부에서 운영하는 Air Korea 웹사이트에서 2015년 1월 2일부터 2018년 12월 31일까지 대한민국 서울에 위치한 측정소에서의 시간별 대기 오염 자료를 일별 평균하여 활용하였다. Air Korea는 전국 112개시, 군에 설치된 398개의 측정망을 통해 수집된 대기환경물질의 농도를 국가 대기 오염 정보 관리시스템(NAMIS)에 누적, 대국민 공개하는 사이트로 분석 대상 자료에 대한 신뢰성을 확보하였다 할 수 있다. 해당 사이트를 통해 대기 환경 인자의 종류는 CO, O_3, SO_2, NO_2 및 미세먼지(PM_{10}, $PM_{2.5}$)를 포함하고 있으며, 변수의 시계별 흐름은 [그림 1]과 같다.

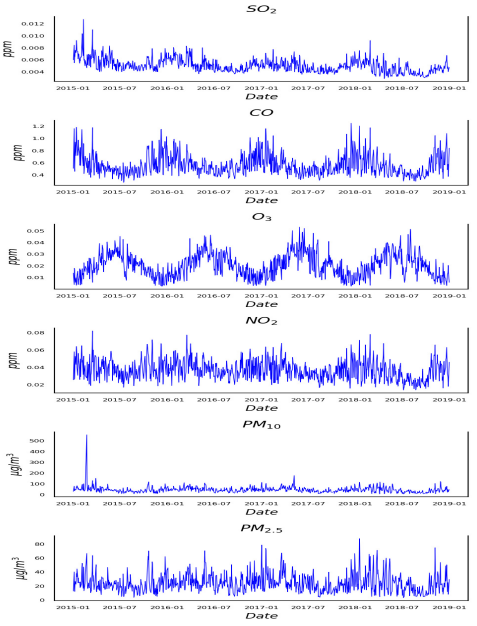

그림 1. 대기 환경 데이터의 일별 흐름

본 연구에서의 예측 목표로 활용할 천식 발병 빈도수 데이터는 건강보험관리공단에서 제공하는 시군구별 과거 진료 건수를 공공데이터포털에서 수집하였고, 2015년 1월 2일부터 2018년 12월 31일까지의 대한민국 서울에서 발생한 진료 건수만을 활용하였다.

수집된 데이터 중, 주말 및 공휴일의 경우 병원 휴진 등으로 인해 실제 발생 가능한 천식 환자 수 대비 과소 계상 될 가능성이 매우 높아 분석의 신뢰성을 확보하고자 제외하였으며, 대기 환경 인자 또한 동 기간의 데이터는 제외하여 자료의 통일성을 확보하였다. [그림 2]는 최종 추출된 천식 발병 빈도의 일별 흐름을 나타낸다.

표 1. 활용된 변수의 기술통계량

Variable	평균	중앙값	첨도	왜도	Percentiles			
					최소	25th	75th	최대
천식 발병 수 (명)	5,859	5,737	0.17	0.56	2,699	4,728	6,724	11,881
대기 환경 인자								
SO_2 (ppm)	0.005	0.005	3.321	1.216	0.003	0.004	0.006	0.013
CO (ppm)	0.564	0.520	1.338	1.217	0.280	0.443	0.64	1.21
O_3 (ppm)	0.020	0.020	-0.377	0.374	0.003	0.012	0.03	0.05
NO_2 (ppm)	0.037	0.036	0.084	0.547	0.014	0.028	0.05	0.08
PM_{10} ($\mu g/m^3$)	46.763	43.393	122.845	7.150	7.975	31.77	57.2	555.69
$PM_{2.5}$ ($\mu g/m^3$)	24.348	22.154	2.249	1.229	3.275	15.37	30.4	87.82

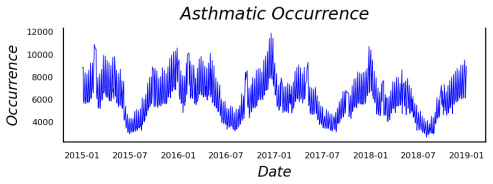

그림 2. 천식 발병 빈도 데이터의 일별 흐름

상기의 과정을 통해 구축된 데이터는 외생변수로 인한 모형 내 영향력을 최대한 배제하고자 하였으며, 대기 환경 인자가 실제 천식 발병에 어떠한 영향을 미치는지 만을 모형에 반영할 수 있도록 설계하였다. 최종적으로 구축된 데이터의 기술통계량은 [표 1]과 같다.

2. 상관분석

앞선 임상 자료 기반 선행연구를 토대로, 본 연구에 활용된 대기 환경 인자들은 천식에 유의한 영향을 미치는 것으로 가정할 수 있다. 본 연구에서는 임상 자료 기반이 아닌, 실제 당일의 대기 환경 데이터를 기반으로 대기 질의 변화가 천식 발병 빈도에 유의한 영향을 주는지 파악하고자 하였다.

이와 더불어, 일반적으로 대기 환경 인자는 전일의 데이터에 후일의 데이터가 영향을 받는 자기상관성이 존재한다고 가정할 수 있다. 또한 대기 환경 인자가 즉각 인체에 유해한 영향을 미칠 수도 있으나, 일반적인 상황에서는 특정 기간 간 환경적 변화의 누적이 인체에 유해한 영향을 미칠 것이라 판단할 수 있다.

이를 종합하여, 본 논문에서는 추출된 대기 환경 인자는 자기상관성을 가지는 시계열(Time-Series) 데이터이며 당일의 대기 환경이 천식 환자에 즉각적 영향을 미칠 수도 있으나, 일정 기간의 누적 시차를 두고 그 영향력이 극대화 될 수 있다고 가정하였다. 이를 모형에 적절히 반영하기 위해 시차 p를 가지는 피어슨 상관관계(Pearson Correlation)을 수행, 실제 천식 발병에 유의성을 나타내는 시차를 검증하고자 하였다. 분석에 활용된 피어슨 상관관계의 산식은 다음과 같다.

$$r_{XY,p} = \frac{\sum_{t=1}^{n}(X_{t-p}-\overline{X})(Y_t-\overline{Y})}{\sqrt{\sum_{t=1}^{n}(X_{t-p}-\overline{X})^2}\sqrt{\sum_{t=1}^{n}(Y_t-\overline{Y})^2}}$$

이때 p는 각 대기 환경 인자의 시차를 의미하며, 본 연구에서는 p의 상한을 일주일, 즉 주말을 제외한 1~5의 범위로 한정하여 단기 시계에서의 영향력을 파악하고자 하였다. 이를 기반으로 분석한 결과는 [표 2]와 같다.

표 2. 대기 환경 인자별 천식 발병 빈도에 대한 피어슨 상관계수 통계량

Variable	1	2	3	4	5
SO_2	0.33 ***	0.32 ***	0.34 ***	0.31 ***	0.33 ***
CO	0.35 ***	0.34 ***	0.37 ***	0.34 ***	0.34 ***
O_3	-0.29 ***	-0.29 ***	-0.29 ***	-0.32 ***	-0.33 ***
NO_2	0.27 ***	0.27 ***	0.32 ***	0.29 ***	0.28 ***
PM_{10}	0.29 ***	0.28 ***	0.23 ***	0.19 ***	0.21 ***
$PM_{2.5}$	0.19 ***	0.18 ***	0.17 ***	0.13 ***	0.14 ***

[1] ***는 1% 유의수준에서 유의함을 의미함.
[2] 변수별 상관계수가 가장 높은 시차는 음영 표기함.

선행연구와 동일하게 모든 대기 환경 인자는 천식 발병에 유의성을 가지는 것이 확인되었다. 또한 대부분의 인자는 천식 발병 빈도에 정(+)의 영향력을 미치는 것으로 파악되었으며, 0_3는 부(-)의 영향을 미치는 것으로 파악되었는데, 이는 기존의 임상 연구 기반의 선행연구[13][14]와 다소 반대되는 결과가 도출되었다. 이는 실제 단일 대기 환경 요소의 영향력이 반영된 결과로, 다양한 변인이 인체에 영향을 미친 결과를 활용한 임상 데이터 기반 분석의 한계점을 인식할 수 있는 결과로 판단할 수 있다. SO_2 및 CO, NO_2는 3일 간의 누적 영향이 천식 발병에 유의성을 가지는 것으로 파악되었으며, O_3는 5일 간의 누적 영향이 천식 발병에 영향을 미치는 것으로 확인된다. 미세먼지의 경우 기체성 물질 대비 입자성 물질의 호흡기에 미치는 영향력을 반영, 전일의 농도가 천식 발병에 즉각 영향을 미치는 것으로 파악되었다.

Ⅲ. 딥러닝 모델

1. 학습 변수의 구성

앞서 본 연구에 활용될 데이터에 대한 피어슨 상관분석을 통해 실제 인자가 천식 발병 빈도에 가장 높은 영향을 미치는 시계를 파악한 바 있다. 본 논문에서 선정한 연구의 목표는 과거의 대기 환경 데이터를 기반으로 미래의 천식 발병 빈도를 예측하는 것이므로, 모형에 학습시킬 변수의 구성에 실제 영향력이 가장 높은 시계를 반영할 필요가 있다. [표 2]의 대기 환경 인자 별 Correlation Coefficient를 토대로 파악한 시차 p는 SO_2, CO, NO_2가 3이며 0_3가 5, PM_{10} 및 $PM_{2.5}$가 1이므로 해당 인자별 시차 p를 기준으로 학습 데이터의 구성을 변경·취합하였다. 예를 들어 천식 발병이 1월 10일의 데이터일 경우 SO_2, CO 및 NO_2는 1월 7일의 데이터를, 03는 1월 5일의 데이터를, PM_{10} 및 $PM_{2.5}$는 1월 9일의 데이터를 같은 행에 배치하여 모든 변수가 천식 발병에 가장 큰 영향을 미치는 시차를 기준으로 학습되도록 하였다. 또한 해당 변수 이외에 주말 및 공휴일 등 병원 휴진이 예상되는 일자 전일 및 후일의 임의적 환자 수 급증을 통제하고자 이에 대한 가변수(*Hol*)를 통제변인으로 추가하였으며, 각 인자가 내포하고 있는 계절성을 감안하여 계절형 가변수 (*Su, Au, Wi*)를 추가하였다. 최종 구축된 학습 변수의 구성은 [표 3]과 같다.

표 3. 시차 p를 반영한 학습데이터의 구성

Original Variable	Final Variable
Asthmatic Occurrence	Asthmatic Occurrence $_t$
SO_2	$SO_{2\ t-3}$
CO	$CO_{\ t-3}$
O_3	$O_{3\ t-5}$
NO_2	$NO_{2\ t-3}$
PM_{10}	$PM_{10\ t-1}$
$PM_{2.5}$	$PM_{2.5\ t-1}$
Hol	$Hol_{\ t}$
Su	$Su_{\ t}$
Au	$Au_{\ t}$
Wi	$Wi_{\ t}$

이와 같이 실제 천식 발병에 가장 연관성이 큰 것으로 판단된 변수별 시차를 기준으로 각 행을 재구성하는 방식은 데이터 내 선형연관성을 예측에 활용된 DNN 모형에 적절히 반영하여 본 연구의 목적인 미래 천식 환자 예측에 대한 성능을 고도화 시켜줄 것이라 기대된다. 구성된 변수 중 2015년~2017년 동안의 데이터는 모형의 구축에 활용될 학습데이터로, 2018년도의 데이터는 모형의 성능 검증을 위한 검증 데이터로 구분하여 추후 모형 성능의 신뢰성 있는 검증에 활용하고자 하였다.

2. 딥러닝 기반 천식 발병 빈도 예측 모델

대기 환경 인자는 천식에 단일 영향력을 미칠 수도 있으나, 인자 간 복합적 상호관계를 기반으로 영향력이 증대 혹은 감소될 가능성이 존재한다. 이러한 복합적 영향력은 일반적으로 결과값에 비선형적 관계를 가질 것이라 판단, 회귀분석 등 선형적 영향관계 기반의 통계적 모델은 예측 모형에의 활용성이 제한적이다. 따라서 본 연구에서는 인자 단일 영향력에 대한 반영 및 인자 간 복합적 영향력을 모두 반영할 수 있는 Deep Neural Networks(DNN) 모델을 활용하여 인자 간 비선형적 패턴을 예측 모형에 복합적으로 반영하고자 하였다.

DNN 모델은 최근 컴퓨팅 파워의 향상 및 빅데이터 확보의 용이성을 기반으로 활용 가능성이 증대된 모델

이다. DNN은 선형 회귀와 같은 일반적인 통계 모델과는 달리 실제 현실의 문제와 유사한 비선형성을 고려한 모델로 평가받는다. DNN은 데이터를 입력받기 위한 입력층, 변수 간 복합적 영향력을 반영하기 위한 복수개의 은닉층과 최종 결과값을 도출해내는 출력층으로 구성된다. 각 은닉층의 노드는 출력층까지 단계적으로 연결되어 있으며, 각 단계별 가중치를 부여하여 이를 토대로 최종 결과값을 도출한다. 상기의 과정은 최종 출력층에서 산출된 오차를 기반으로 각 층 사이의 가중치를 갱신, 모형의 예측 성능에 대한 향상을 목표로 최적화된다.

Ⅳ. 실험 및 결과

본 연구의 목표로 삼은 일별 천식 발생 빈도 예측을 위해 구축된 DNN 모형은 4개의 Hidden Layer에 각 Hidden Layer 별 64개의 노드를 포함, 인자 간 복잡한 패턴을 모형에 반영하고자 하였다. 이외 Hyper Parameter는 최대 200 Epochs, 0.001 Learning Rate로 구성하였다. 이때, 학습 시 Epoch별 Loss의 감소세가 확연히 줄어들 경우, Early Stopping을 반영하여 최적의 Epochs를 학습에 자동 반영하도록 하였다. 또한 학습 데이터에 대한 과적합을 방지하기 위해 첫 번째 Fully Connect 층 이후 Drop-out 층을 추가하여 학습 시 10%의 노드를 배제하도록 하였다. 최종적인 DNN 모형의 구조는 [표 4]와 같다.

표 4. DNN 모형의 구조

Type	Size	Param
Fully Connect	(10 X 64)	704
Drop-out		0
Fully Connect	(64 X 64)	4160
Fully Connect	(64 X 64)	4160
Fully Connect	(64 X 64)	4160
Fully Connect	(64 X 1)	65

초기 데이터가 가진 시계열적 특성은 시차 p를 활용한 변수의 재구성에서 모두 반영 되었다고 판단, Random Sampling을 통해 모형의 과적합을 최소화하고자 하였다. 또한 Adam Optimizer를 활용하여 모델의 최적화를 수행하였으며 손실함수로는 타 지표 대비 학습 과정 및 결과에 대한 평가가 즉각적으로 판단 가능한 MAE(Mean Absolute Error)를 활용하였다. 구축된 예측모형의 Loss Graph는 [그림 3]과 같다. Loss가 학습데이터 및 검증데이터에서 모두 균일하게 감소하고 있는 것을 확인할 수 있으며, 그 차이가 매우 적어, 모형의 과적합이 안정적으로 방지된 것을 확인할 수 있다.

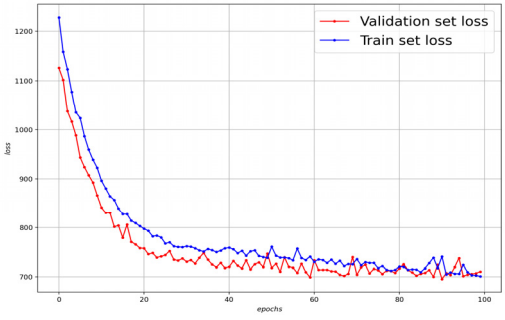

그림 3. DNN의 Loss Graph

구축된 DNN 모형의 검증 데이터를 활용한 실제 예측 결과는 [그림 4]와 같다. 실제 검증데이터 상의 1년간 일별 천식 발병 빈도와 예측 결과가 전체적 흐름이 매우 유사함을 확인할 수 있으며, 변동성 또한 적절히 예측되고 있어 모형의 예측 성능이 높음을 확인할 수 있다.

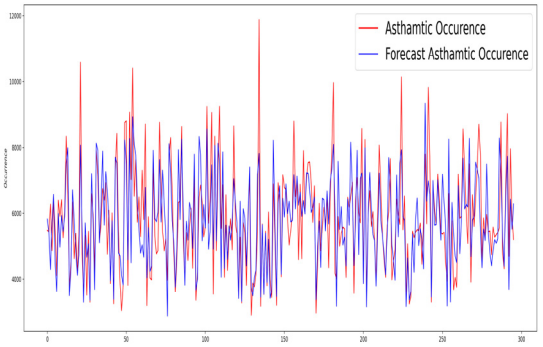

그림 4. 구축된 DNN 모형을 활용한 천식 발병 빈도 예측 결과

최종적으로 Multiple Linear Regression, 시차 p를 활용하지 않은 일반 DNN모형과의 비교를 통해 구축된 모형의 성능을 비교·검증하였다. 성능 평가의 척도로는 회귀 예측모델로 구현된 본 연구의 목적성을 반영, 일

반적으로 활용되는 MAE(Mean Absolute Error), MSE(Mean Squared Error), RMSE(Root Mean Squared Error) 및 MAPE(Mean Absolute Percentage Error)를 모두 도출하여 정밀한 평가를 수행하고자 하였다. [표 5]에서 확인할 수 있듯이 타 모델 대비 본 연구에서 제안한 시차p를 활용한 DNN 모델의 성능이 MAE 기준 일별 평균 천식 발병 빈도 대비 오류율이 약 11.86%로 가장 우수한 것을 확인하였다. 이는 앞서 언급한 바와 같이 천식 발병 빈도에 가장 높은 유의성을 보이는 시차 p를 활용하여 예측성능을 극대화한 점, Multiple Linear Regression에서의 선형성만을 반영한 모형 대비 비선형적 패턴을 모형에 모두 반영할 수 있는 DNN의 특성이 복합적으로 반영된 결과로 판단된다.

표 5. 구축된 모형과 MLR, DNN의 성능 평가

Method	MAE	MSE	RMSE	MAPE
(Proposed)DNN	694.95	846504.75	920.06	11.72
MLR	755.35	967260.21	983.49	13.14
DNN	802.63	1404395.75	1185.07	13.28

V. 결론

산업의 고도화로 인해 인간의 활동성이 증대, 이로 인해 파생된 대기 오염에 관한 문제가 지속적으로 대두되고 있다. 그 중 천식과 같은 호흡기 질환은 이러한 문제에 즉각적 영향을 받을 수 있는 고위험군 임에도 불구하고, 대부분의 국가에서 질병 관리에 우선순위를 두고 있지 않다는 문제가 있다. 본 연구에서는 신산업 시대에 발맞추어 활용성이 증대되고 있는 딥러닝 알고리즘을 활용하여 대기 환경 인자에 대한 천식 발병 빈도를 예측하는 모델 구축에 연구의 주안점을 두었다.

제안한 시차 p 기반의 DNN 모델은 모형 내 학습에서의 인자별 선형 및 비선형적 패턴이 복합적으로 반영될 수 있다는 장점이 있다. 실제 대기 환경 인자가 천식 발병에 어느 정도의 시차를 두고 영향력이 극대화되는지를 파악하였다는 점에서 연구의 의의가 있다. 피어슨 상관계수를 통해 Coefficient가 극대화 되는 시차 p를 파악한 결과, 대부분의 인자는 전일 환경이 당일 천식 발병 빈도에 유의한 영향을 주는 것으로 파악되었으나, SO_2 및 0_3의 경우 약 5일 간의 환경적 누적이 천식 발병에 유의성이 가장 높은 것으로 파악되었다. 그 중 O_3는 기존 임상 실험 기반 선행연구와는 영향력의 방향이 반대되는 결과가 도출되었는데, 이는 임상 실험 기반의 복합적 환경에서 파생되는 질병에 대한 영향력 대비 실제 단일 환경데이터를 활용한 결과의 해석적 특성으로 간주할 수 있다.

최종적으로 구축된 시차 p를 반영한 DNN은 모델은 MAE가 약 11.86%로 예측 성능이 우수함을 입증하였으며, 학습데이터에서의 성능과 검증데이터에서의 성능이 유사한 것으로 보아 모형의 실제적 활용성이 매우 높을 것으로 기대된다. 일반적인 선형 모형인 회귀분석 모형, 시차를 반영하지 않은 DNN 모형과의 성능 비교 결과 또한 구축된 모델이 가장 우수한 것으로 판명되어 모형에의 영향력 있는 시차에 대한 반영이 성능 향상에 유의할 것이라는 가정이 옳음을 입증하였다.

본 연구는 기존의 천식 환자에 대한 임상 데이터 기반의 한정된 표본으로 진행된 연구와는 달리, 상시 업데이트 되는 대규모 데이터를 활용하여 분석을 수행하였다는 점에서 분석 결과 및 성능에의 신뢰성을 확보할 수 있다. 또한 해당 데이터를 확보할 수 있는 모든 국가에서 모형의 일부 응용을 통해 별도 예측 모델을 구축할 수 있다는 점에서 질병 연구에 대한 새로운 방향성을 제시하였다 할 수 있다. 본 모형을 활용하여 미래 발생할 질병 예측을 통해 국가 의료보험 체계 및 예산 배정의 효율성, 의료인력 배치의 효율화를 추구할 수 있을 것으로 기대되며 기존 관리가 미흡했던 기저질환자에 대한 모니터링 시스템 구축에 기준점으로 활용되어 미래 발생할 위험 징후에 대한 사전 Alarming 방안으로 활용 가능할 것이라 기대한다.

참 고 문 헌

[1] 권호장, 조수헌, 김선민, 하미나, 한상환, “설문지에 의한 대기오염의 호흡기계 증상 발현에 관한 조사연구,” 예방의학회지, 제27권, 제2호, pp.313-325, 1994.

[2] B. Bert and T. H. Stephen, "Air pollution and health," The Lancet, Vol.360, pp.1233-1242,

2002.

[3] J. G. Andrew and B. D. Robert, "Inflammatory Lung Injury after Bronchial Instillation of Air Pollution Particles," American Journal of Respiratory and Critical Care Medicine, Vol.164, pp.704-708, 2001.

[4] N. Fredrik, G. Per, J. Lars, B. Tom, B. Niklas, J. Robert, and P. Göran, "Urban Air Pollution and Lung Cancer in Stockholm," Epidemiology, Vol.11, pp.487-495, 2000.

[5] P. Laura, D. Christophe, I. Carmen, A. Inmaculada, B. Chiara, B. Ferran, B. Catherine, C. Oliver, C. B. Francisco, F. Francesco, F. Bertil, H. Daniela, H. Britta, C. Koldo, L. Marina, M. Hanns, O. Peter, R. B. Miguel, M. Sylvia, and K. Nino, "Chronic burden of near-roadway traffic pollution in 10 European cities," European Respiratory Journal, Vol.42, pp.594-605, 2013.

[6] 최기운, 백도명, "우리나라에서의 천식과 대기오염에 관한 연구," 한국역학회지, 제17권, 제1호, pp.64-75, 1995.

[7] 김상헌, 손지영, 이종태, 김태범, 박흥우, 이재형, 김태형, 손장원, 신동호, 박성수, 윤호주, "서울지역 대기오염이 성인 천식 급성 악화에 미치는 영향: 환자교차연구," 대한내과학회지, 제78권, 제4호, pp.450-456, 2010.

[8] O. C. Piotr, D. Piotr, O. J. Aneta, B. Michalina, C. Ernest, O. Tomasz, R. K. Patrycja, and B. Artur, "A Preliminary Attempt at the Identification and Financial Estimation of the Native Health Effects of Urban and Industrial Air Pollution Based on Agglomeration of Gdansk," Sustainability, Vol.12, No.42, pp.1-28, 2020.

[9] 진영주, 박남규, 이현숙, 김대수, 엄재호, 조명찬, 윤세진, 정화숙, 송형근, 성노현, 이상도, "급성 아황산가스 폭로후 흰쥐의 폐에 유발된 염증반응에 관한 연구," 결핵 및 호흡기질환, 제41권, 제4호, pp.328-338, 1994.

[10] G. Henry Jr., S. L. William, L. T. Sheryl, R. A. Karen, and W. C. Kenneth, "Anti-inflammatory and Lung Function Effects of Montelukast in Asthmatic Volunteers Exposed to Sulfur Dioxide," CHEST, Vol.119, pp.402-408, 2001.

[11] 홍수종, 서주희, "기후변화와 건강영향," 대한의사협회지, 제52권, 제2호, pp.149-155, 2011.

[12] Q. K. Jane, "Air pollution and asthma," Journal of Allergy and Clinical Immunology, Vol.104, pp.717-722, 1999.

[13] J. S. Matthew, A. D. Lyndsey, K. Mitchel, W. D. Flanders, A. S. Jeremy, A. W. Lance, E. S. Stefanie, A. M. James, and E. T. Paige, "Short-term Associations between Ambient Air Pollutants and Pediatric Asthma Emergency Department Visits," American J. of Respiratory and Critical Care Medicine, Vol.182, pp.307-316, 2010.

[14] A. S. Robert and K. Ito, "Age-related association of fine particles and ozone with severe acute asthma in New York City," J. of Allergy and Clinical Immunology, Vol.125, pp.367-373, 2010.

[15] 임형준, 이상윤, 윤기정, 주영수, 강대희, 조수현, "대기오염과 천식증상에 의한 응급실내원과의 연관성에 관한 환자교차연구," 대한직업환경의학회지, 제12권, 제2호, pp.249-257, 2000.

[16] Y. Hao and X. Linyu, "Comparative study of PM10/PM2.5 - bound PAHs in downtown Beijing, China: Concentrations, sources and health risk," J. of Cleaner Production, Vol.177, pp.674-683, 2018.

[17] K. M. Jennifer, R. B. John, A. B. Tim, M. M. Kathleen, G. M. Helene, P. Boriana, S. H. Katharine, W. L. Frederick, and B. T. Ira, "Short-Term Effects of Air Pollution on Wheeze in Asthmatic Children in Fresno, California," Environmental Health Perspectives, Vol.118, pp.1497-1502, 2010.

[18] B. Katayoun, D. W. Mary, M. Carlo, L. Larry, A. Kadria, S. John, and J. F. Mark, "Economic burden of asthma: a systematic review," BMC Pulmonary Medicine, Vol.9, pp.1-16, 2009.

[19] M. Matthew, F. Denise, H. Shaun, and B. Richard, "The global burden of asthma: executive summary of GINA Dissemination

Committee Report," Allergy, Vol.59, pp.469-478, 2004.

[20] I. G. McKendry, "Evaluation of artificial neural networks for fine particulate pollution(PM10 and PM2.5) forecasting," J. of the Air&Waste Management Association, Vol.52, No.9, pp.1096-1101, 2002.

[21] H. T. Shahraiyni and S. Sodoudi, "Statistical Modeling Approaches for PM10 Prediction in Urban Areas; A Review of 21st-Century Studies," Atmosphere, Vol.7, No.2, pp.1-24, 2016.

저 자 소 개

성 태 응(Tae-Eung Sung) **정회원**

- 2002년 2월 : 서울대학교 전기공학부(공학사)
- 2004년 5월 : 美) 텍사스오스틴 주립대학교 전기컴퓨터공학과(공학석사)
- 2010년 1월 : 美) 코넬대학교 전기컴퓨터공학과(공학박사)
- 2010년 5월 ~ 현재 : 연세대학교(원주) 컴퓨터정보통신공학부 부교수

〈관심분야〉 : 빅데이터분석, 머신러닝/딥러닝, 기술가치평가, 지능형정보시스템