의료 수요 예측을 위한 딥러닝: 천식 발병을 중심으로

김용선[1], 강현욱[2], 김정연[3], 윤현식[4]
[1]전남대학교 경영학과 박사과정
[2]전남대학교 경영학과 학부생
[3]전남대학교 컴퓨터정보통신공학과경영학과 학부생
[4]전남대학교 경영학과 교수
myhapia@gmail.com, khu147159@naver.com, kjykjy020916@jnu.ac.kr, Dr.yoon@jnu.ac.kr

요 약

질병은 인간의 삶과 사회에 큰 영향을 미치며, 천식은 주변 사람들이 쉽게 그 심각성을 인지할 수 있는 질환이다. 천식의 유병률에도 불구하고 광범위한 데이터를 사용하여 천식 발병을 예측하는 연구가 부족하다. 그리하여 본 연구는 딥러닝 모델을 활용하여 기상 데이터, 대기 오염 물질 수준, 천식 환자 수를 딥러닝 모델로 학습하고자 한다. 날짜, 위도, 경도 좌표를 기준으로 데이터가 병합되며 생성된 통합 데이터세트에서는 천식 환자 수를 결과변수로 사용하고, DNN과 LSTM 모델을 사용해 지도학습을 수행하며 모델의 성능은 MSE 및 MAE를 지표로 사용하여 평가한다. 본 연구는 질병 발생을 사전에 예측함으로써 잠재적인 질병 발생에 대한 사전 정보를 제공하고 의료인력 배치를 최적화하며 의료 서비스의 전반적인 효율성을 향상시켜 사회적 이익을 향상시키고자 한다.

1. 서론

생명체는 생명유지를 위해 일정한 음식과 물, 적절한 온도, 산소 등의 조건이 필요하다. 이러한 조건들이 제대로 공급되지 않거나 신체 내부의 균형이 깨졌을 때 질병은 더 쉽게 발생할 수 있으며 신체의 특정 기관이나 시스템에 영향을 미친다. 그중 대기질이라는 조건은 도시화 과정을 통해 극적인 변화가 있었고 대기오염이 조기 사망의 주요 원인이 되었다. 천식은 흔한 호흡기 질환 중 하나이며 대기오염 물질 노출에 영향을 받는다.[1]

그중 소아 천식은 흔히 알레르기와 연관되어 있으며 대규모 코호트에서 알레르기 감작이 소아 천식의 발병 및 지속의 위험인자로 나타났다. 성인의 경우 알레르기 감작이 천식 및 성인발병 천식의 위험인자로 보고되고 있지만, 성인 천식은 알레르기보다 비알레르기인 경우가 더 많다.[2]

한편 알레르기에 대한 인간의 질병과 건강 영향을 파악하고자 노력이 증가하고 있으며, 한국 또한 알레르기를 관리하기 위한 정책 개발을 진행하고 있다.[3]

알레르기 원인으로 언급되는 요인 중 사람이 항상 접하는 환경, 즉 대기환경, 오염물질, 꽃가루 등은 알레르기에 나쁜 영향을 미치는 것으로 보인다.[4][5]

언급한 것처럼 천식을 포함한 알레르기는 다양한 위험인자에 의해 발생한다. 이러한 문제를 완화하고자 질병 상황에 따라 원인을 파악하려는 많은 연구가 이루어졌지만 천식 발병에 대해 대용량 데이터를 이용하여 딥러닝 모델로 예측하는 연구는 많지 않았다.

2. 선행연구

1)천식 특성

천식 증상은 기침, 쌕쌕거림, 호흡 곤란, 가슴 답답함으로 이어지는 기도의 만성 염증성 질환이다. 기도의 염증에 의해 유발되며, 이는 점액 생성, 기도 벽의 리모델링, 기관지 과민반응(BHR, 냉기와 같은 비특이적 자극에 평활근 세포가 반응하는 경향) 등의 과정을 유발한다. 알레르기 천식은 소아기에 시작되는 경향이 있으며, 아토피 피부염이나 알레르기 비염과 같은 다른 알레르기 질환에서도 볼 수 있는 Th2 세포 반응과 관련이 있으며 이러한 형태의 천식은 집먼지진드기(HDM), 꽃가루와 같은 환경성 알레르겐과의 조우에 의해 유발된다.[6]

2)천식 또는 호흡기 질환 발병 연구

천식의 발생 빈도를 예측하기 위해 독립 변수로 초미세먼지(PM2.5), 미세먼지(PM10), 오존(O3), 아황산가스(SO2), 일산화탄소(CO), 이산화질소(NO2)와 같은 대기오염 데이터를 사용했으며, 시간 지연

요소를 반영한 딥러닝 모델을 통해 회귀 모델과 기존의 딥러닝 모델보다 좋은 성능을 보여주었다.[7]

한편 천식이 다인성 질환이며 공기 중 오염 물질에 대한 노출은 천식 진단에 대한 임상적 포인트 중 하나로 보고 있다.[8]

미세먼지(PM10)로 인한 호흡기 질환, 만성폐쇄성 폐질환, 간질성 폐질환과 관련된 의료비와 환자 수를 추적한 Panel VAR 모형 연구에서는 미세먼지 농도가 증가함에 따라 환자 수와 의료비가 동시에 증가하는 양상을 보였다.[9]

3. 데이터 수집

1)데이터 구성

본 연구에서 수집되는 데이터는 크게 2017년 1월 1일~2017년 12월 31일 기상 데이터, 대기오염 데이터, 천식 발병 데이터로 구성하였다. 기상청 기상자료개방포털에서 기상데이터, 한국환경공단 에어코리아에서 대기오염 데이터, 국민건강보험공단 NHISS에서 천식환자수 데이터를 수집하였다.

<표 1> 2017년 천식 raw데이터

순서	속성명	내용	형식
1	DT	날짜	문자
2	DAY	요일	숫자
3	ADDR_NUM	주소 일련번호	숫자
4	AGE	연령대	숫자
5	SEX_TYPE	성별	숫자
6	SYD	거주기간	숫자
7	ASTHMA_INOUT	천식 입원·외래 실인원수	숫자

2)데이터 전처리와 병합

기상 데이터, 대기오염 데이터, 천식 환자수 데이터를 일자, 주소 기준으로 병합하였다.

수도권 주변을 중심으로 데이터를 병합하였고 기상데이터의 경우 11 지점, 대기오염 데이터의 경우 101 지점, 천식 환자수 데이터의 경우 1208 지점의 데이터가 존재한다.

기상 데이터의 지점을 중심으로 최단거리로 매칭하여 병합하였다. 좌표값의 경우 위경도 도분초, WGS84, UTM을 WGS84로 변환하여 활용하였다. 각 데이터셋 좌표값을 유클리디안 거리로 최단거리를 파악하여 매칭하였다.

<표 2> 구성 데이터

순서	변수명	비고
1	측정일	메타데이터
2	위치	메타데이터
3	평균기온(℃)	위험인자
4	최고최저기온차(℃)	위험인자
5	최대 순간 풍속(m/s)	위험인자
・・・・・・・・・・・・생략・・・・・・・・・		
21	SO_2	위험인자
22	CO	위험인자
23	O_3	위험인자
24	NO_2	위험인자
25	PM_{10}	위험인자
26	Asthma	질병

4. 연구모델

1)표준 딥러닝

딥러닝은 복잡한 수준의 추상화를 통해 데이터를 학습하는 강력한 도구이다. 표준 딥러닝 모델을 사용하면, 다수의 과정으로 구성된 계산을 통해 데이터를 학습할 수 있다. 이 과정에서 역전파 알고리즘을 활용하여 각 계층의 값을 계산하는 데 사용되는 내부 매개변수를 조정한다. 이를 통해 대규모 데이터 세트에서 복잡한 구조를 발견할 수 있다. 그리고 딥러닝 아키텍처는 간단한 모듈들의 다층 스택으로 구성되며, 대부분의 모듈이 학습 대상이다. 이 중 상당수는 비선형 입출력을 계산한다.[10]

딥러닝에는 여러 유형이 있는데 표준 DNN(Deep Neural Network) 모델, 시차 상관관계 데이터를 이용한 DNN모델, LSTM(Long short-term memory) 모델로 학습하여 예측 성과를 비교하고자 한다.

2)시차 상관관계 적용 딥러닝

시차 상관관계를 적용한 딥러닝은 기상 데이터와 대기오염 데이터가 특정 간격으로 천식 환자 수에 영향을 미칠 수 있다는 가정을 바탕으로 생각된 모델이다. 이 모델은 천식 환자 수와 31일 전부터 당

일까지의 독립변수 데이터 간의 상관관계를 분석하여, 독립변수 시차 중 상관계수가 가장 높은 데이터를 선택하여 학습시키고자 한다

3)LSTM

LSTM은 Long Short-term memory의 약자이며 순차 데이터와 관련된 여러 학습 문제에 대한 효과적이고 확장 가능한 모델로 등장했고 장기적인 시간적 의존성을 포착하는 데 일반적이고 효과적이다. LSTM 아키텍처의 핵심 아이디어는 시간이 지나도 상태를 유지할 수 있는 메모리 셀과 셀 안팎으로의 정보 흐름을 조절하는 비선형 게이팅 장치이다 (Greff et al., 2016).

이에 기상 데이터와 대기오염 데이터가 천식 데이터에 미치는 영향이 순차적이며 누적적이라고 가정하여 고려한 모델이다. 6일 과거 데이터가 누적적으로 영향을 미치도록 조정하였다.

4)가설 도출

표준 DNN, 시차 상관관계 DNN, LSTM으로 대기환경 데이터와 천식 의료 입원 이용수 데이터 간의 관계를 파악하고자 하였고 각 모델의 시계열 구조가 다르므로 간접적인 가설을 도출하였다.

<표 3> 도출된 가설

순서	가설 내용
H1	대기환경 데이터가 천식 외래 환자수에 일정 시차를 두고 영향을 미칠 것이다.
H2	대기환경 데이터가 천식 외래 환자수에 누적적으로 영향을 미칠 것이다.

5. 성능평가

1)평가 지표

본 연구에서는 MSE(Mean Squared Error), MAE(Mean Absolute Error)를 사용하여 모델의 성능을 평가하였다.

<Eq 1> MSE 공식

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})}{n}$$

<Eq 2> MSE 공식

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y_i}|}{n}$$

천식 환자수를 예측하기 위한 모델 구조는 64개 노드로 시작하여 은닉층 6부분을 Dropout 30% 2번으로 실행하였다. 그리고 하이퍼파라미터의 경우 Epochs 100, Learning rate 0.01, Batch size 32, Optimizer adam으로 작동시켰다.

검증 데이터를 통해 각 모델을 평가한 결과 아래 <표 4>를 통해 확인할 수 있다. MSE 기준 LSTM 모델이 가장 좋은 성능을 보여주고 있다. 그리고 표준 딥러닝 모델과 시차 상관관계를 이용한 딥러닝 모델 간의 MSE 차이는 약 40 정도로 나타났다.

모델	MSE	MAE
DNN	497.48	14.56
CODNN	448.16	13.74
LSTM	223.16	9.43

6. 결론

자연환경의 변화에 따른 사회현상을 높은 정확도로 예측하는 것은 매우 중요하고 가치 있는 일이다. 더불어 중요한 것은 이러한 예측의 근본적인 이유를 이해하는 것인데 딥러닝 모델의 한계는 해석 가능성이 부족하다는 것이며, 이는 지속적인 문제이다.

이를 완화하기 위해 딥러닝 표준 모델, 시차 적용 모델, 순차 및 누적 모델의 세 가지 모델을 적용하여 성능을 평가함으로써 이러한 현상 반영한 데이터를 해석하는 방법을 더 잘 이해할 수 있다. LSTM 모델이 상대적으로 우수한 성능을 보여 기상 조건과 대기 오염이 시간이 지남에 따라 누적된 영향을 미친다는 것을 추정할 수 있다.

하지만 모델 간의 성능 차이가 유의미한 것인지

추가적인 시험과 해석이 필요하다. 그리고 본 연구에서 활용한 지역 제한적인 데이터를 전국으로 확장하 연구하면 더 정확한 결과가 도출될 수 있을 것이다.

7. 사사

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 소프트웨어중심대학사업의 연구결과로 수행되었습니다.(2021-0-01409)

참고문헌

[1] Eguiluz Gracia, I., Mathioudakis, A. G., Bartel, S., Vijverberg, S. J., Fuertes, E., Comberiati, P., ... & Hoffmann, B. The need for clean air: the way air pollution and climate change affect allergic rhinitis and asthma. Allergy, 75(9), 2170-2184. 2020.

[2] Cevhertas, L., Ogulur, I., Maurer, D. J., Burla, D., Ding, M., Jansen, K., ... & Akdis, C. A. Advances and recent developments in asthma in 2020. Allergy, 75(12), 3124-3146. 2020.

[3] Yoo, S., Kim, D. W., Kim, Y. E., Park, J. H., Kim, Y. Y., Cho, K. D., ... & Lee, E. J. “Data resource profile: the allergic disease database of the Korean National Health Insurance Service.” Epidemiology and Health, 43. 2021.

[4] 홍소영, 손동국, & 권호장. “기후변화와 알레르기 질환”. Pediatric allergy and respiratory disease, 제20권, 제3호, 151-158. 2010.

[5] Singh, A. B., & Kumar, P., “Climate change and allergic diseases: An overview.” Frontiers in Allergy, 3, 964987. 2022.

[6] Hammad, H., & Lambrecht, B. N. The basic immunology of asthma. Cell, 184(6), 1469-1485. 2021.

[7] 성태용. “딥러닝 알고리즘을 활용한 천식 환자 발생 예측에 대한 연구”.한국콘텐츠학회논문지, 제20권 제7호, 674-682. 2020.

[8] Chatkin, J., Correa, L., & Santos, U. External environmental poll-ution as a risk factor for asthma. Clinical reviews in allergy & immunology, 62(1), 72-89. 2022.

[9] 이해춘, 안경애, & 김태영. “미세먼지로 인한 호흡기 질환 발생의 사회경제적 손실가치 분석: Panel VAR 모형을 중심으로.” 경영컨설팅연구, 제18권 4호, 173-186. 2018.

[10] LeCun, Y., Bengio, Y., & Hinton, G. “Deep learning”. nature, 521(7553), 436-444. 2015.

[11] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. “LSTM: A search space odyssey.”, IEEE transactions on neural networks and learning systems, 28(10), 2222-2232. 2016.