

AI 기반 호흡기 질환 진단 기술 개발을 위한 ICBHI Challenge 데이터 분석 및 고찰

양해리, 이성엽*, 우영춘*, 백옥기*

(주)에비드넷, *한국전자통신연구원

haeri.yang@evidnet.co.kr, *sylee549@etri.re.kr, *ycwoo@etri.re.kr, *ok.baek@etri.re.kr

Analysis on ICBHI Challenge Data for AI-based Respiratory Disease Diagnosis

Yang Hae Ri, Lee Sung Yup*, Woo Young Choon*, Baek Ock Kee*

EvidNet, *Electronics Telecommunications Research Institute

요 약

전자 청진기가 널리 보급되고 상용화되면서 많은 양의 청진음 데이터를 간편하게 수집할 수 있게 되었고, 수집된 데이터를 활용한 AI 기반 호흡음 분류 및 진단 기술 개발 연구도 활발히 이루어지고 있다. 전자 청진기로 수집된 ICBHI Challenge Respiratory Sound 데이터 셋은 공개된 호흡음 데이터 중에서 가장 규모가 크며 호흡음 관련 연구에 널리 활용되고 있다. 본 논문에서는 ICBHI Challenge 데이터의 특성에 대해 살펴보고, 이 데이터를 활용하여 AI 기반 호흡기 질환 진단 기술을 연구할 때 유의할 점들을 제시한다. 관련 유의 사항을 적절히 반영한 모델과 그렇지 않은 모델을 서로 비교 분석하여 그 중요성을 확인하였고, 이를 바탕으로 추후 관련 연구에 ICBHI Challenge 데이터 셋을 활용할 때 고려해야 할 점들을 제시한다.

I. 서 론

호흡기 질환의 진단 방법에는 청진, 폐 기능 검사, CT, X-ray 등이 있으며 그 중 청진은 최소비용으로 가장 기본이 되면서도 중요한 진단 방법이다. 하지만 청진은 의사의 진료 경험, 숙련도 등 청각에만 의존한 주관적인 판단에 의한 오진의 가능성도 크며, 현재 코로나 확산으로 인한 호흡기 관련 질환 증가로 인해 청진을 통한 모니터링 및 조기진단을 할 수 있는 의료 인력은 매우 부족한 실정이다. 최근 이러한 문제를 해결하기 위해 AI를 활용한 호흡음 분석을 통해 폐 질환을 진단하는 연구가 활발히 진행되고 있다.[1]. 전자 청진기의 기술적인 발달로 인해 간단하고 편리하게 청진음을 수집할 수 있게 되었고, 수집된 호흡음을 전처리하고 특징 추출하는 단계를 거쳐 기계학습 기반의 모델을 이용하여 호흡음을 분류하거나 질병을 진단하는 식의 연구가 다각도로 이루어지고 있다.[2]

이와 관련하여 AI 기반 호흡음 진단 기술 개발 연구에 쓰이는 대표적인 공개된 데이터 중에서 가장 규모가 큰 호흡 데이터 셋인 ICBHI Challenge는 호흡음 샘플 뿐만 아니라 환자의 기본적인 정보, 수집에 대한 부가 정보들을 포함하고 있어 관련 연구에 널리 활용되고 있다. 본 논문에서는 ICBHI Challenge가 제공하는 데이터 형태와 내용을 간략히 짚어보고, ICBHI Challenge 데이터를 활용할 때 유의해야 할 사항들을 비교 실험을 통해 확인한다. 이를 바탕으로 추후 연구에 이 데이터를 활용할 때 고려해야 할 점들을 제시한다.

II. 본 론

ICBHI Challenge 데이터 셋에는 표 1과 같이 126명의 환자로부터 수집한 920개의 호흡음 파일이 있고 총 6898개의 호흡 주기 샘플로 구성되어 있다.[3] 모든 호흡 주기들의 시작 시점과 끝 시점이 주석 처리되어있고 각 샘플은 정상 호흡음, 수포음, 천명음, 수포음+천명음의 4개 클래스로 구분되어있다. 환자마다 진단 질병과 기본 통계정보(나이, 성별, 성인 BMI, 청소년 키, 청소년 몸무게)도 제공하고 있어 질병 진단을 위한 모델

을 만드는 데에 유용하게 활용될 수 있다. 이 외에도 이상 호흡음이 발생한 10,775개의 이벤트에 대한 정확한 시점과 어떤 이상 호흡음이 발생했는지도 라벨링 되어 있어 추후에 이상음 감지 연구 등에 활용될 수 있다.[4]

실험 모델 설정

ICBHI Challenge 데이터는 wav 형태의 음원으로 제공되며, 기계학습 모델에 활용할 특징 추출을 위해서 스펙트로그램 형태로 변환하고 합성곱 신경망(Convolutional Neural Network)같은 딥러닝 모델을 이용해 분류 모델을 만드는 방식으로 많이 시도되고 있다.[5] 본 논문에서는 ICBHI Challenge 데이터 셋의 특성을 파악하기 위한 비교 실험을 위해 전통적인 머신러닝 분류기를 사용하였고, 모델 설정은 다음과 같다.

호흡음 데이터를 활용하기 위해 librosa 라이브러리를 이용해 스펙트로그램으로 변환 후 40개의 MFCC(Mel-frequency spectral coefficient)를 추출하여 사용했다. 데이터의 변수 개수가 일관되도록 프레임 길이가 80 이상인 오디오 신호는 MFCC의 중간에 80 프레임을 샘플링하고, 80 프레임보다 짧은 오디오 신호는 부족한 부분의 프레임에 0을 패딩했다. 그 후 MFCC를 1차원 벡터로 펼치고 차원 축소를 위해 주성분분석(PCA)을 적용했다.[6] 최종적으로 추출된 35개의 변수를 학습 데이터의 특성으로 활용했다. 분류 모델 알고리즘은 XGBoost (eXtreme Gradient Boosting)이며, 기계학습 자동화 프레임 워크인 Pycaret을 사용해 10-겹 교차 검증(10-fold Cross Validation)으로 하이퍼 파라미터를 최적화한다.

표 1. ICBHI Challenge 데이터 셋의 주요 속성[3]

환자 수	음원 수	호흡 주기 수	클래스 구분
126명	920개	6898개	정상 (3632개, 53%)
			수포음 (1864개, 27%)
			천명음 (886개, 13%)
			수포음+천명음 (506개, 7%)

수집 도구에 따른 데이터 품질 차이

ICBHI Challenge는 서로 다른 두 국가의 연구팀이 독립적으로 수집한 오픈 데이터베이스이므로 각기 다른 환경에서 수집되었고 수집 도구로 사용된 전자 청진기의 종류에 따라라도 데이터의 음질이나 노이즈 형태들이 일관되지 않다는 특성이 있다.[3] ICBHI Challenge는 수집에 사용된 네 가지의 도구를 명시하였는데, 그 중 Meditron, Litt3200, LittC2SE는 전자 청진기이며, AKGC417L 모델은 전자 청진기가 아닌 일반 마이크이다. 마이크로 수집된 소리는 전자 청진기로 수집된 소리에 비해 배경 노이즈가 크고 호흡음보다 심음이 비교적 크게 들려 호흡음 소리를 분석하는데에 어려움이 있다. 해당 마이크로 수집된 파일은 전체 데이터의 약 70%를 차지하고 있어 전반적인 분류 모델의 성능을 저해하는 원인이 될 수 있다.

전체 데이터를 마이크로 수집된 데이터와 전자 청진기로 수집된 데이터로 나누어 각각의 분류 정확도를 비교해 보았다. 표 2에서 볼 수 있듯이 전자 청진기로 수집된 데이터만을 테스트 세트로 사용했을 때의 분류 정확도가 더 좋은 것을 볼 수 있는데, 이는 전자 청진기가 마찰음, 말 소리 등의 잡음을 어느 정도 제거할 수 있고 심장음은 감쇄, 호흡음은 증폭시키는 기능이 있어 일반 마이크보다 품질이 높은 호흡음을 수집하기 때문이다. 보다 나은 모델을 얻기 위해서 수집 도구의 차이에 따른 데이터의 품질 차이를 고려하여 마이크로 수집된 저품질의 데이터에는 추가적인 전처리를 적용하는 것도 하나의 방법이다.

호흡음 특성을 고려한 데이터 분할 기준

평가에 쓰일 데이터는 모델이 보지 못한 데이터를 사용해야 정확한 성능 평가를 할 수 있으므로 훈련 데이터와 평가에 사용할 테스트 데이터를 명확히 구분하여 사용해야 한다는 것은 잘 알려진 사실이다. 이 때 데이터의 특성을 정확히 파악한 후 적절한 데이터의 분할 기준을 마련해야 실험에서 왜곡되지 않은 평가를 얻을 수 있다.[6]

ICBHI Challenge 데이터는 126명의 환자에서 한 환자당 약 55개의 샘플을 수집해 총 6898개의 샘플이 수집되었다. 기계학습 기반 실험에서 전체 데이터 셋을 임의로 분할하여 훈련/테스트 셋을 나누는 것은 매우 일반적인 방법이나, 같은 환자에서 수집된 샘플들끼리는 유사한 클래스 양상을 보이므로 같은 환자에서 나온 호흡 주기 샘플들은 같은 세트에 포함되도록 훈련/테스트 데이터를 분할하는 것이 바람직하다. ICBHI Challenge에서는 같은 환자 단위로 훈련 데이터와 테스트 데이터를 6:4 비율로 나눈 공식 분할 정보를 제공한다. 표 3은 ICBHI Challenge의 공식 분할과, 환자를 고려하지 않고 전체 데이터 셋에 대한 임의 분할을 적용한 데이터 셋을 사용해 모델을 학습하고 분류 성능을 비교한 결과이다. 임의 분할로 실험한 결과에서 볼 수 있듯이 같은 환자로부터 수집된 호흡 주기 샘플들이 훈련 데이터와 테스트 데이터에 섞여 들어가 결과가 왜곡되어 정확도가 높게 나온 것으로 보인다. 결과적으로 이러한 데이터의 도메인 특성을 고려하여 올바른 분할 기준을 세워서 실험에 적용해야 유의미한 결과를 도출할 수 있을 것이다.

표 2. 수집 도구 별 샘플 분포와 분류 정확도

수집 도구	모델명	샘플 분포	분류 정확도
일반 마이크	AKGC417L	4346 (63%)	45%
전자 청진기	Meditron	1467 (21%)	68%
	LittC2SE	583 (9%)	
	Litt3200	502 (7%)	

표 3. 분할 기법에 따른 모델 성능 평가

분할 기법	정확도	거시 평균 F1-Score
임의 분할	70%	57%
ICBHI Challenge 공식 분할	53%	23%

III. 결 론

본 논문에서는 비교 실험을 통해 ICBHI Challenge 데이터가 가진 특성들을 파악하고 호흡음 분류 연구에 ICBHI Challenge 데이터 셋을 활용하기 위해 고려해야 할 점들을 제시했다.

ICBHI Challenge 데이터 셋은 다른 호흡음 데이터 셋과는 달리 호흡 주기마다 주석 처리가 되어 있는 점, 호흡음 뿐만 아니라 청진 위치, 환자의 기본 정보들까지도 제공하고 있다는 장점들이 있어 호흡음 분류를 통한 호흡기 질환 진단 모델을 만드는 연구에 쓰기에 충분히 매력적인 데이터이다. 반면에 일반적으로 기계학습 모델은 대규모의 데이터를 필요로 한다는 점에서 ICBHI Challenge 데이터 수는 좋은 성능을 내기에 충분하지 않고, 대부분의 의료 데이터가 가지고 있는 데이터 불균형 문제 또한 가지고 있다. 또한 기계학습에 적용하기 위해선 노이즈 핸들링, 특징 추출, 데이터 증강 등 까다로운 데이터 전처리 과정이 요구될 수 있고 이러한 한계점을 해결하기 위해 전처리 관련 연구들도 활발히 이루어지고 있다.[7]

본 논문에서 제시한 ICBHI Challenge 데이터의 고유한 특성과 한계점을 고려하여 올바른 방법으로 데이터를 사용한다면 유의미한 연구 결과를 도출할 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음. [22ZS1100, 자율성장형 복합인공지능 원천기술 연구]

참 고 문 헌

- [1] Renard X, S. B, and Esther Rodriguez-Villegas. 2017. Automatic adventitious respiratory sound analysis: A systematic review. PLoS One 12, 5. <https://doi.org/10.1371/journal.pone.0177926>
- [2] Tong X, Jing H. "Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues" Experimental Biology and Medicine 2022
- [3] B. M. Rocha et al., "An open access database for the evaluation of respiratory sound classification algorithms," Physiol. Meas., vol. 40, no. 3, p. 35001, 2019.
- [4] "ICBHI 2017 Challenge," 2022, (https://bhichallenge.med.auth.gr/ICBHI_2017_Challenge).
- [5] Chanane, Hassen, and Mohammed Bahoura. "Convolutional neural network-based model for lung sounds classification." 2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS). IEEE, 2021.
- [6] F. Demir, A. Sengur, and V. Bajaj, "Convolutional neural networks based efficient approach for classification of lung diseases," Heal. Inf. Sci. Syst., vol. 8, no. 1, pp. 1-8, 2020.
- [7] Saldanha J, Chakraborty S, Patil S, Kotecha K, Kumar S, Nayyar A (2022) Data augmentation using Variational Autoencoders for improvement of respiratory disease classification. PLoS ONE 17(8): e0266467. <https://doi.org/10.1371/journal.pone.0266467>