https://doi.org/10.7840/kics.2025.50.12.1974

온디바이스 AI를 위한 딥러닝 모델 성능 비교 연구

박 진 호◆, 홍 혁 기°

Performance Analysis of Deep Learning Models for On-Device AI

Jinho Park◆, Hyuck Ki Hong°

요 약

COVID-19에 따른 사람들의 호흡기 질환에 대한 관심도의 증가로 다양한 Artificial Intelligence (AI) 기반 질병 탐지 연구가 진행되었다. AI 기반 질병 탐지 기술은 청진기를 통해 측정되는 폐 호흡음을 통해 질병 상태를 분류한다. 기존 AI 기반 질병 탐지 기술은 높은 정확도와 빠른 추론 결과를 보여주기 위해 컴퓨팅 리소스가 풍부한 서버를 활용한다. 서버의 이용은 폐 호흡음과 같은 여러 정보가 네트워크를 통해 전달되어야 한다. 네트워크를 통한 정보 전달에 따른 개인정보의 문제가 발생할 수 있다. 이러한 문제를 해결하기 위해 기기 자체에서 AI 모델이 동작하여 결과를 보여주는 온디바이스 AI가 주목받고 있다. 온디바이스 AI는 내부에서 데이터를 수집 및 처리하여 개인정보 문제가 적다. 온디바이스 AI에 다양한 딥러닝 모델이 적용되어 활용될 수 있지만 리소스 부족으로 인한 성능 저하로 적절한 모델이 선택되어야 한다. 본 논문은 모델을 서버와 온디바이스에서 실행하였을 때 질병 분류 성능과 탐지 성능을 분석하고 평가한다. 실험 결과 딥러닝 모델은 서버에서 동작하는 것과 비교하여 온디바이스에서 동작할 때 성능 저하가 발생하는 것을 확인하였다.

키워드 : 온디바이스 AI, 딥러닝, 질병 진단, 폐 호흡음
Key Words : On-device AI, Deep learning, Disease diagnosis, Lung respiratory sound

ABSTRACT

Due to the increased public interest in respiratory diseases following the outbreak of COVID-19, various artificial intelligence (AI)-based disease detection studies have been actively conducted. AI-based disease detection classification by analyzing lung sounds measured through stethoscopes. Conventional AI-based detection schemes typically rely on resource-rich servers to achieve high accuracy and fast inference times. Utilizing servers requires transmitting information such as lung sounds over a network, which raises concerns about personal data privacy. To address this issue, on-device AI—where the AI model runs locally on the device—has been gaining attention. On-device AI collects and processes data internally, thereby minimizing privacy concerns. Although various deep learning models can be deployed for on-device AI, performance degradation due to limited computing resources necessitates careful model selection. This study analyzes and evaluates the disease classification and detection performance of models executed on both server and on-device environments. Experimental results show that deep learning models have lower performance when operated on-device compared to when operated on a server.

※ 본 연구는 과학기술정보통신부 고신뢰의료기기 위험관리를 위한 다차원 공격 표면관리 핵심기술 개발(RS-2024-00399373) 지원으로 수행되었습니다.
◆ First Author : Korea Electronics Technology Institute, parkjh@keti.re.kr, 정회원
° Corresponding Author : Korea Electronics Technology Institute, hkhong@keti.re.kr, 정회원
논문번호 : 202507-179-D-RU, Received July 29, 2025; Revised September 12, 2014; Accepted September 23, 2025

Ⅰ. 서 론

호흡기 질환은 전 세계적으로 가장 흔한 질환 중 하나이며, COVID-19 유행에 따라 전 세계적으로 관심이 증가되었다[1]. 호흡기 질환에 대한 관심의 증가에도 폐 호흡음은 비주기적이고 일정하지 않은 특징으로 분별하기 어렵다. 폐 호흡음 기반으로 정확한 질병 분류를 위해 딥러닝을 이용한 연구들이 진행되었다.

폐 호흡음을 이용한 AI 기반 딥러닝 모델은 오디오 데이터로부터 mel-spectrograms, Mel-Frequency Cepstral Coefficients (MFCC) 등과 같은 특징을 추출하여 입력 데이터로 사용한다. 딥러닝 모델은 오디오 데이터의 차별적 특징을 효과적으로 추출하고 패턴을 인식함으로써 높은 정확도를 보여준다. 그러나 이러한 딥러닝 모델은 컴퓨팅 리소스가 풍부한 서버 인프라가 필요하다[2]. 서버 인프라를 활용하기 위해 환자로부터 측정된 폐 호흡음은 네트워크를 통해 전달되어야 한다. 하지만 환자의 폐 호흡음은 개인정보로 분류되어 네트워크를 통한 데이터 송수신은 정보보호 문제가 발생할 수 있다.

정보보호 문제를 최소화하기 위해 온디바이스 AI가 주목받고 있다. 온디바이스 AI는 기기 내에서 데이터를 수집하고 처리한다. 따라서 데이터가 기기 외부로 전달되지 않는다. 이 구조에 의해 정보보호 문제가 완화된다. 그럼에도 여전히 온디바이스 AI는 딥러닝 모델을 기기에서 동작시키기 위한 과정이 필요하며, 높은 성능을 위해 적절한 모델 선택이 필요하다.

본 논문에서 온디바이스 AI를 위한 딥러닝 모델 성능을 비교한다. 딥 러닝 모델은 ResNet50, MobileNetV2, InceptionV2, 그리고 Stacked를 사용하였다[3]. 성능 평가를 위해 폐 호흡음 기반 질병 분류 모델을 생성하고 각 환경에 맞추어 모델을 최적화하는 테스트 베드를 구축하였다. 또한, 테스트 베드는 각 모델의 분류 정확도 및 추론 속도 계산 모듈을 통한 성능 지표를 보여준다

Ⅱ. 관련 연구

2.1 딥러닝 기반 호흡 질환 분류

폐 호흡음 분류에서 높은 성능을 달성하기 위해 다양한 딥러닝 기반 접근 방식이 연구되었다[4,5]. Basu는 폐 호흡음 녹음에서 MFCC를 추출하여 다섯 가지 호흡기 질환을 분류하였다[6]. Bardou은 오디오 파일에서 12개의 MFCC 계수를 계산하고 이러한 계수에서 6개의 통계적 특징을 추출하였다[7]. 또한, 스펙트로그램 시각화에서 로컬 이진 패턴 특징을 추출하고 CNN (Convolutional Neural Network) 기반 모델을 사용하여 높은 분류 정확도를 달성하였다.

Petmezas은 CNN과 Long Short-Term Memory (LSTM) 유닛을 결합한 하이브리드 아키텍처를 제안하였다[8]. 이 기법은 CNN을 사용하여 Short-Time Fourier Transform (STFT) spectrogram에서 특징을 추출한 다음 LSTM 모듈로 전달하여 시간 종속성을 포착하고 분류를 수행한다. Li는 Residual 블록 내에 증강된 주의 합성곱을 통합하여 폐 호흡음 분류 성능을 개선하는 아키텍처를 제안하였다[9]. 그들의 기법은 특징 추출을 위해 가변 Q-factor wavelet transform과 삼중 STFT를 활용하여 폐 호흡음의 특성을 효과적으로 표현할 수 있다.

2.1 딥러닝 기반 호흡 질환 분류

AI 기반 디지털 헬스케어 기술은 최근 큰 주목을 받으며 광범위한 연구가 진행되고 있다. AI의 기반이 되는 심층 신경망(DNN)은 대규모 데이터에서 복잡한 패턴을 학습하는 데 특히 효과적이다. 그러나 기존 DNN 모델은 일반적으로 효율적인 작동을 위해 연산 집약적인 서버 환경을 필요하다[2]. 이러한 요구 사항으로 인해 통신 인프라가 제한된 지역과 같이 리소스가 제한된 환경에서 서비스 제공에 대한 제한이 있다. 또한 네트워크를 통한 데이터의 전송은 개인정보의 문제가 발생할 수 있다. 이러한 한계를 해결하기 위해 의료 분야에서 온디바이스 AI 기술이 연구되고 있다. 온디바이스 AI는 웨어러블 기기, 임상 기기 또는 스마트 의료 장비에 통합된 임베디드 시스템에서 AI 추론 알고리즘을 직접 실행하는 것을 의미한다. 인터넷 연결을 통한 원격 서버에 의존하는 기존의 클라우드 기반 시스템과 달리, 온디바이스 AI는 실시간 추론, 향상된 데이터 보안, 운영 자율성, 중단 없는 서비스 연속성 등 여러 가지 이점을 제공한다.

온디바이스 AI는 디바이스에서 직접 실시간 추론을 수행하기 때문에 전력 효율성이 중요하다. 클라우드 서버 환경에서 일반적으로 사용되는 그래픽 처리 장치(GPU)는 범용 하드웨어 가속기로 기능하지만 상당한 전력이 필요하다. 이러한 한계를 해결하기 위해 신경망 처리 장치(NPU)가 온디바이스 AI를 위한 저전력 AI 가속기로 주목받고 있다[10]. NPU는 인간의 뇌에서 영감을 받아 구조적으로 설계되었으며, 수많은 상호 연결된 노드를 사용하여 정보를 효율적으로 처리한다. 인공지능 및 머신러닝 워크로드에 최적화되어 GPU보다 전력 소모가 훨씬 적고 기존 중앙처리장치(CPU)보다 작업에 따른 효율성이 높다. 온디바이스 AI 도입이 지속

적으로 확대됨에 따라, 엣지 디바이스에서 에너지 효율적인 고성능 추론을 지원하기 위해 NPU 기반 온디바이스 AI 연구가 주목받고 있다.

Ⅲ. 실 험

3.1 환경 구성

그림 1은 딥러닝 모델의 성능 평가를 위한 환경 구성을 보여준다. 폐 호흡음 기반 딥러닝 모델을 생성하기 위해 ICBHI 2017 및 KAUH 데이터셋을 사용하였다. ICBHI 2017 데이터셋은 920 개의 오디오 데이터를 포함하고 있으며, 오디오의 길이는 10초에서 90초로 구성되어 있다[11]. 이 데이터셋에 포함된 질병의 종류는 8가지(천식, 만성폐쇄성폐질환, 상기도염, 기관지확장증, 폐렴, 세기관지염, 하기도염, 건강상태)를 포함한다. KAUH 데이터셋은 308개의 데이터셋으로 이루어져 있으며, 오디오 길이는 5초에서 30초로 구성되어 있다. KAUH 데이터 셋에 포함된 질병의 종류는 7 가지(천식, 심부전, 폐렴, 기관지염, 흉막 삼출, 폐 섬유증, 만성폐쇄성폐질환)이 포함되어 있다.

본 논문에서 사용한 모델들은 11가지의 질병 분류를 수행한다. 질병 분류를 수행하기 위해 서로 다른 길이를 가지는 오디오를 5초 간격으로 나눈다. 예를 들어, 12초의 길이를 가지는 오디오의 경우 5초 길이의 오디오 2개로 나뉘며 2초의 길이는 버려진다. 이렇게 나뉜 오디오를 기반으로 부족한 데이터 수를 증가시키기 위해 기존 데이터에 시간 스트레칭, 피치 시프팅, 노이즈 삽입, 시간 시프팅, 그리고 볼륨 스케일링을 통해 데이터를 증강하였다. 입력 데이터는 오디오 데이터에서 mel-spectrogram, MFCC, 그리고 chroma를 추출한다. 추출된 데이터는 128x216 크기의 2D 이미지로 변환하고, 2D 이미지를 쌓아 사용하였다. 서버 환경은 NVDIA GeForce RTX 4090에서 모델을 실행하였으며, 온디바이스는 Raspberry Pi 5에서 Hailo 8 기반으로 모델을 실행하였다. Hailo 8은 딥러닝 모델 동작에 최적화된 Neural Process Unit (NPU) 모듈로 26 Tera Operations Per Second (TOPS)의 성능을 가진다.

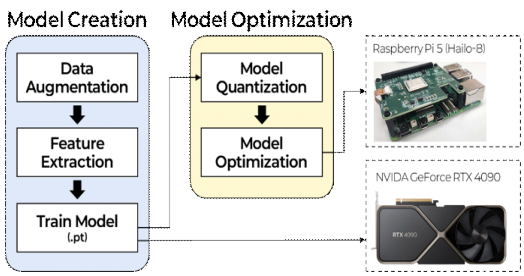


그림 1. 딥러닝 모델의 성능 평가를 위한 환경 구성
Fig. 1. Environment configuration for performance analysis of deep learning model.

온디바이스에서 모델을 동작시키기 위해 model optimization 과정을 수행한다. 기존 모델이 온디바이스에서 동작하기 위해 Float32 기반으로 학습된 파라미터를 UINT8로 양자화한다[12]. 모델 양자화 이후 Hailo 8에서 동작하기 위한 모델에 대한 정보를 생성한다.

이러한 테스트 환경은 모델의 성능을 보여주기 위해 분류 성능을 나타내는 정확도, F1-score, precision, 그리고 recall을 계산하여 보여준다. 또한, 온디바이스에서 모델의 효율성을 평가하기 위해 모델의 크기 및 추론 속도를 계산한다. 또한, 본 논문은 데이터 불균형으로 인한 잠재적 성능 편향을 완화하기 위해 10회 반복 학습에 따른 결과의 평균을 계산하였다. 실험 결과의 모든 지표는 매크로 평균 점수로 산출하였다.

3.2 서버 기반 딥러닝 모델 성능 평가

표 1은 서버 환경 기반 딥러닝 모델의 질병 분류 성능을 보여주며, 표 2는 모델 크기 및 추론 속도 성능을 보여준다. Residual 블록을 사용하는 ResNet50은 오디오의 시간-주파수 특징을 추출하고 많은 수의 레이어를 통해 특징을 학습하여 가장 높은 분류 성능을 보여준다. 하지만 ResNet50은 심층 구조로 인해 다른 모델들과 비교하여 크기가 가장 크다. MobileNetV2는 모바일 및

표 1. 서버 환경 기반 딥러닝 모델의 질병 분류 성능
Table 1. Disease classification performance based on server environment.

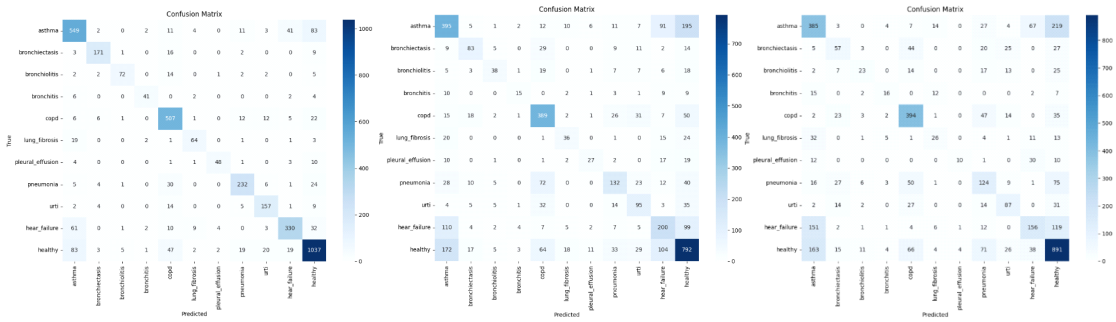
Model	Accuracy	F1-score	Precision	Recall
ResNet50	80.59%	0.798	0.801	0.759
MobileNetV2	55.43%	0.495	0.541	0.798
InceptionV2	55.48%	0.427	0.493	0.41
Stacked	50.16%	0.243	0.267	0.224
EfficientNet	60.44%	0.552	0.576	0.535

표 2. 모델 크기 및 추론 속도 성능 결과
Table 2. Performance result of model size and inference time.

Model	Parameter	Model size	Inference time
ResNet50	23.53 M	90 MB	18.5 ms
MobileNetV2	2.24 M	8.75 MB	11.2 ms
InceptionV2	5.98M	22.8 MB	12.8 ms
Stacked	0.36 K	1.4 MB	9.9 ms
EfficientNet	2.45M	9.55 MB	35 ms

임베디드 플랫폼을 위해 설계된 모델이다. MobileNetV2는 역잔차와 선형 병목 현상에 기반한 아키텍처를 사용한다. 이 모델은 감소된 파라미터 수와 모델 크기를 가지지만, 제한된 채널 용량으로 인해 복잡한 오디오 시나리오에서 ResNet50에 비해 분류 성능이 낮다. InceptionV2는 모델 복잡도를 줄이기 위해 인셉션 블록을 사용한다. 그러나 MobileNetV2보다 더 많은 매개변수와 더 큰 모델 크기를 가짐에도 불구하고, 심층 레이어에서 오디오 특징에 대한 인식 및 학습이 제대로 이루어지지 않아 낮은 성능을 보여준다. Stacked는 폐 호흡음 기반으로 질병 분류를 위한 경량화 모델이다. Stacked는 다른 모델들과 비교하여 파라미터의 수와 모델의 크기가 가장 작은 것을 보여준다. 하지만 Stacked는 노이즈가 포함된 오디오 데이터에서 특징을 인식 및 학습에 대해 효율적이지 않아 낮은 성능을 보여준다. EfficientNet은 다른 경량화 모델보다 더 높은 성능을 보이며, 이는 모델의 깊이, 너비, 해상도에 따라 확장하기 때문이다. 그러나 이 모델은 영상 내 특정 지점에 집중하는 특성을 가지므로 성능이 낮다.

그림 2는 각 모델의 confusion matrix를 보여준다. ResNet50은 건강한 사람, 만성 폐쇄성 폐질환, 폐렴, 천식과 같이 샘플 수가 많고 특징이 잘 정의된 클래스에서 높은 정확도를 보여준다. 그러나 상기관지염, 기관지염, 기관지확장증과 같이 증상이 겹치는 호흡기 질환에서 분류의 정확도가 낮다. 상기관지염과 만성 폐쇄성 폐질환 그리고 폐렴과 상기관지염의 오분류는 상기도 감염의 중첩된 음향 특징이 있음을 나타낸다. 특히, 폐섬유증과 천식은 호흡음 중 천명음이 부분적으로 중첩되어 오분류가 나타난다. MobileNetV2는 천식과 심부전에 대해 낮은 정확도를 보여준다. 천식과 심부전 간 오류는 심혈관/호흡기 신호의 경계 모호성으로 오분류가 나타난다. 또한, 데이터가 부족한 클래스는 건강상태 또는 심부전으로 분류된 것을 보여준다. InceptionV2는 심부전, 천식, 건강에 대해 오분류가 빈번하게 나타나며, 상기관지염, 폐렴, 기관지염 간에도 오분류가 발생한다. ResNet50과 MobileNetV2와 같이 심혈관계와 호흡기계의 음향 신호 특성에 대한 경계를 인지하지 못하여 오분류가 발생하며, 질환 간 음향적 특징 중첩에 의한 성능 저하를 보여준다. Stacked는 천식과 건강, 심부전과 천식, 폐렴과 건강에 대해 정확하게 분류하지 못한다. EfficientNet은 건강상태 클래스뿐만 아니라 만성폐쇄성폐질환 그리고 천식과 같이 데이터가 충 분하고 음



(a) ResNet50 (b) MobileNetV2 (c) InceptionV2

(d) Stacked (e) EfficientNet

그림 2. 서버 환경 기반 딥러닝 모델의 혼동 행렬
Fig. 2. Confusion matrix of deep learning model based on server environment.

표 3. GPU 환경에서 각 기법의 p-value
Table 3. P-value of each scheme in GPU environment.

Model	P-value
ResNet50	1.80×10^{-70}
MobileNetV2	1.69×10^{-5}
InceptionV2	1.81×10^{-130}
Stacked	2.78×10^{-106}
EfficientNet	4.85×10^{-81}

향적으로 뚜렷하게 구분되는 클래스에 대해 높은 정확도를 보여준다. 반면, 천식, 심부전, URTI 간에는 증상과 관련된 음향 특징이 유사하여 오분류가 높은 것을 보여준다. 또한, 기관 지염, 흉막 삼출, 그리고 폐 섬유증은 학습에 사용되는 데이터의 수 부족으로 이 클래스들에 대한 분류 정확도가 낮게 나타난다.

표 3은 GPU 환경에서 각 기법의 p-value를 나타낸다. 각 기법의 p-value는 0.001보다 낮아 통계적으로 유의미하며, 이것은 모델의 성능이 우연에 의해 발생했을 가능성이 매우 낮음을 의미한다.

3.3 온디바이스 기반 딥러닝 모델 성능 평가

GPU 기반으로 동작하는 모델을 온디바이스에서 동작시키기 위해 모델을 onnx 포맷으로 변환하여 UINT8로 양자화한다. 양자화된 모델을 Hailo-8에서 동작시키기 위해 hef 포맷으로 변환한다. 모델을 hef 포맷으로 변환하는 과정은 python 기반으로 설계된 모델을 온디바이스에서 동작시키기 위해 C 기반 모델로 변환하는 과정을 포함한다. 이 과정에서 모델의 레이어 제거 및 수정에 따른 변화가 발생하지 않았다. 표 4는 onnx 포맷 변환에 따른 각 기법의 성능을 나타낸다. 모든 기법은 GPU 기반 평가와 유사한 성능을 보였으나, 다소간의 성능 저하가 관찰되었다. 이는 기준 모델과 ONNX 모델 모두 FLOAT32 정밀도로 동작하지만, 변환 과정에서 매개변수 스케일링에 따른 미세한 변동이 발생하여 ONNX 모델이 세밀한 표현에 덜 민감하게 되기 때문이다.

표 5는 온디바이스 환경 기반 딥러닝 모델의 질병 분류 성능을 보여주며, 표 6은 온디바이스에서의 모델 크기 및 추론 속도 성능 결과를 보여준다. EfficientNet은 양자화 이후 분류 성능이 저하되었는데, 이는 모델이 입력 영상의 제한된 영역에 집중하는 경향을 보여 오디오 특징 인식의 전체적인 정확도가 감소하기 때문이다. InceptionV2와 Stacked 모델 역시 8비트 양자화후 성능 저하를 보였다. 구체적으로, InceptionV2는 GPU 기반 성능 대비 NPU 평가에서 정확도가 6.47% 감소하였으며, Stacked 모델은 동일 조건에서 2.29% 감소하였다. 흥미롭게도, 정확도는 낮아졌음에도 불구하고 두 모델 모두 F1-score, precision, recall에서는 향상을 보였다. 이러한 결과는 데이터셋에 존재하는 클래스 불균형 때문으로 해석될 수 있다. 반면, ResNet50와 MobileNetV2는 GPU 대비 NPU에서 전반적인 성능 향상을 보였다. GPU 기반 학습 환경에서 부동소수점 모델은 데이터가 많은 클래스에 과도하게 학습되어 편향된 예측을 초래하여 성능이 저하될 수 있다. 그러나 이러한 모델을 정수 기반 NPU에 양자화하여 배포할 경우, 수치적 정밀도의 감소가 정규화 역할을 수행하여 과적학을 완화하여 성능을 향상시킨다. 양자화는 파라미터의 범위를 압축하여 질병 분류에 대한 영향도를 감소시켜 모델의 민감도를 낮춘다. 이것은 클래스별 편향을 완화하여 모든 클래스에 대해 균형 잡힌 추론을 보여준다.

표 4. ONNX 포맷 기반 딥러닝 모델의 질병 분류 성능
Table 4. Disease classification performance based on ONNX format.

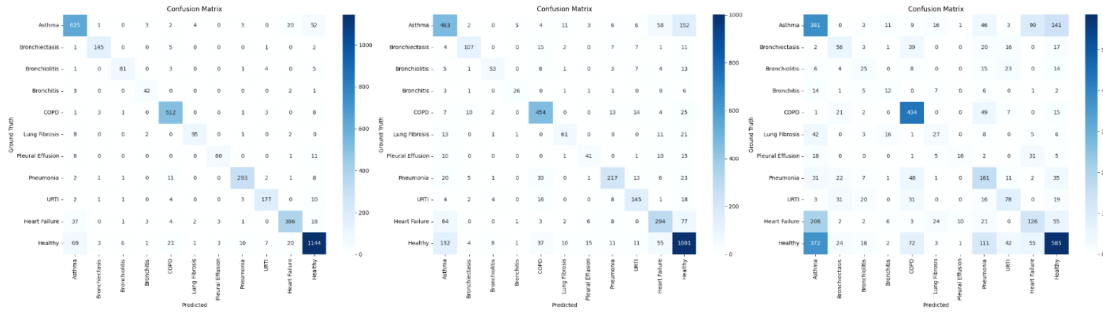
Model	Accuracy	F1-score	Precision	Recall
ResNet50	80.42%	0.781	0.788	0.78
MobileNetV2	55.24%	0.484	0.522	0.397
InceptionV2	54.36%	0.415	0.477	0.392
Stacked	50.02%	0.228	0.256	0.204
EfficientNet	60.26%	0.534	0.562	0.523

표 5. 온디바이스 환경 기반 딥러닝 모델의 질병 분류 성능
Table 5. Disease classification performance based on on-device environment.

Model	Accuracy	F1-score	Precision	Recall
ResNet50	89.41%	0.89	0.88	0.87
MobileNetV2	71.89%	0.66	0.73	0.66
InceptionV2	47.74%	0.42	0.41	0.36
Stacked	55.95%	0.28	0.24	0.32
EfficientNet	40.63%	0.273	0.45	0.265

표 6. 온디바이스에서 모델 크기 및 추론 속도 성능 결과
Table 6. Performance result of model size and inference time in on-device

Model	Model size	Inference time
ResNet50	37.6 MB	245.9 ms
MobileNetV2	3.8 MB	169.77 ms
InceptionV2	6.47 MB	150.27 ms
Stacked	835 KB	257.74 ms
EfficientNet	3.85 MB	166.84 ms



(a) ResNet50 (b) MobileNetV2 (c) InceptionV2

(d) Stacked (e) EfficientNet

그림 3. 온디바이스 환경 기반 딥러닝 모델의 혼동 행렬
Fig. 3. Confusion matrix of deep learning model based on on-device environment.

그림 3은 온디바이스 환경 기반 딥러닝 모델의 혼동 행렬을 나타낸다. 모든 모델은 서버 환경과 비교하여 오분류가 나타나는 클래스가 같게 나타나는 것을 확인하였다. 하지만 InceptionNetV2는 서버 환경과 비교하여 천식과 만성 폐쇄성 폐질환에 대한 분류 성능이 높지만, 심부 전과 건 강 클래스에 대한 분류 정확도가 낮다.

표 7은 NPU 환경에서 각 기법의 p-value를 나타낸다. 각 기법의 p-value는 양자화를 진행하기이전과 다르지만 UNIT8 기반으로 모델이 양자 화되었을 때 p-value가 0.05보다 낮으므로 우연에 의한 성능변화가 아닌 것을 보여준다.

표 7. NPU 환경에서 각 기법의 p-value
Table 7. P-value of each scheme in NPU environment.

Model	P-value
ResNet50	0.04
MobileNetV2	3.11×10^{-99}
InceptionV2	0
Stacked	1.65×10^{-277}
EfficientNet	0

Ⅳ. 결 론

본 연구에서는 폐 호흡음을 기반으로 질병 분류를 위한 AI 모델을 서버 환경과 온디바이스 환경에서 각각 실행하고, 그 성능을 비교 및 분석하였다. 실험 결과, 온디바이스 환경에서 서버 대비 딥러닝 모델의 일부 성능 저하가 발생하였으나, 실시간 처리가 가능한 것을 확인하였다. 이러한 결과는 제한된 자원에서도 실용적인 AI 기반 질병 탐지 시스템을 구현할 수 있는 가능성을 보여준다. 또한 모델의 크기 및 정확도에 따른 상관관계성을 가져 목적에 따라 적절한 모델 선택이 필요하다.

향후 연구에서는 온디바이스 환경에 특화된 경량 모델을 직접 설계하고 제안함으로써 성능과 처리 속도의 균형을 더욱 정교하게 맞추는 방향으로 확장하고자 한다. 또한, 다양한 디바이스 환경에서의 일반화 성능을 추가적으로 검증할 예정이다.

References

[1] J. B. Soriano, P. J. Kendrick, K. R. Paulson,

V. Gupta, E. M. Abrams, R. A. Adedoyin, T. B. Adhikari, S. M. Advani, A. Agrawal, and E. Ahmadian, "Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: A systematic analysis for the global burden of disease study 2017," *The Lancet Respiratory Med.*, vol. 8, no. 6, pp. 585-596, Jun. 2020.

[2] X. Wang, Z. Tang, J. Guo, T. Meng, C. Wang, T. Wang, and W. Jia, "Empowering edge intelligence: A comprehensive survey on on-device AI models," *ACM Computing Surv.*, vol. 57, no. 9, pp. 1-39, Apr. 2025. (https://doi.org/10.1145/3724420)

[3] T. Wanasinghe, S. Bandara, S. Madusanka, D. Meedeniya, M. Bandara, and L. D. L. T. Diez, "Lung sound classification with multi-feature integration utilizing lightweight CNN model," *IEEE Access*, vol. 12, pp. 21262-21276, Feb. 2024. (https://doi.org/10.1109/ACCESS.2024.3361943)

[4] E. Messner, M. Fediuk, P. Swatek, S. Scheidl, F.-M. Smolle-Juttner, H. Olschewski, and F. Pernkopf, "Multi-channel lung sound classification with convolutional recurrent neural networks," *Computers in Biology and Med.*, vol. 122, p. 103831, Jul. 2020. (https://doi.org/10.1016/j.compbiomed.2020.103831)

[5] K. K. Lella and A. Pja, "Automatic diagnosis of COVID-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: Cough, voice, and breath," *Alexandria Eng. J.*, vol. 61, no. 2, pp. 1319-1334, Feb. 2022. (https://doi.org/10.1016/j.aej.2021.06.024)

[6] V. Basu and S. Rana, "Respiratory diseases recognition through respiratory sound with the help of deep neural network," *Int. Conf. CINE*, pp. 1-6, Apr. 2020. (https://doi.org/10.1109/CINE48825.2020.234388)

[7] D. Bardou and K. Zhang, "Lung sounds classification using convolutional neural networks," *Artificial Intelligence in Med.*, vol. 88, pp. 58-69, Jun. 2018. (https://doi.org/10.1016/j.artmed.2018.04.008)

[8] G. Petmezs, G.-A. Cheimariotis, L. Stefanopoulos, B. Rocha, R. P. Paiva, A. K. Katsaggelos, and N. Maglaveras, "Automated lung sound classification using a hybrid CNN-LSTM network and focal loss function," *Sensors*, vol. 22, no. 3, pp. 1-13, Jan. 2022. (https://doi.org/10.3390/s22031232)

[9] S. B. Shuvo, S. N. Ali, S. I. Swapnil, T. Hasan, and M. I. H. Bhuiyan, "A lightweight CNN model for detecting respiratory diseases from lung auscultation sounds using EMD-CWT-Based hybrid scalogram," *IEEE J. Biomed. and Health Inf.*, vol. 25, no. 7, pp. 2595-2603, Dec. 2020. (https://doi.org/10.1109/JBHI.2020.3048006)

[10] L. Xun, J. Hare, and G. V. Merrett, "Dynamic DNNs and runtime management for efficient inference on mobile/embedded devices," *arXiv preprint arXiv:2041.08965*, Jan. 2024. (https://doi.org/10.48550/arXiv.2401.08965)

[11] *ICBHI 2017 Challenge*, [Online]. Available: https://bhichallenge.med.auth.gr/ICBHI_2017_Challenge

[12] S. Lee, S. Jeong, and I. Anh, "Trends in lightweighting and optimizing on-device AI models for carbon reduction," *J. KICS*, vol. 42, no. 5, pp. 22-28, Apr. 2025.

박 진 호 (Jinho Park)

2018년 2월 : 광운대학교 전자통신공학과 졸업
2023년 8월 : 광운대학교 전자통신공학과 박사
2023년 8월~현재 : 한국전자기술연구원 재직
<관심분야> 전자공학, 통신공학, 의용공학, 컴퓨터공학
[ORCID:0000-0002-1742-6839]

홍 혁 기 (Hyuck Ki Hong)

2001년 8월 : 단국대학교 화학과 졸업
2003년 8월 : 단국대학교 분석화학 석사
2003년 8월~현재 : 한국전자기술연구원 재직
<관심분야> 전자공학, 의용공학