creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사 학위논문

Vision Transformer 기반 심음 분류 연구

2023년 2월

서울대학교 융합과학기술대학원

지능정보융합학과

김준엽

Vision Transformer 기반 심음 분류 연구

지도 교수 서봉원

이 논문을 공학석사 학위논문으로 제출함
2022년 12월

서울대학교 융합과학기술대학원
지능정보융합학과
김준엽

김준엽의 공학석사 학위논문을 인준함
2022년 12월

위 원 장 _____ 이교구 _____ (인)

부위원장 _____ 서봉원 _____ (인)

위 원 _____ 이원종 _____ (인)

초 록

심장 질환의 진단은 중요한 의학적 정보인 심장의 상태 및 기능에 대한 정보를 제공하는 심음(heart sound)을 측정하는 것으로부터 시작된다. 청진기를 통해 심음을 듣고 진단하는 것은 정밀검사에 비해 정확도가 떨어지지만, 용이하고 비용이 거의 들지 않기 때문에 진단에 있어 필수적으로 사용된다. 건강한 사람에게서는 규칙적이고 명확한 심장 박동 소리가 들리지만, 그렇지 않은 사람에게서는 심장 소리와 함께 잡음이 함께 들리기도 한다. 이러한 잡음을 심잡음(heart murmur)이라고 하며, 심잡음의 특징과 잡음이 들리는 위치 등을 사용하여 심장병을 판단할 수 있다. 심음 데이터를 녹음해서 만들어진 심음도(PCG; Phonocardiograms) 데이터로 이 사람이 심장병 환자인지 유무를 탐지하거나, 심장 소리에 이상이 있는지 등을 탐지할 수 있다. 심음에서 비정상적인 심장 기능을 감지하기 위한 자동화된 모델을 개발하는 것은 개발도상국과 같이 전문가와 자본이 부족한 나라의 심장병으로 고통받는 사람들에게 중요한 연구 주제이다.

의료데이터를 모델링 하는 방법에는 고전적인 기계학습(Machine Learning) 기반의 모델을 활용하는 방법과, 딥러닝(Deep Learning) 기반의 모델을 활용하는 방법으로 나눌 수 있다. 기계학습 기반의 의료데이터 모델링은 특성(feature)을 직접 추출해야 하므로, 사용하는 데이터에 대해 연구자의 사전지식과 전처리(pre-processing) 방법들이

큰 영향을 미친다. 반면 딥러닝 기반의 모델은 이러한 특성까지도 모델이 직접 학습하여 추출하기 때문에, 사용하는 데이터에 대한 연구자의 사전지식과 전처리 방법의 영향이 비교적 낮다. 전통적으로 의료데이터 모델링에는 기계학습 모델들이 많이 사용되었으나, 최근 의료데이터 분야에서는 딥러닝 분야에서 좋은 성능을 보이는 모델을 사용해서 기존 의료데이터 모델링에서 성능을 높이는 연구들이 진행되고 있으며, 고전적인 기계학습들보다 더 좋은 성능을 보인다. 하지만 딥러닝 기반의 모델들은 대부분 결과에 대한 해석을 할 수 없어, 전문가의 진단에 도움을 주는 것에 어려움을 겪고 있다.

본 논문에서 사용되는 Vision Transformer 구조의 경우에는 셀프 어텐션(self-attention)이 포함되어 있으며, 이는 결과를 이해하는 데 도움이 되는 어텐션 점수(attention score)를 계산하여 이를 어텐션 마스크(attention mask)로 시각화 할 수 있다는 장점과, 컴퓨터 비전(Computer Vision)분야의 이미지 분류(Image Classification) 태스크(task)에서 높은 성능을 보이는 장점을 가진 모델이다.

본 논문에서는 여러 청진 위치에서 측정된 심장 소리 녹음에서 잡음의 유무와 임상 결과를 잘 감지하는 모델을 제안한다. 보다 높은 분류 성능과 결과의 해석에 도움을 주기 위해 시각적인 접근 방식을 사용했다. 제안된 모델은 신호의 리샘플링(resampling)이나 필터링(filtering) 없이 심장 소리 신호를 스펙트로그램(spectrogram)으로 변환하고, 변환된 이미지를 환자의 인구통계학적 정보와 함께 입력 받아 심잡음과 임상 결과를

추론(inference)한다. 임상 결과 식별 작업의 경우 테스트 데이터에 대한 대회 비용 함수 점수 11943을, 심잡음 분류에 대해서는 0.69의 가중치 정확도를 기록했다. 또한 모델에는 어느 부분을 보고 판단했는지 시각화가 가능한 어텐션 마스크가 포함되어 있다.

주요어 : 심음도, 심잡음, Vision Transformer, 스펙트로그램, 어텐션 마스크
학 번 : 2021-29046

목 차

제 1 장 서 론 1

제 1 절 연구의 배경 1

제 2 절 연구의 내용 4

제 2 장 선행 연구 7

제 1 절 심장 소리 분류 7

제 2 절 소리의 시각화 12

제 3 절 Vision Transformer 16

제 4 절 데이터 증강 22

제 3 장 연구 방법 및 설계 28

제 1 절 데이터셋 28

제 2 절 전처리 및 이미지 데이터셋 생성 32

제 3 절 모델 학습 37

제 4 장 실험 및 결과 41

제 1 절 평가 지표 41

제 2 절 훈련 데이터에 대한 절제 연구 43

제 3 절 대회 데이터셋에 대한 실험 결과 47

제 5 장 논의 50

제 1 절 대회 결과에 대한 고찰 50

제 2 절 어텐션 마스크 52

제 3 절 향후 계획 56

제 4 절 한계 58

제 6 장 결론 59

참고문헌 61

Abstract 68

그림 목차

[그림 1] PCG와 ECG 신호 3

[그림 2] 일반적인 자동 심장 소리 분류 메커니즘 5

[그림 3] 기계학습 기반 모델의 모델링 과정 9

[그림 4] 딥러닝 기반 모델의 모델링 과정 11

[그림 5] 국소 푸리에 변환 과정 12

[그림 6] 심장 소리 스펙트로그램 15

[그림 7] 어텐션 메커니즘 17

[그림 8] 트랜스포머 구조 19

[그림 9] ViT 구조 20

[그림 10] 어텐션 지도 예시 21

[그림 11] 이미지 데이터 증강 기법 종류 23

[그림 12] SpecAugment 예시 27

[그림 13] ViT 구조를 통한 심장 소리 분류 메커니즘 28

[그림 14] 의사 결정 프로세스 및 비용 29

[그림 15] 심장 소리 데이터 및 인구통계학적 정보 변환 과정 33

[그림 16] 스펙트로그램 이미지 변환 과정 34

[그림 17] 데이터 증강 기법을 적용한 스펙트로그램 35

[그림 18] 본 논문에서 제안하는 ViT 모델 38

[그림 19] 심잡음이 없는 스펙트로그램 및 어텐션 시각화 52

[그림 20] 심잡음 유무를 판단할 수 없는 스펙트로그램 및 어텐션 52

[그림 21] 심잡음 스펙트로그램 및 어텐션 시각화 53

[그림 22] 심장병이 없는 사람의 스펙트로그램 및 어텐션 시각화 53

[그림 23] 심장병이 있는 사람의 스펙트로그램 및 어텐션 시각화 53

[그림 24] 전문가에게 모델의 결과를 해석하는데 도움을 주는 방법 제안 54

수식 목차

[수식 1] 이산 푸리에 변환 계산.....13

[수식 2] 국소 푸리에 변환 계산.....13

[수식 3] 가중 범주형 교차 엔트로피 계산.....39

[수식 4] 대회 비용 함수 계산.....42

[수식 5] 대회 가중된 정확도 계산.....42

표 목차

[표 1] PhysioNet 2022 데이터셋에서 사용하는 변수.....30

[표 2] PhysioNet 2022 데이터셋의 클래스별 개수.....20

[표 3] PhysioNet 2022 데이터셋 범주형 데이터 특성 요약.....36

[표 4] PhysioNet 2022 데이터셋 수치형 데이터 특성 요약.....36

[표 5] 심장병 환자 분류 모델 혼동 행렬.....41

[표 6] 심잡음 검출 모델 혼동 행렬.....42

[표 7] PhysioNet challenge 2022 심장병 환자 분류 공식 점수.....47

[표 8] PhysioNet challenge 2022 심음 분류 공식 점수.....47

[표 9] 심장병 환자 분류 모델에 대해 제출된 최종 모델의 학습, 검증, 평가 데이터에 대한 결과.....47

[표 10] 심잡음 분류 모델에 대해 제출된 최종 모델의 학습, 검증, 평가 데이터에 대한 결과.....47

그래프 목차

[그래프 1] 심장병 환자 검출 절제 연구.....44

[그래프 2] 심집음 검출 절제 연구.....44

제 1 장 서 론

제 1 절 연구의 배경

심음은 심장 박동에서 판막을 통과하는 혈류 및 심장이 만드는 소리로, 청진기를 통해 측정할 수 있다. 심음은 중요한 의학적 정보인 심장의 상태 및 기능에 대한 여러 정보를 제공한다. 심장 질환[①]의 진단은 보통 심음을 측정하는 것으로부터 시작되지만, 진단하는 전문가별 청각특성에 따른 편향(bias)이 생기면서 정밀검사에 비해 정확도가 떨어진다. 하지만 진단이 용이하고 비용이 거의 들지 않기 때문에 진단에 있어 필수적으로 사용된다. [1]

건강한 사람에게서는 규칙적이고 명확한 심장 박동 소리가 들리지만, 그렇지 않은 사람에게서는 심장 소리와 함께 잡음이 함께 들리기도 한다. 이러한 잡음을 심잡음이라고 하며, 심장 판막의 협착과 불완전한 폐색, 심장의 정상적이지 못한 통로의 생성 등으로 인한 혈류의 방해로 생긴다. 심잡음 자체가 병은 아니지만, 심장에 이상이 있을 확률이 높아진다. 이에 심잡음이 뚜렷하게 들리는 경우에는 더 정밀한 검사가 필요하다. 심잡음의 특징과 잡음이 들리는 위치 등 심장 소리에서 얻을 수 있는 정보를 분석하면, 심장병이 있는지에 대해 판단할 수 있다. [2]

세계보건기구(WHO)의 조사에 의하면 [3], 2019년에 허혈성 심장병으로 인해 사망한 사람은 890만명으로, 심혈관병은 전 세계

---

① 부정맥, 심근증, 고혈압 등 심장과 관련된 모든 질환을 포함한다.

사망원인의 1순위로 뽑힌다. 일반적으로 심장 검사에 사용되는 방법은 심전도(ECG; electrocardiogram), CT(Computed Tomography)등의 정밀검사가 있다. 하지만 이러한 정밀검사는 비용과 시간 많이 들며, 이 결과를 해석하는 전문가가 필수적이다. 또한 측정 장비의 비용이 비싸며, 전문가가 필요하기 때문에 개발도상국 등 많은 나라에서 정밀검사를 진행하는 것에 어려움이 있다. 반면 청진기를 통해 심장 소리를 측정하는 방식은 간단하고, 비용이 거의 들어가지 않는다. 다만, 심장 소리를 듣고 의사가 진단할 때, 사람에게 병이 있는지를 판단하기에는 정확도가 떨어진다. 이는 심장 소리는 측정 시 소음이 함께 기록되기에, 측정하는 환경에 따라 영향을 많이 받기 때문이다. 또한, 심혈관 질환에는 협심증, 심근경색, 동맥경화증, 심장판막증, 심부전, 고혈압, 심실세동 등 다양한 질병이 포함되며, 이러한 증상들의 다양함이 소음과 함께 분석에 어려움을 야기한다. 그러므로 전문가가 심장 소리를 청진하는 것 만으로 진단을 내리는 것에는 어려움이 있다. 이에 컴퓨터를 통해 심장 소리로 심장병 환자를 분류하는 것은 개발도상국 뿐만 아니라, 심혈관질환으로 고통받을 수 있는 사람들에게도 편의성을 가져다 줄 수 있는 중요한 연구문제이다[4,5].

심전도나 CT같은 정밀검사의 경우에도 전문가가 데이터를 통해 사람에게 병이 있는지를 판단한다. 이러한 탐지에 인공지능 기반의 모델을 활용한다면, 더 높은 정확도와 빠른 속도 그리고 비용을 절감할 수 있다. 실제로 딥러닝 및 기계학습 모델을 사용하면 의료 데이터 기반의 질병에 대한 모델링을 할 수 있으며, 정밀검사를 통한 진단에

비해서는 정확도가 낮지만 심음 데이터를 녹음해서 만들어진 심음도 데이터로도 환자의 심장병 유무를 탐지하거나, 심장 소리에 이상이 있는지 심음을 탐지할 수 있다. 따라서 심장 소리의 PCG로 비정상적인 심장 기능을 감지하기 위한 자동화된 모델을 개발하는 것이 심장 상태를 진단하고 치료하는 데 도움이 될 것이다.

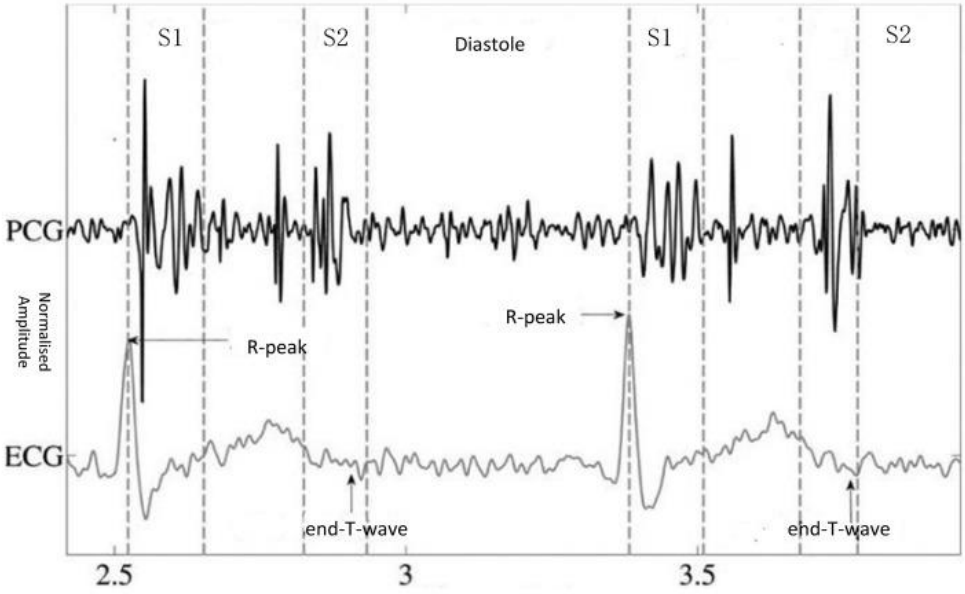


그림 1. PCG와 ECG 신호[6]

제 2 절 연구의 내용

본 논문은 의료데이터 중 PCG에서 심장병 환자 및 심잡음을 탐지하는 모델에 대해 연구한다. 컴퓨터를 통해 의료데이터를 모델링하는 방법에는 크게 고전적인 기계학습 기반의 방법과, 딥러닝 기반의 방법으로 나눌 수 있다. 기계학습 기반의 의료데이터 모델링 방법에는 랜덤 포레스트(Random Forest), 결정 트리(Decision Tree), 서포트 벡터 머신(Support Vector Machine), K-최소 근접(K-Nearest Neighbor), 익스트림 그라디언트 부스팅(XGBoost) 등의 기법들이 있다. 이러한 모델을 사용하기 위해서는 연구자가 직접 특성을 추출하고 선택해야한다. 이에 기계학습 기반 방법은 해당 의료데이터에 대한 사전지식이 요구되며, 이에 따른 적절한 전처리 방법들이 성능에 큰 영향을 미친다. 반면 CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), 트랜스포머(Transformer)[6]와 같은 딥러닝 모델 기반 방법은 이러한 특성까지도 모델이 직접 추출하고 선택하기에, 연구자의 의료데이터에 대한 사전지식과 전처리 방법의 영향이 비교적 낮다. 전통적으로 의료데이터 모델링에는 기계학습 모델들이 많이 사용되어왔으나, 최근에는 딥러닝 모델을 사용해서 성능을 높이는 연구들이 진행되고 있으며, 이러한 방법들은 고전적인 기계학습보다 더 좋은 성능을 보인다. 하지만 딥러닝 기반의 모델들은 대부분 결과에 대한 해석을 할 수 없어, 전문가의 진단에 도움을 주는 것에 어려움을 겪고 있다[7-9].

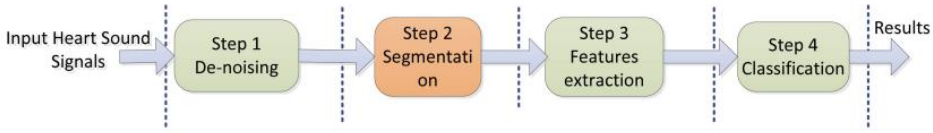


그림2. 일반적인 자동 심장 소리 분류 메커니즘[7]

최근 딥러닝 연구에서는 Transformer[10] 구조를 다양한 분야에 적용하면서 좋은 성능들을 보여주며, 이 예로 자연어 처리 분야의 BERT[11]가 있다. 컴퓨터 비전 분야에서도 역시 Transformer 구조를 기반으로 하는 이미지 분류 모델인 ViT(Vision Transformer)[12]가 등장하였으며, 이미지 분류 분야에서 CNN 기반 모델들보다 더 좋은 성능을 보인다. 또한 Transformer 모델을 적용한 경우, 셀프 어텐션(self-attention) 점수 계산을 통해 결과를 이해하는 데 도움이 되는 시각화인 어텐션 지도(attention map)와 어텐션 마스크를 생성할 수 있다는 장점이 있다. ViT는 이미지를 여러개의 패치(patch)로 분할하고 임베딩(embedding)하는 방식을 통해 우수한 성능을 보여주었다. 그 외에도 다양한 방식의 모델이 의료데이터 분석에 사용되었지만, 아직까지 PCG에 ViT를 적용해 분류한 연구는 없다.

이러한 아이디어에 착안하여, 본 논문에서는 PCG 데이터를 통해 심장병 유무로 정상(normal)과 비정상(abnormal)을 판단하는 모델과 전문가가 판단한 심잡음의 유무 및 잘 모르겠다고 응답한 총 세 가지 없음(absent), 있음(present), 알수없음(unknown)으로 분류하는 모델 총 두 가지 모델을 연구한다. PCG를 ViT 구조로 모델링하기 위해

1)겹치는 부분 없이 신호를 분할하고, 2)분할된 신호를 스펙트로그램으로 변환하며, 3)데이터를 증강하고, 4)인구통계학적 특성 데이터를 사용하기 위한 전처리의 진행과정을 거쳤다. 또한 모델의 셀프 어텐션 부분을 통해 모델의 결과를 이해하는 데 도움이 될 수 있는 어텐션 마스크를 출력한다. 본 논문에서 제안하는 모델은 Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022에서 39팀중 6등을 기록하였으며, 이에 시각적인 접근 방식을 통해서도 PCG에 기록된 심장 잡음과 환자를 더 높은 정확도로 구별할 수 있을 뿐만 아니라 스펙트로그램에서 모델의 어텐션을 시각화 할 수 있음을 보였다. 본 연구의 기여 부분을 정리하자면 다음과 같다.

- 처음으로 ViT 기반 심음을 분류할 수 있는 전략을 제시함
- 시각적 접근을 통한 좋은 성능의 심잡음 검출 및 환자를 검출함

제 2 장 선행 연구

본 연구의 내용은 Vision Transformer 모델을 사용해서 심잡음 및 심장병 환자를 탐지하는 분류모델을 만드는 것이다. 이에, 이번 장에서는 본 연구의 선행연구로 첫째로 심장 소리를 분류하는 연구에서 사용되는 기법을 소개한다. 다음으로 본 논문의 기반 모델인 ViT에 사용할 수 있는 이미지를 만들기 위한 소리의 시각화와 관련된 방법과 모델을 소개한다. 마지막으로는 데이터 증강 기법에 대해 소개한다.

제 1 절 심장 소리 분류 연구

컴퓨터를 사용해 자동으로 병을 탐지하는 것은 빠르고 경제적이며, 효율적인 방법이다. 현재까지 다양한 의료데이터 분야에서 컴퓨터를 사용한 질병 탐지는 전문가보다 높은 정확도를 보인다. 심음도 에서도 주요 특성들을 추출한 뒤 이를 환자의 데이터와 함께 비교함으로써 직관적인 판단이 가능하며, 임상 지식을 갖춘 전문가가 잠재적인 심혈관 질환을 추가로 추론할 수 있다. 컴퓨터를 통한 자동 병 탐지 모델을 사용할 경우 원본 소리 신호를 그대로 사용할 수도 있지만, 심장 소리의 주요 대역폭인 10~400Hz로 대역 필터링(bandpass filtering)하는 전략[13] 등 다양한 방법으로 신호를 처리한 뒤 사용하는 방법도 있다. 소리 신호를 그대로 입력할 수도 있지만, 스펙트로그램으로 변환하여 입력하거나 이미지로 변환하여 분류 모델을 사용하는 방법 등 다양한

방법의 모델링이 가능하다. 컴퓨터를 사용한 의료 데이터의 모델링 방법은 기계학습 기반의 방법과 딥러닝 기반의 방법으로 나눌 수 있다. 이번 절에서는 각 모델링 방법에서의 일반적인 모델링 과정과 어떠한 방법들을 사용할 수 있는지에 대해 소개하고자 한다.

기계학습 기반 심장 소리 분류

기계학습 기반의 모델을 통해 심음에서 심혈관 질환 여부를 판단하는 과정은 일반적으로 전처리, 특성 추출, 특성 선택, 분류 및 추론의 과정을 거친다. 데이터 전처리 과정에서는 심장 소리가 아닌 소리들을 걸러내는 작업과 함께, 사용되는 모델에 알맞은 특성을 추출하기 위한 작업으로 주로 노이즈 필터링(noise filtering), 리샘플링(resampling), 분할(segmentation)등의 작업을 선택적으로 거친다. 이후 전문가에 의해 필요한 특성들이 추출 및 선택되며, 모델에 입력되어 심장병이 있는지 없는지를 추론하는 되는 과정을 거쳐 환자를 탐지한다.

기존 연구들은 공개된 데이터셋인 PhysioNet2016[14] 데이터를 이용해 심장병 환자를 분류하는 모델을 연구했다. 이 데이터는 5초에서 120초 이상 지속되는 총 3126개의 심장 소리로 구성되며, 2000Hz로 리샘플링한 심음 데이터셋을 제공하며, 대회의 민감도와 특이도의 산술평균을 대회 점수로 정의하고 평가한다.

기계학습 기반의 모델을 사용해 PhysioNet 2016 데이터셋 심음을

분류한 선행연구 몇 개를 소개하면 다음과 같다. 심장 소리를 스펙트로그램으로 변환하고, 행렬 분해를 사용해 특징을 추출하고, 3000개의 트리(Tree)를 갖는 랜덤 포레스트(Random Forest) 모델을 통해 0.83의 특이도와 0.92의 민감도 및 0.88의 점수로 분류한 연구[15], 심음에서 신호의 웨이블릿 분해 계수를 추출하여 서포트 벡터 머신(Support Vector Machine) 기반의 트윈 서포트 벡터 머신(Twin Support Vector Machine) 모델을 사용하여 0.9의 정확도와 0.946의 민감도 및 0.855의 특이도로 분류한 연구[16], 심음 신호를 분할하지 않고 XGBoost를 적용하여 0.929의 정확도 및 0.945의 민감도와 0.913의 특이도로 분류한 연구[17] 등 다양한 모델 기반의 심음 분류 연구들이 진행되어왔다.

이러한 기계학습 기반의 모델을 사용할 경우, 사람이 소리를 듣고 판단하는 것 보다 높은 성능으로 판별이 가능하다. 하지만, 이러한 성능을 내기 위해서는 심음을 전처리 하는 과정이 필수적이며, 전문가의 도움을 필요로 한다. 연구별로 처리하는 방식과 사용하는 특성이 상이하기에 딥러닝 모델보다 번거로운 과정을 거치게 되지만, 최근 성능이 딥러닝 기반 모델들에 미치지 못해 효율이 떨어진다.

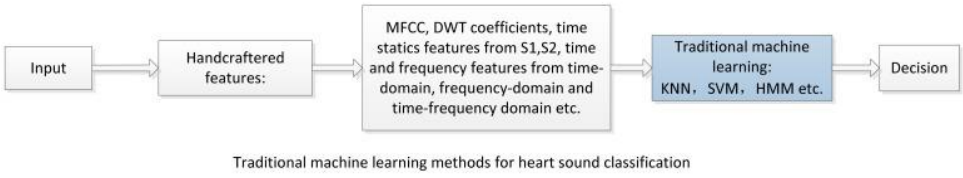


그림 3. 기계학습 기반 모델의 모델링 과정[7]

딥러닝 기반 심장 소리 분류

딥러닝 기반의 모델을 통해 심음에서 심혈관 질환 여부를 판단하는 과정 역시 기계학습 기반의 방법과 비슷한 과정을 거친다. 일반적으로 데이터 전처리, 특성 추출, 분류의 과정을 거치며, 앞의 기계학습 기반의 방법과 다르게 딥러닝 모델의 경우에는 전문가에 의한 특성 추출과 선택이 필요 없고, 이 역할을 모델이 대신해서 학습하고 수행한다. 추출된 특성을 통해 모델이 최종적으로 심장병이 있는지 없는지를 추론하게 된다. 앞의 머신러닝 기법들처럼 대부분의 딥러닝 기반의 심장 분류 모델들 역시 PhysioNet 2016 데이터를 통해 평가되었다.

딥러닝 기반의 모델을 사용해서 심음을 분류한 다음과 같은 연구들이 있다. MFCC(Mel-frequency cepstrum coefficient)로 심음의 특징을 추출한 뒤 CNN 기반 모델인 심층 컨볼루션 순환 신경망(CRNN)을 사용하여 0.987의 민감도와 0.98의 민감도 및 0.983의 정확도로 분류한 연구[18], MFCC로 심음의 특징을 추출한 뒤 RNN 기반의 모델인 LSTM을 통해 0.9995의 민감도와 0.9671의 특이도의 결과를 얻은 연구[19]가 있으며, 이미지를 입력으로 사용한 연구의 예시로 연속 웨이블릿 변환(CWT) 결과와 함께 짧은 5초의 PCG 기록을 사전 처리한 후 2D 스칼로그램(scalogram) 이미지로 2d AlexNet[20]에 입력하여 0.9의 민감도 및 0.9의 특이도를 얻은 연구[21]가 있다.

이러한 딥러닝 기반의 모델을 사용할 경우, 다른 방법들보다 높은

성능으로 심장병 환자의 판별이 가능하다. 또한 직접 특성을 추출하고 선택할 필요가 없어 전문가의 도움을 비교적 덜 필요로 한다. 하지만, 딥러닝 모델링을 통한 모델들은 대부분 블랙박스(black box) 라는 단점을 갖는다. 이는 전문가가 이 모델이 어떻게 환자를 분류했는가에 대해 알 수 없어, 이를 통해 진단하기에는 설명에 대한 어려움이 있다. 이러한 단점에도 불구하고, 최근 연구에는 성능이 좋은 딥러닝 모델을 사용한 연구가 많이 진행되고 있다.

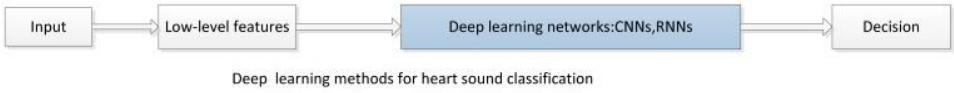


그림 4. 딥러닝 기반 모델의 모델링 과정[7]

제 2 절 소리의 시각화

본 논문에서 사용하는 ViT의 입력 형태를 만족하기 위해서는 사용하는 데이터인 소리 데이터를 이미지로 변환하는 시각화 과정이 필요하다. 소리의 음파 신호를 그대로 시각화 하여 이미지로 만들 수 있지만, 소리에서 중요한 특성을 표현하는 스펙트로그램(Spectrogram) 및 멜 스펙트로그램(Mel-spectrogram)의 시각화를 사용할 수도 있다. 이에 이번절에서는 본 논문에서 사용한 스펙트로그램을 이용한 소리를 시각화 하는 방법과, 소리를 스펙트로그램으로 변환하는 과정에서 사용되는 푸리에 변환(Fourier Transform)에 대해 소개하고자 한다.

푸리에 변환

소리에서의 푸리에 변환은 소리의 시간과 진폭으로 표현되는 함수를 다양한 주파수를 갖는 신호의 합으로 분해하여 표현하는 변환이며, 컴퓨터에서 이를 구현할 때는 불연속적으로 변하는 함수에 대한 푸리에 변환인 이산 푸리에 변환(Discrete Fourier Transform, DFT)으로 계산한다. 하지만 이산 푸리에 변환은 $O(N^2)$의 계산복잡도를 가져서 실제 사용에는 비효율적인 알고리즘이다. 이러한 점이 착안되어 고속 푸리에 변환(Fast Fourier Transform) 알고리즘을 사용하여, 이산 푸리에 변환 및 역 변환 계산에 대칭성을 이용하여 계산복잡도를 $O(N \log N)$으로 줄일 수 있다. [22]

$$f_k = \sum_{n=0}^{N-1} x_n e^{\frac{-2\pi}{N} ikn}, \ \ k = 0,1,2,\dots,N-1$$

수식 1. 이산 푸리에 변환 계산

국소 푸리에 변환(Short Time Fourier Transform)이란, 주파수 특성이 시간이 따라 달라지는 신호를 분석하기 위한 방법으로, 소리와 같은 성질을 가진 신호들에 사용된다. 시간에 따라 변화하는 함수를 일정한 시간 간격으로 분할하며, 이를 윈도우(window)라고 한다. 분할된 신호의 분석을 위한 윈도우를 적용하여 신호 비정상성(non-stationary)을 처리하는 푸리에 변환의 자연스러운 확장이며, 아래 식과 같이 표현된다. STFT는 푸리에 변환에 비해 더 나은 시간 및 주파수 위치의 특성을 갖지만, 시간 해상도(resolution)와 주파수 해상도의 곱이 일정하여 이를 통해 생성된 특징으로는 시간 및 주파수의 즉각적인 위치를 파악할 수 없으며, 고정된 창 길이와 고정된 기준 함수를 사용하기 때문에 STFT로는 지속 시간이 다르거나 신호에 빠른(예리한) 이벤트가 포함된 이벤트를 캡처할 수 없다. [23, 24]

$$STFT\{x(t)\}(\tau, w) = \int_{-\infty}^{\infty} x(t) w(t-\tau) e^{-iwt}\, dt$$

수식 2. 국소 푸리에 변환 계산

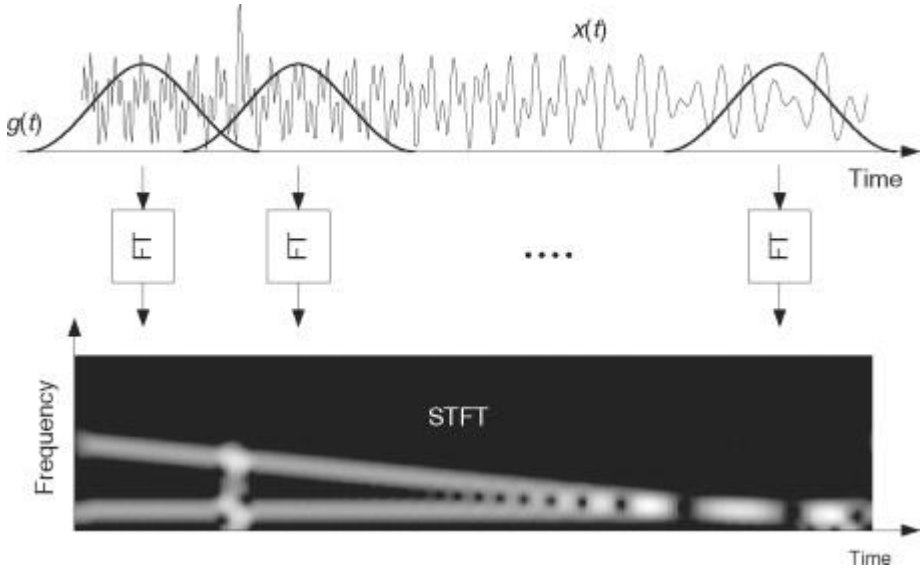


그림 5. 국소 푸리에 변환 과정[25]

스펙트로그램

스펙트로그램(spectrogram)은 신호를 시각화 하여 분석하는 방법으로, 파형(waveform)과 스펙트럼(spectrum)의 특징을 갖는 데이터이다. 특정 파형에 존재하는 다양한 주파수에서 시간이 경과함에 따라 변하는 신호의 세기(음량)을 시각적으로 나타내는 방법이다. 원 신호의 함수에서는 시간의 변화에 따른 진폭의 변화를 볼 수 있지만, 스펙트로그램에서는 주파수 변화에 따른 진폭의 변화를 볼 수 있게 변환된다. 스펙트로그램을 계산하기 위해서는 각 구간별로 윈도우를 씌운 뒤 빠른 푸리에 변환(FFT)을 계산해 평균을 내어 구한다. [26]

스펙트로그램은 기본적으로 2차원 그래프이며, 색의 차이를 통해 세기를 나타낸다. 가로축은 시간을 나타내며, 세로축은 주파수를 나타내며 피치 또는 톤으로도 생각할 수 있다. 특정 시간에 특정

주파수의 세기(음량)는 3차원인 색상으로 표시되며 어두운 색은 낮은 세기에 해당하고 더 밝은 색상은 점점 더 강한 세기에 해당한다. [27]

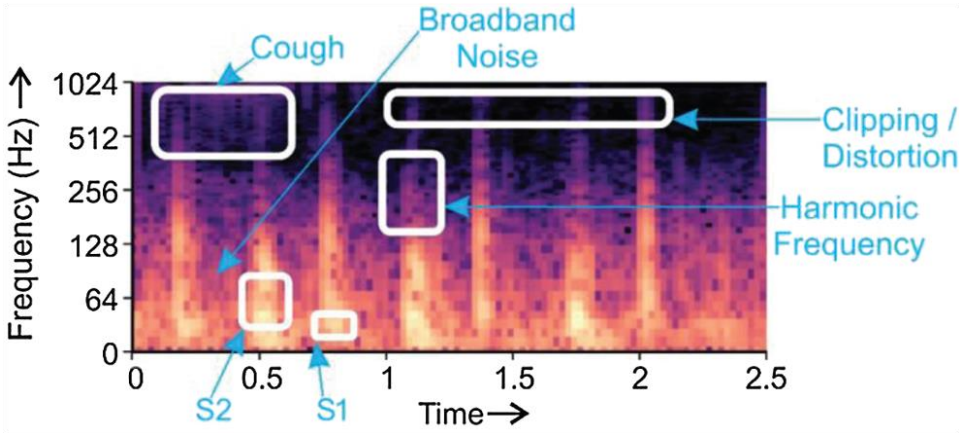


그림 6. 심장 소리 스펙트로그램[27]

제 3 절 Vision Transformer

이번 절에서는 본 논문에서 사용하는 모델인 Vision Transformer(ViT)[12]에 대해 설명하며, Transformer[10]의 구조와 어텐션(attention)기법에 대해서 소개한다. Transformer 구조는 딥러닝 다양한 분야에 적용되면서 좋은 성능을 보여주며, 대표적으로 자연어 처리 분야에 사용되는 BERT[11]가 있다. 컴퓨터 비전 분야에서도 CNN을 사용하지 않고 Transformer를 기반으로 하는 ViT가 이미지 분류에서 좋은 성능을 보인다. 본 논문에서는 추가적인 인구통계학적 정보를 사용하기 위해 분류기 부분의 개조가 이루어졌으며, 이러한 과정을 설명하기 위해 ViT 구조를 소개하고자 한다.

Transformer

Transformer는 2017년 구글이 발표한 모델로, RNN기반의 모델의 장기의존성을 개선하기 위한 구조이다. Transformer는 seq2seq 구조를 따르며, 인코더와 디코더로 이루어져 있다. 인코더에서는 입력된 데이터를 시퀀스에서 벡터로 압축하며, 디코더에서는 반대로 압축된 벡터를 시퀀스로 변환해서 출력하게 되는데, 기존 모델의 경우 인코더의 압축과정에서 정보의 손실이 생기는 단점이 있었는데, 이를 보완하기 위해 Transformer에서는 어텐션(Attention)을 사용한다. [10]

다음으로는 어텐션 기법에 대해 설명한다. 어텐션이란 디코더에서 시퀀스를 출력하고 이 단어를 예측하는 단계(step)별로 인코더의 입력

시퀀스를 사용하며, 이 때 입력된 시퀀스를 모두 동일한 가중치로 사용하지 않고, 예측하는 시퀀스와 관련 있는 입력 시퀀스를 더 높은 가중치로 사용한다. 어텐션 함수의 키(Key)와 쿼리(Query)의 유사도를 계산하는데, 이때 쿼리에 대해 모든 키의 유사도를 계산하게 되며, 계산된 유사도를 키가 가리키는 각각의 값(Value)에 반영한다. 이러한 값들을 모두 합하면 주의 값(Attention value)가 된다. 셀프 어텐션(Self-Attention)은 키, 쿼리, 값이 동일한 경우를 의미한다. 셀프 어텐션을 사용하면 입력되는 시퀀스 내에서의 유사도를 계산하여, 각 층(layer)별 계산 비용을 줄이고, 병렬처리를 가능하게 해주며 장기의존성 문제를 완화할 수 있다. [10]

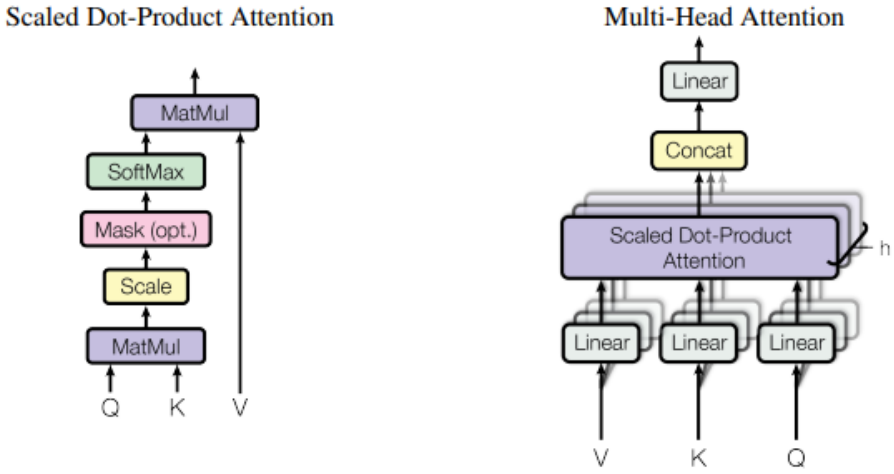


그림 7. 어텐션 메커니즘 [10]

아래의 그림 8을 보면 왼쪽이 Transformer의 인코더, 우측이

디코더를 나타내며, 원 논문에서는 각 6장의 층을 쌓았다. 입력 데이터가 인코더에 들어가기 전 임베딩과 위치 인코딩(positional encoding)이 진행되며, 인코더에 입력된 후 다중 헤드 셀프어텐션(multi-head self attention)과 피드포워드 네트워크(Feed forward network)를 통과한다. 다중 헤드 셀프어텐션에서는 8개의 어텐션 층을 두고, 각각 다른 초기값을 사용해 학습한다. 마지막으로 각 층에서 나온 출력을 모두 더한 뒤 가중치 벡터(weight vector)를 곱해줌으로써 하나의 벡터로 변환하고, 피드포워드 네트워크를 통과하여 임베딩을 출력하며, 여기서 가장 상위의 출력이 키(Key)와 값(Value)이 된다. [10]

디코더는 가려진 멀티 헤드 어텐션(masked multi head attention), 멀티 헤드 어텐션 및 피드포워드 네트워크로 구성된다. 가려진 멀티 헤드 어텐션은 출력 시퀀스(sequence)의 이전 위치만을 어텐션 하기 위해 앞으로의 위치(future position)을 음의 무한대로 마스킹하며, 이후의 동작은 인코더의 멀티 헤드 어텐션과 동일하게 이루어진다. 디코더의 멀티 헤드 어텐션은 인코더의 출력값과 디코더의 입력값을 이용하는 인코더-디코더 어텐션이며, 이때의 쿼리(query)는 디코더의 가려진 멀티 헤드 어텐션에서의 출력 값이며, 키와 값은 인코더에서의 출력 값이다. 피드포워드 네트워크의 동작 역시 인코더와 동일하다. [10]

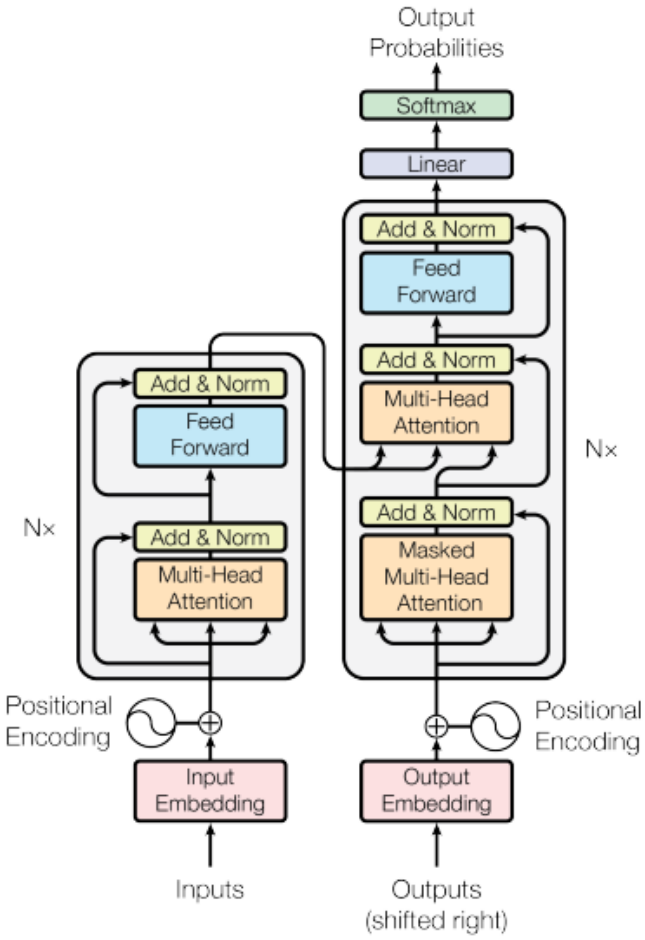


그림 8. 트랜스포머 구조 [10]

Vision Transformer

컴퓨터 비전 분야에서도 Transformer 구조를 활용하기 위한 시도들이 이루어졌다. ViT 모델은 기존 컨볼루션(convolution) 연산에 의존하던 CNN(Convolution Neural Network)을 쓰지 않고, 성공적으로 Transformer 구조를 이미지 분류 모델로 적용한 모델이다. 이 모델은 이미지를 작은 패치로 분할하고 임베딩하여 이미지 분류를 수행하는

모델로, CNN 기반 모델보다 미세 조정(fine tuning)과정에서 더 적은 데이터로도 더 높은 성능을 보여주는 모델이다. 또한, 모델의 추론 결과를 시각적으로 이해하는 데 도움이 될 수 있는 어텐션 지도를 생성할 수 있으며, ViT의 구조는 그림9와 같다.

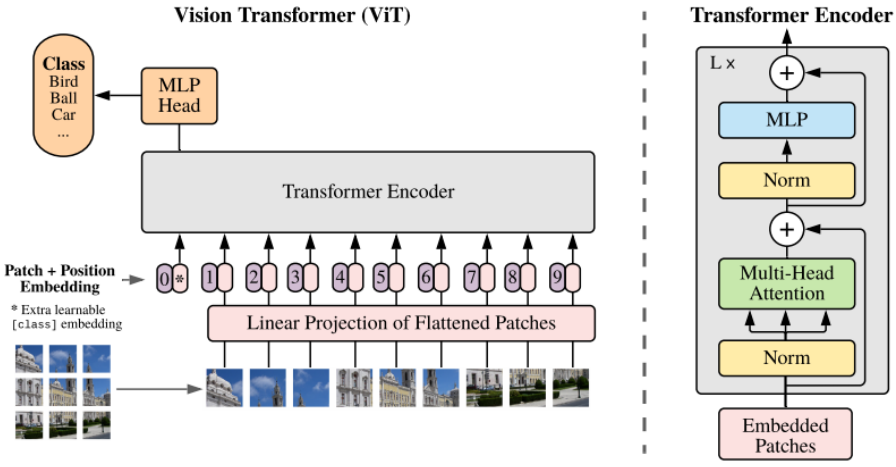


그림 9. ViT 구조[12]

ViT는 이미지를 입력의 기본단위인 패치(patch)들로 분할하며, 각 패치를 임베딩한다. 본 논문에서 사용된 ViT-B16의 경우, 입력되는 이미지의 크기는 224×224 pixel, 입력의 기본단위인 패치의 크기는 16×16 pixel로, 하나의 이미지는 총 228개의 patch로 분할되며, 패치별 임베딩 차원은 16×16×3=768이 된다. 다음으로 임베딩된 패치에 클래스 토큰(class token)을 붙이고, 각 패치마다 순서 정보를 담은 위치 임베딩(Position Embedding)을 더한다. 이 때, 클래스 토큰 및

위치 임베딩 모두 학습되는 파라미터(parameter)이며, ViT B-16 모델의 경우 Transformer의 인코더가 12층으로 쌓여 있는 구조이며, 각 인코더는 앞서 설명한 Transformer의 인코더에서의 동작과 동일하다. 이렇게 모든 인코더를 지나면, 분류기에서 최종적으로 추론을 하며, 이 때 선택적으로 주의 지도와 마스크를 출력할 수 있다. [12]

ViT에서 어텐션 지도와 마스크는 Attention Rollout[28]에 의해 계산된다. 이는 모든 헤드에서 ViT의 주의 가중치를 평균화하고, 모든 레이어의 가중치 행렬을 재귀적으로 곱하는 방법이며, 모든 레이어에서의 토큰에 대한 어텐션을 복합적으로 나타낸다. 어텐션 지도는 추론하고자 하는 이미지에 어텐션 마스크를 곱해서 구할 수 있으며, 그림 10에서 볼 수 있듯 어텐션 지도는 모델이 이미지의 어느 부분을 어텐션 했는지를 확인할 수 있는 시각화 기법이다. [12]

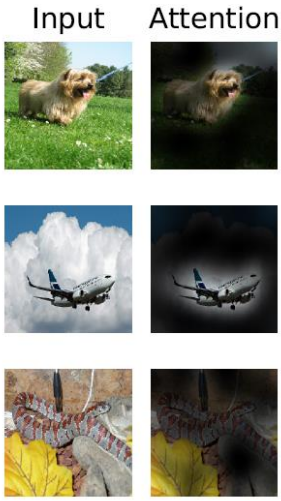


그림 10. 어텐션 지도 예시[12]

제 4 절 데이터 증강 기법

이번 절에서는 PCG 연구의 한계점 중 하나인 데이터가 부족하다는 부분을 보완하여 모델의 성능을 향상시키기 위한 데이터 증강 기법을 소개하고자 한다. 현실의 데이터셋을 컴퓨터로 학습하고 이를 평가할 때 데이터 셋이 부족한 경우가 많다. 모델을 학습할 때 데이터가 부족하면, 만들어진 모델은 실제 데이터의 특징을 잘 반영하지 못하고 학습된다. 예를 들어 훈련 데이터에 대해 너무 많은 학습이 될 경우에는 과대적합(overfitting)될 가능성이 높다. 이러한 과적합을 데이터 증강(Data Augmentation) 기법을 통해 완화할 수 있으며, 이는 모델의 성능 향상에 큰 영향을 미치는 정규화(regularization) 기법이다.

본 논문의 기반모델인 ViT는 주로 이미지 분류작업에서 사용되는 모델이기에 이미지를 위한 증강 기법이 사용된다. 또한, 스펙트로그램 이미지의 경우 소리 데이터의 성질을 갖고 있기에 이러한 증강 기법들이 오히려 중요한 정보를 손상시킬 수 있다. 이러한 점에 비추어 이번 절에서는 컴퓨터 비전에서 사용되는 이미지 증강 기법과, 음성 데이터에서 사용되는 소리 증강 기법에 대해서 소개하고자 한다.

그림 11. 이미지 데이터 증강 기법 종류[29]

이미지 데이터 증강

컴퓨터 비전에서는 이미지의 데이터를 증가시켜 모델의 성능을 높이기 위한 기법들이 보편적으로 사용된다. 전통적인 데이터 증강에는 해당 이미지가 나타내는 물체를 최대한 보존하는 방식으로, 뒤집기(flipping), 색 공간 증대(color space augmentations), 임의 자르기(random cropping), 회전(rotation) 등의 방법이 있다. 이런 방법들은 이미지의 일부에만 변화를 주어 추가적인 데이터를 만드는 기법들이며, 해당 이미지의 본질과 중요한 정보를 크게 해치지 않아 보편적으로 사용된다. 하지만 이미지 별 포함하는 정보(동물, 숫자,

의료영상)별로 증강 기법에 따라 효율이 달라지므로, 컴퓨터 비전 분야에서는 주로 사용되는 데이터셋들에 대해 효율이 좋은 증강 방법에 관한 연구가 많이 이루어졌다. [29]

전통적인 데이터 증강 기법 이외에도 많이 사용되는 증강 방법으로는 이미지 혼합 기반의 방식이 있다. 이미지 혼합 방식에는 두가지 이미지와 라벨을 일정 비율로 섞어 새로운 이미지(데이터)를 만드는 mixup[30]의 방식이 포함된다. 그 외에도 GAN[31] 기반의 방법으로 생성 모델(Generative model)을 활용하여 추가적인 의료 영상 데이터를 생성하는 방식의 데이터 증강 접근법도 제안되고 있다.[32] 생성 모델 기반의 방식에서 기본이 되는 생성모델은 생성자(generator)와 구분자(discriminator)로 이루어진 구조이며, 생성자에서는 실제 데이터에 변화를 준 가짜 데이터를 생성하며, 구분자에서는 이렇게 생성된 데이터가 진짜인지를 구분하는 역할을 한다. 이렇게 GAN 기반의 모델을 통해 데이터를 직접 생성해서 사용하는 방법도 있다.

[33]에 따르면, ViT의 학습에는 많은 양의 데이터가 필요로 되며, ViT에서 mixup은 효과적인 데이터 증강 기법이다. 하지만 mixup을 포함한 이미지에서 사용되는 대다수의 데이터 증강 기법은 데이터의 특성 따라 중요한 정보를 손상시킬 수 있으며, 이러한 경우에는 모델의 학습을 방해하여, 모델이 데이터의 특징을 잘 반영하지 못하는 과소적합(underfitting)의 원인이 되기도 한다.

소리 데이터 증강

소리데이터에서도 데이터를 증가시켜 모델의 성능을 높이기 위해 증강 기법들이 사용된다. 소리 데이터와 이미지 데이터는 데이터의 특징이 많이 다르기 때문에 이미지 증강 기법이 소리 데이터의 중요한 정보를 손상시켜 데이터 증강의 효과가 떨어진다. 이에 소리 데이터를 위한 다양한 데이터 증강 기법들이 연구되어왔다. 소리 데이터에 직접 사용하기도 하지만, 소리 신호에서 추출할 수 있는 특성 중 하나인 스펙트로그램으로 변환하여 사용할 수 있는 기법도 연구되어왔다.

소리 데이터에 보편적으로 사용되는 데이터 증강 방법에는 임의의 잡음을 추가하는 잡음 주입(noise injection), 소리의 시간 축에 변화를 주는 시간 이동(time shifting), 소리의 속도를 빠르게 하거나 느리게 하는 속도 변화(speed changing) 등의 기법이 있다. 하지만 스펙트로그램은 시간에 따라 중요한 정보를 담고 있는 소리의 특징 데이터이자 시간 순차 데이터(time-sequential-data)여서, 이러한 방법을 적용시키면 중요한 정보를 손상시킬 수 있다. 이에 스펙트로그램의 중요한 정보 손실을 최소화하고, 효과적으로 데이터를 증강시킬 수 있는 기법인 SpecAugment[34]가 연구되었다. SpecAugment에는 시간 뒤틀림(time warping), 주파수 마스킹(frequency masking), 시간 마스킹(time masking)의 세 가지 기본 방법으로 데이터를 증강시킨다. 그림 12의 두번째 이미지에 시간 뒤틀림의 예시가 있으며, 이 기법은 좌우로 소리를 늘리며 임의로 3개의

좌표(중심, 좌, 우)를 선택하여 중심 좌표를 중심으로 좌측 좌표와 우측 좌표를 뒤틀림 요소(warping factor) α만큼 이동하는 기법이다. 그림 12의 세번째 이미지인 주파수 마스킹은 주파수축의 일부 영역들의 정보를 가리는 기법이며, 그림 12의 네번째 이미지인 시간 마스킹은 시간축의 특정 영역들의 정보를 가리는 기법이다. 이 두가지 기법에 비해 시간 뒤틀림 기법은 비효율적이다. 소리의 크기를 점차적으로 키우거나 줄이는 방법인 페이드 인, 아웃(fade in and out) 기법도 종종 사용된다. 이러한 증강 기법은 데이터의 중요한 정보를 유지시킴으로써, 모델이 데이터의 특성 학습에 있어 문제가 생기지 않으며, 모델의 정규화를 돕는다. SpecAugment는 간단하고 적은 컴퓨팅 자원(computing resource)을 필요로 하는 효율적인 방법이다. [34]

ViT의 학습에는 많은 양의 데이터가 필요로 되며, 스펙트로그램 이미지를 사용한다면 데이터의 정보를 손상시키지 않는 선에서의 소리 증강 및 SpecAugment 방법을 사용하는 것이 적절할 것이다. 이러한 기법이 모델이 데이터의 특징을 반영하는데 방해하지 않고, 모델의 일반화를 도와 과적합(overfitting)을 방지할 수 있을 것이다.

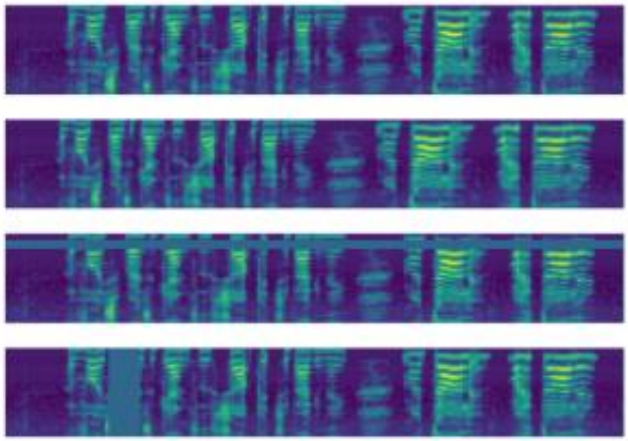


그림 12. SpecAugment 예시[34]

제 3 장 연구 방법 및 설계

이번 절에서는 앞에서 소개한 ViT를 심장 소리 데이터를 통해 학습하기 위해 심장 소리를 적합한 이미지로 변환 한 과정과, 환자 분류 모델의 성능을 높이기 위해 진행 한 데이터 전처리 과정들에 대해서 설명하고, 이러한 방법을 통해 심장 소리 데이터를 ViT를 통해 분류할 수 있음을 보이고자 한다. 본 논문의 모델의 구조는 ViT 논문에서 설명되는 모델 구조를 기반으로 하며, 인구통계학적 정보를 사용하기 위해 모델에서의 분류기의 수정을 거쳤다. 이러한 기법들에 대한 설명을 하기 이전에 본 논문에서 사용한 데이터셋에 대해 소개하고자 한다. 그림 13과 같이 입력 데이터를 데이터 전처리, 데이터 증강을 거친 이후 ViT 모델에 입력하여 결과를 출력하는 과정을 거친다.

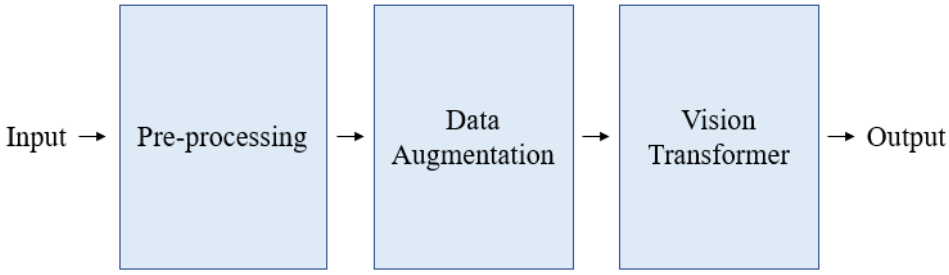


그림 13. ViT 구조를 통한 심장 소리 분류 메커니즘

제 1 절 데이터셋

본 논문의 실험에는 The George B. Moody Challenge 2022

데이터셋이 사용되며, 이 데이터는 2014년 7월부터 8월과 2015년 6월부터 7월의 브라질 북동부에서 소아로부터 수집되었다. 대회의 전체 데이터에는 1568명의 환자에 대한 하나 이상의 심장 소리 녹음과 환자에 대한 일상적인 인구 통계 정보가 포함되어 있다. 대회 레이블은 두 가지 유형으로 구성되는데, 임상결과 관련 라벨은 의료 전문가가 진단한 정상 또는 비정상 임상 결과를 나타내며, 심잡음 관련 레이블은 전문가가 녹음에서 환자의 잡음을 감지했는지 혹은 확신하지 못했는지 여부를 나타낸다. [35]

알고리즘의 출력이 비정상인 경우 환자를 전문가에게 의뢰하며, 알고리즘의 출력이 정상인 경우 환자를 전문가에게 의뢰하지 않는다. 환자가 전문가에게 의뢰되면 전문가가 환자를 선별하며, 이 때 임상 결과가 비정상이면 환자는 치료를 받는다. 전문가는 진단 오류를 범하지 않는다고 가정하며, 환자가 전문가에게 의뢰되지 않은 경우와, 전문가가 정상으로 임상 결과를 판단했다면 환자는 치료를 받지 않는다. [35]

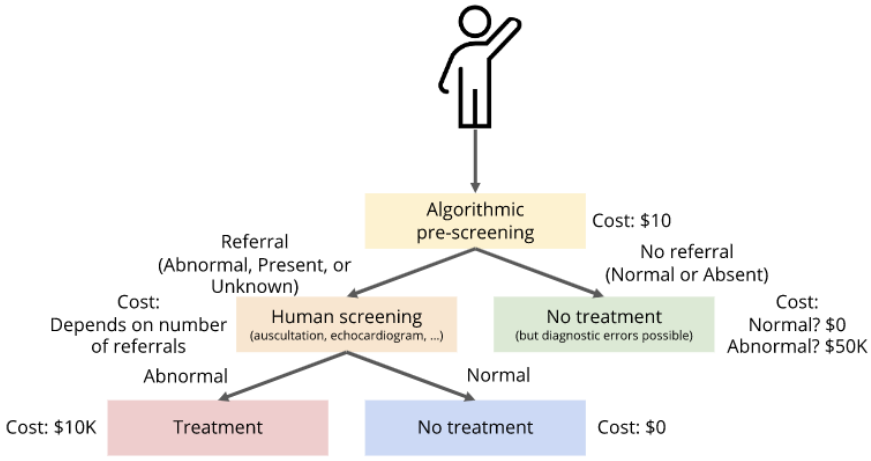


그림 14. 의사 결정 프로세스 및 비용[35]

대회 데이터에서 각 환자는 폐동맥판(PV), 대동맥판(AV), 승모판(MV), 삼첨판(TV) 및 기타(Phc) 중 하나 이상의 청진 위치에서 기록을 측정했다. 심장 소리 녹음 데이터는 서로 다른 청진 위치에서 디지털 청진기를 사용하여 순차적으로 측정되었다. 녹음 데이터의 개수, 측정한 위치 및 시간은 환자마다 상이하며, 대회에서 공개한 데이터인 학습 데이터는 전체 데이터의 약 60%로 942명으로부터 얻은 3163개의 심장 소리 데이터를 포함하며, 나머지 40%의 데이터는 공개되지 않으며, 확인 및 평가데이터로 대회에 제출된 코드들의 성능 측정에 사용된다. 본 논문에서 제안하는 방법으로 Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022에 참여하였으며, 공개되지 않는 확인 및 평가데이터에 대한 결과를 포함한다.

본 논문에서 사용하는 해당 데이터셋이 포함하는 정보는 아래의 표1과 같으며, 심잡음 태스크의 경우 표2와 같이 데이터가 불균형 하다.

변수	가능한 값
Age	Neonate, Infant, Child, Adolescent, Young adult
Sex	Female, Male
Height	> 0
Weight	> 0
Pregnancy status	True, False
Murmur	Present, Absent, Unknown
Outcome	Normal, Abnormal

표 1. PhysioNet 2022 데이터셋에서 사용하는 변수

Clinical Outcome	
Normal	486
Abnormal	456

Murmur	
Absent	695
Present	179
Unknown	68

표2. PhysioNet 2022 데이터셋의 클래스별 개수

제 2 절 전처리 및 이미지 데이터셋 생성

이번 절에서는 심음을 ViT 모델을 잘 학습할 수 있는 이미지 데이터로 변환하는 과정을 소개하고자 한다. 데이터 전처리 과정은 길이가 서로 다른 심장 소리 데이터들을 균등한 시간으로 만들기 위해 신호를 분할하는 과정으로 시작한다. 그 다음으로 심음 데이터를 로그 스펙트로그램으로 변환한 뒤, ViT의 성능을 이끌어 낼 수 있는 크기와 형태의 이미지로 변환한다. 또한 데이터셋에 포함된 환자의 인구통계학적 정보를 사용하기 위해 간단한 방법으로 특성을 추출한다. 마지막으로 의료데이터의 공통적인 한계점인 데이터가 부족하다는 부분을 개선하기 위해 PCG 스펙트로그램 이미지에 적합한 데이터 증강 기법을 사용한다.

본 논문에서는 주어진 데이터에 대한 전문 지식이 없어 심장 소리 데이터의 라벨 재지정(re-labeling)이나 제외(excluding)는 진행하지 않았다. 마찬가지로 최적 값을 알 수 없어 복잡한 과정이 될 수 있는 전처리 과정인 신호 필터링이나 리샘플링 역시 진행하지 않았다. 기반이 되는 모델인 사전 훈련된(pre-trained) ViT 모델이 224×224 pixel의 이미지를 입력 받기 때문에 동일한 크기인 224×224 pixel 로 스펙트로그램 이미지를 만든다. 이는 다른 크기의 이미지를 크기 조정(resizing)하게 되면 이미지가 포함한 정보를 손실하게 되므로, 이를 최소화하기 위함이다. 같은 이유로 패치 크기는 ViT B-16에서 사용되는 크기인 16×16 pixel을 사용했다.

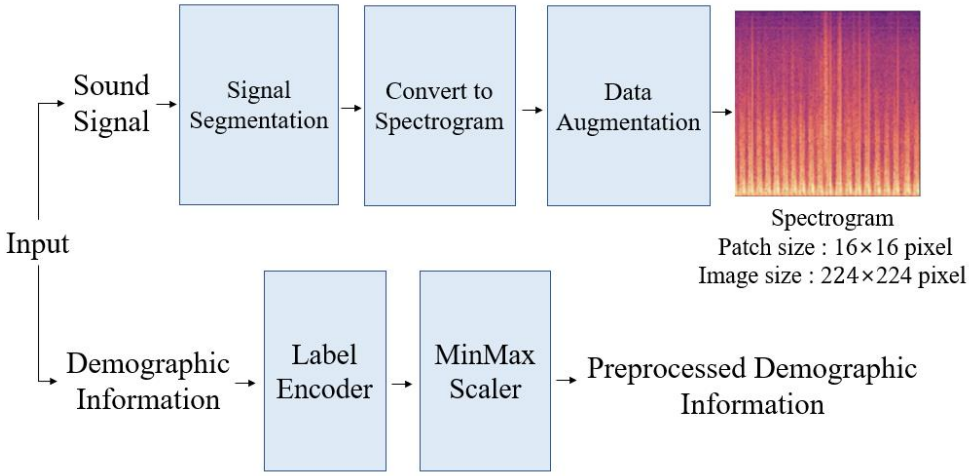


그림 15. 심장 소리 데이터 및 인구통계학적 정보 변환 과정

2.1 신호 분할 및 스펙트로그램 변환

신호 분할 과정은 길이가 일정하지 않은 심장 소리를 겹치는 부분 없이 일정한 시간으로 분할하여 모델이 잘 학습할 수 있게 만들기 위함이다. Raza[36]는 심음 분류 딥러닝 모델의 입력으로 3초, 7초, 12.5초, 25초로 심음도를 분할해서 성능을 비교했는데, 12.5초가 가장 높은 성능을 보였다. 본 논문에서도 이러한 연구 결과를 참고하여 심음도를 12.5초로 분할하여 사용했다.

스펙트로그램 변환 과정에서는 심음이 ViT로 입력되는 단위인 패치(patch)단위로 나뉠 수 있는 적합한 크기와 형태의 이미지로 변환한다. 각 심장 소리 데이터를 224×224 pixel의 스펙트로그램으로 변환하기 위해 창 크기(window size)는 0.11을 사용했고, 겹침 비율(overlap ratio)는 0.5를 사용했다. 대부분의 선행연구에서 사용된 전처리 기법인 리샘플링 및 필터링은 전문 지식 없이 본 모델에 적합한

값을 찾는 과정이 복잡하며, 심음도가 가진 정보의 손실을 최소화하기 위해 수행하지 않았다.

원 심음 데이터가 12.5초보다 작을 경우에는 검정 pixel 이미지를 사용해 부족한 부분을 채움으로써 제로 패딩(zero padding)의 효과를 주었다. 또한 12.5초가 넘을 경우, 12.5초로 나눠지는 만큼을 제외하고 앞 뒤 같은 길이만큼 잘라서 사용하였다. 예를 들어 그림16과 같이 8.5초의 신호가 있다면, 신호를 스펙트로그램으로 변환한 뒤 부족한 4초를 검정 pixel 이미지로 채워준다. 다음으로 24초의 신호가 있다면 앞 5.75초, 뒤 5.75초를 버리고 12.5초에 해당하는 신호만을 사용한다. 마지막으로 32초의 신호가 있다면, 앞 3.5초, 뒤 3.5초를 버리고 25초에 해당하는 신호를 12.5초 두개의 신호로 분할하여 사용한다. 신호를 분할하는 방법을 사용함으로써 부족한 데이터를 증강시키는 효과도 기대할 수 있다.

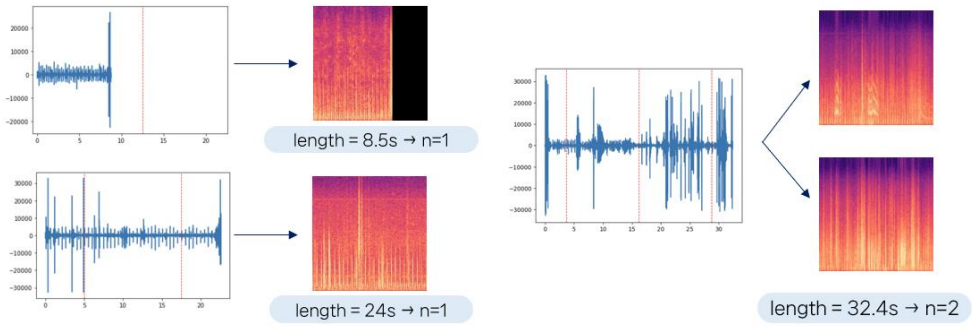


그림 16. 스펙트로그램 이미지 변환 과정

2.2 데이터 증강

본 연구에서는 의료데이터 중 하나인 심음 데이터를 사용하며, 모델을 만드는 과정에서 대부분의 의료데이터가 겪는 데이터의 부족 문제가 발생했다. 이 문제를 해결하기 위한 효과적인 방법으로 데이터 증강기법을 채택했다. 앞서 소개한 이미지에 전통적으로 사용되는 뒤집기, 회전, 색 반전 등의 증강 방법을 본 데이터에 사용해본 결과, 데이터 증강의 효과보다 심음 스펙트로그램 이미지의 중요한 정보를 손상시켜 모델의 성능을 향상시키지 못했다. 이에 스펙트로그램 신호의 중요한 정보를 손상시키지 않고, 데이터 증강의 효과를 주는 SpecAugment를 적용했다. SpecAugment에는 주파수 마스킹, 시간 마스킹이 포함되며, 그 외의 증강 방법으로 페이드 인/아웃 기법을 적용했다. 대회 훈련 데이터와 검증 데이터에서 시간 마스킹이 다른 어떤 증강 기법보다 높은 성능을 보였기에, 시간 마스킹을 사용하여 데이터를 2배로 증강시켰다. 그림 17의 (a)는 기본 스펙트로그램 이미지, (b)는 주파수 마스킹, (c)는 시간 마스킹, (d) 페이드 인/아웃 기법을 적용한 이미지이다.

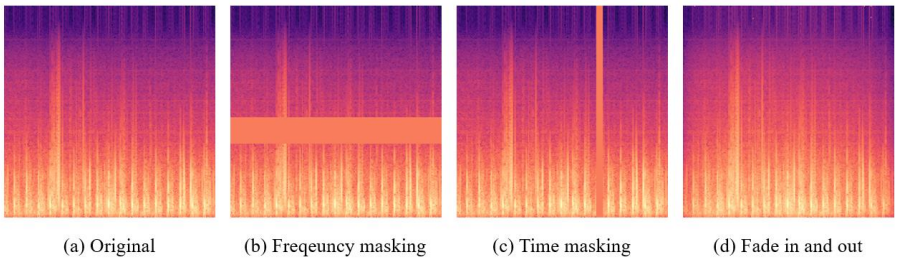


그림 17. 데이터 증강 기법을 적용한 스펙트로그램

2.3 인구통계학적 정보 변환

데이터 전처리 대회 데이터 세트에서 인구통계학적 특성 중 나이, 성별, 키, 체중, 임신 정보를 모델에 적용하기 위해 전처리를 진행했다. 범주형 데이터의 경우 레이블 인코더를 사용하여 변환한 뒤, 각 인코더를 저장하여 테스트 데이터 변환에 사용하였다. 결측 값의 경우 새로운 클래스로 변환했다. 수치형 데이터의 경우 최대 최소 스케일러를 이용하여 변환하였으며, 각 스케일러를 저장하여 테스트 데이터 변환에 사용하였다. 수치형 데이터의 결측 값은 최빈값을 사용해서 채웠다. 이렇게 채워진 값은 테스트 데이터에도 사용되었다. 최종적으로 인구통계학적 특성은 [Ages, Sexs, Heights, Weights, Pregs]로 표현되며, 각 특성 별 통계 수치는 아래와 같다.

Age	
Child	664
Infant	126
Adolescent	72
Neonate	6

Sex	
Female	486
Male	456

Pregnancy	
True	872
False	70

표3. PhysioNet 2022 데이터셋 범주형 데이터 특성 요약

Weight (kg)	
Mean	23.63
Mode	24.8
Median	20.4

Height (cm)	
Mean	110.8
Mode	115
Median	115

표4. PhysioNet 2022 데이터셋 수치형 데이터 특성 요약

제 3 절 모델 학습

이번 절에서는 모델의 구조를 설명하면서 모델을 학습시킨 방법에 대해 소개하고자 한다. ImageNet-21K(이미지 1400만 개, 클래스 21843개)으로 사전 훈련된 B-16 ViT 모델을 사용했으며, 이 모델은 패치 크기가 16×16 pixel이고 이미지 크기가 224×224 pixel이다. 이미지와 함께 전 처리된 인구 통계 정보를 활용하여 모델을 학습하기 위해 분류기 부분을 개조했으며, 심잡음 및 환자 탐지 모델에 대부분 같은 하이퍼파라미터를 사용했다. ViT의 배치 크기(batch size)가 클수록 성능이 좋아지므로 테스트 환경에서 가능한 최대값을 사용하였으며, 단계(step)의 경우 실험적으로 최적 값을 찾아냈다. Transformer 구조는 많은 양의 데이터를 통해 높은 성능을 만들어 내기 때문에, 데이터가 부족하면 성능이 떨어지므로 사전 훈련 없이는 모델이 잘 동작하기 힘들다.

3.1 모델의 구조

본 연구에서 제안하는 모델은 그림 18과 같다. 앞의 과정을 따라 전 처리된 심음 신호는 스펙트로그램 이미지로 변환되어 ViT 모델에 입력되며, 가장 먼저 ViT 특징 추출기(ViT feature extractor)를 거친다. ViT 특징 추출기에서는 이미지를 정해진 패치 단위로 쪼개고, 각 패치를 1차원 임베딩 차원으로 변환한다. 만들어진 임베딩 패치에

클래스 토큰과 위치 임베딩을 더해 인코더로 전달한다. 다음으로 이미지 분류를 위해 ViT의 Transformer 인코더에 입력된다. 여기에 입력된 특징은 총 12개의 ViT 인코더 층을 거치면서 학습되며, 마지막 인코더에서 출력된 특징에 앞서 전처리한 인구통계학적 정보 특징을 연결한다. 이렇게 연결된 특징은 마지막 층인 분류기에 입력되며, 각 모델의 분류기에서 심장병 유무를 분류하거나 심잡음을 분류하게 된다. 이러한 과정을 통해 본 논문에서 제안하는 모델은 클래스를 학습하고 추론을 수행한다.

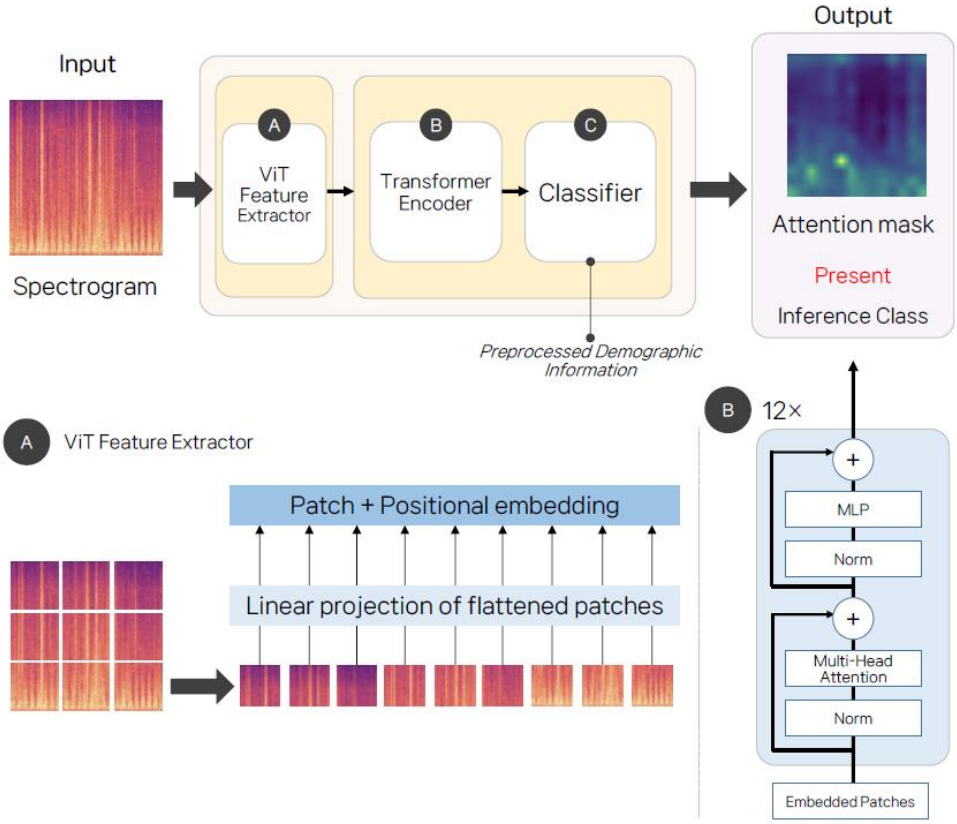


그림 18. 본 논문에서 제안하는 ViT 모델

3.2 가중 범주형 교차 엔트로피 손실

교차 엔트로피는 서로 다른 확률 분포 간의 차이를 측정하는 데 사용되며, 주로 분류에서 기본적인 손실 함수로 사용된다. 대회 데이터 세트의 클래스가 불균형하고, 점수를 계산하는데 있어 각 클래스의 중요도가 다르기 때문에 가중 범주형 교차 엔트로피 손실을 사용했다. 가중 범주형 교차 엔트로피 손실은 다음과 같이 표현된다:

$$L = -\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M} y_m^k \times w_k \times \log\left(h_\theta(x_m, k)\right)$$

수식 3. 가중 범주형 교차 엔트로피 계산

여기서 M은 훈련 예제의 수, K는 클래스의 수(환자 분류 모델에서 2, 심잡음 모델에서 3), y_m^k는 각 훈련 예제 m에서 클래스 k에 대한 대상 레이블, w_k는 가중치(환자 분류 모델에서 클래스별 비율 1.2:1, 심잡음 모델에서 Absent, Present, Unknown 클래스별 비율 1:5:3), h_θ는 가중치가 θ인 모델, x_m은 학습 예제 m에 대한 입력이다. [37]

일반적으로 가중치를 결정하는 방법에는 모두 동일한 가중치를 사용하거나 클래스 비율의 역수를 사용하거나 비용 함수 가중치의 비율을 사용하는 방법들이 있다. 환자 탐지 모델의 경우에는 1.2:1의 가중치를 미세 조정을 통해 선택했으며 이는 숨겨진 검증 세트에서 경험적으로 최고의 성능을 낸 하이퍼 파라미터이다. 심잡음 탐지의 경우,

앞의 경우들 중 대회의 잡음 비용 함수의 비율과 동일한 1:5:3의 가중치 비율이 가장 좋은 성능을 보였다.

제 4 장 실험 및 결과

이번 장에서는 본 논문에서 제안하는 모델의 PhysioNet challenge 2022 학습 데이터셋으로 진행된 절제 연구(ablation study)의 결과와 공개되지 않은 검증 및 평가 데이터를 통한 결과를 보이며 이를 설명한다. 이러한 결과를 통해 본 논문에서 제안하는 방법들이 ViT에 적절한 학습 방법임을 보이고자 한다.

제 1 절 평가 지표

본 절에서는 실험에 사용된 평가 지표를 설명한다. 심장병 환자를 분류하는 모델의 성능은 PhysioNet challenge 2022의 공식 평가지표인 대회 지표(challenge metric)를 사용했으며, 심잡음을 분류하는 모델의 결과는 대회에서 정한 가중치로 가중된 정확도(Weighted Accuracy)를 사용했다. 대회 지표는 비용 함수 기반의 지표이며, 알고리즘의 사전심사(prescreening) 비용, 전문가의 심사(screening) 및 처리(treatment) 그리고 진단시의 오류와 값의 누락을 모두 고려하는 비용의 합이며 다음과 같이 계산한다.

	전문가의 판단		
모델의 결과		Abnormal	Normal
	Abnormal	n_{TP}	n_{FP}
	Normal	n_{FN}	n_{TN}

표5. 심장병 환자 분류 모델 혼동 행렬

	전문가의 판단			
모델의 결과		Present	Unknown	Absent
	Present	m_{PP}	m_{PU}	m_{PA}
	Unknown	m_{UP}	m_{UU}	m_{UA}
	Absent	m_{AP}	m_{AU}	m_{AA}

표6. 심잡음 검출 모델 혼동 행렬

$$\boldsymbol{C_{outcome}^{total} = C_{algorithm}(N) + C_{expert}(n_{TP} + n_{FP}) + C_{treatment}(n_{TP}) + C_{error}(n_{FN})}$$

수식 4. 대회 비용 함수 계산

여기서 $C_{algorithm}$ 은 10s, C_{expert} 는 $\left(25 + 397\frac{s}{t} - 1718\frac{s^2}{t^2} + 11296\frac{s^4}{t^4}\right)t$, $C_{treatment}$ 는 10000s, C_{error} 는 50000s로 표현되며, $\boldsymbol{C_{outcome}^{total}}$은 t명의 환자 모집단 중 전문가가 s번 검진하는 총 비용이 된다. $C_{outcome}$ 을 통해 심장병 환자를 분류하는 모델을 평가하며, 값이 작을수록 성능이 좋음을 나타낸다.

심잡음 검출 모델의 경우에는 가중된 정확도를 지표로 사용하며, 값이 클수록 성능이 좋음을 나타내며, 대회에서 정한 가중치를 통해 아래와 같이 계산한다.

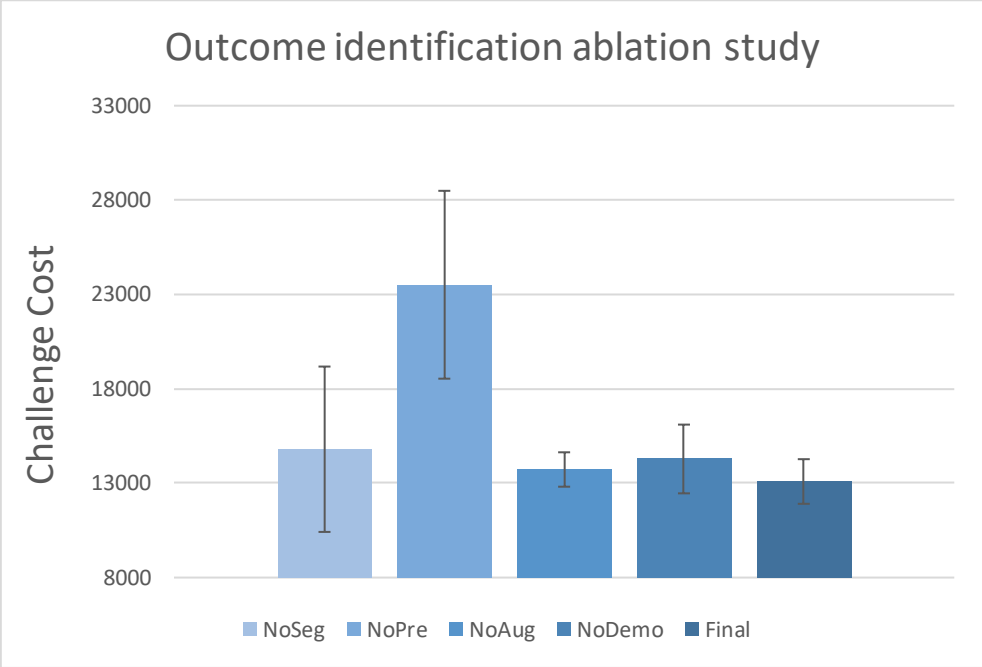
$$S_{murmur} = \frac{5m_{PP} + 3m_{UU} + m_{AA}}{5(m_{PP} + m_{UP} + m_{AP}) + 3(m_{PU} + m_{UU} + m_{AU}) + (m_{PA} + m_{UA} + m_{AA})}$$

수식 5. 대회 가중된 정확도 계산

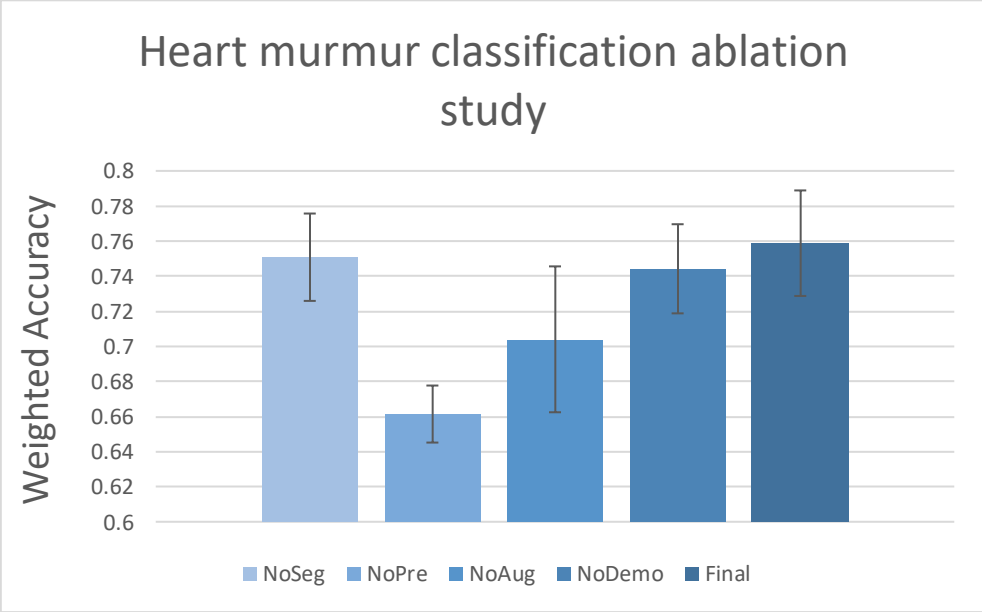
제 2 절 훈련 데이터에 대한 절제 연구

이번 절에서는 공개된 훈련 데이터에 대한 심장병 환자 검출 모델 및 심잡음 검출 모델에서 본 논문에서 제안한 방법들의 효과를 확인하기 위한 절제 연구(ablation study)가 시행되었다. 절제연구의 결과는 각각 그래프1 및 그래프2에 보고된다.

아래의 그래프에서 Final은 앞에서 소개한 방법을 토대로 신호 분할, 데이터 증강, 인구통계학적 정보를 사용, 이미지넷으로 사전 학습된 ViT 모델을 사용한 대회에 최종 제출한 모델을 의미한다. NoSeg는 Final에서 사용한 방법 중 신호 분할을 하지 않은 경우이며, 심음 데이터에서 가장 긴 길이의 데이터 길이보다 적을 경우 black pixel 이미지를 붙여 실험했다. NoPre는 사전 학습된 가중치(pretrained weight) 없이 학습한 경우이며, NoAug는 데이터 증강 기법을 제외한 경우, NoDemo는 분류기에서 인구통계학적 정보를 추가하지 않은 경우를 의미한다.



그래프 1. 심장병 환자 검출 절제 연구



그래프 2. 심잡음 분류 절제 연구

본 논문에서 제안한 환자 검출 모델의 경우 공개된 훈련 데이터에서 13091의 평균 대회 점수를 기록했으며, 신호 분할 과정을 제외한 경우는 14800, 사전훈련된 가중치 없이 진행한 경우는 23514, 데이터 증강 기법 없이 진행한 경우는 13726, 인구통계학 정보를 사용하지 않은 경우는 14283의 점수를 기록했다. 환자 검출 모델의 경우 사전 훈련된 가중치의 사용 유무가 성능에 가장 큰 영향을 미쳤으며, 이는 대회 비용 함수에서 44%의 점수차이를 보였다. 다음으로는 신호 분할과 인구 통계 데이터의 사용 유무가 각각 13%와 10%의 영향을 미쳤으며, 데이터 증강은 성능에 4.9%의 영향을 미쳤다.

심잡음 분류 모델의 경우에는 본 논문에서 제안한 모델의 경우 0.759의 평균 가중된 정확도를 기록했으며, 신호 분할을 제외한 경우 0.751, 사전훈련된 가중치를 사용하지 않은 경우 0.661, 데이터 증강을 하지 않은 경우 0.704, 인구통계학적 정보를 사용하지 않은 경우 0.744의 점수를 기록했다. 심잡음 모델에서도 사전 훈련된 가중치의 사용 유무가 가장 큰 영향인 12.9%의 정확도 향상을 이끌어냈으며, 데이터 증강 유무가 7.2%의 향상을, 나머지 두 기법 역시 각각 1~2%의 향상을 이끌어냈다.

이를 통해 대부분의 전처리 기술은 모델의 성능 향상에 기여했음을 확인할 수 있었다. 두 가지 작업에서 훈련 데이터 세트에서의 성능은 사전 훈련된 모델을 사용했는지 여부가 모델에 가장 큰 영향을 미쳤다. 나머지 기법의 경우 두 모델에서의 양상이 달랐는데, 심잡음 분류 모델에서는 데이터 증강 기법이 그 다음으로 성능 향상에 기여했지만,

환자를 분류하는 모델에서는 신호 분할과 인구통계학적 정보가 더 큰 기여를 했다. 반면에 신호 분할 및 인구 통계 데이터는 훈련 데이터에서 심잡음의 성능 개선에 유의미한 영향을 미치지는 못했다. 신호 분할은 훈련 데이터에서 성능에서의 큰 효과가 없었지만, 숨겨진 검증 데이터에서는 13.3%의 점수 향상을 보여 ViT 모델에 대한 PCG 잡음 검출에 신호 분할이 효과가 있음을 보여준다.

제 3 절 대회 데이터셋에 대한 실험 결과

이번 절에서는 본 논문에서 제안한 모델을 PhysioNet 2022 대회에 제출하여 받은 숨겨진 검증 및 평가데이터에 대한 등수를 포함한 결과를 설명한다. 대회 데이터 세트에 대한 심장병 환자 식별 모델의 결과는 표7에 기록했으며, 심잡음 분류 모델에 대한 결과는 표8에 기록되었다. 앞 장에서 설명한 최종 모델을 결과는 5중 교차 검증(5-fold cross validation)을 통해 훈련되었으며, 공개된 훈련 데이터에 대해 가장 높은 성능을 보이는 모델을 공개되지 않은 검증 및 평가데이터 추론에 사용하였다. 최종 제출한 모델은 실격된 팀을 제외하고 심장병 환자 분류에서 39개 팀 중 6위, 심잡음 분류에서 40개 팀 중 21위를 기록했다. 심장병 환자 분류의 경우 훈련 데이터보다 검증데이터에서 더 좋은 점수를 기록했으며, 평가데이터에서는 점수가 소폭 하락했다. 심잡음 환자 분류의 경우 훈련데이터에 비해 검증 및 평가데이터의 점수가 현저히 떨어짐을 확인할 수 있었다.

Training	Validation	Test	Ranking
10615	9903	11943	6/39

표7. PhysioNet challenge 2022 심장병 환자 분류 공식 점수

Training	Validation	Test	Ranking
0.936	0.671	0.69	21/40

표8. PhysioNet challenge 2022 심잡음 분류 공식 점수

표9 및 표10에는 대회에 제출한 최종 모델에 대한 공개된 훈련 데이터셋 및 비공개 검증 및 평가 데이터셋에 대한 결과가 있다. AUROC(Area Under Receiver Operating Characteristic)는 ROC 곡선 아래의 영역의 면적, AUPRC(Area under the Precision-Recall Curve)는 PRC 곡선 아래의 면적을 의미한다. 또한, F1 점수, 정확도, 가중된 정확도 및 대회 비용 점수를 보여주며, 이러한 지표들 또한 심장병 환자 분류의 경우에 훈련 데이터에서의 결과가 가장 잘 나왔지만, 평가 데이터에서의 결과 역시 준수함을 확인할 수 있다. 반면 심잡음 환자 분류의 경우 훈련데이터셋보다 검증, 평가데이터셋에서는 성능이 큰 폭으로 떨어짐을 확인할 수 있었는데, 이는 제안된 모델이 공개된 훈련 데이터에 대해 과적합 되었음을 의미한다. 이는 제출 전 실험이 훈련데이터에서만 이루어져서 제안된 모델의 데이터 증강이 훈련데이터셋에 대해서는 충분하지만, 검증 및 평가데이터에 대해서는 부족하였음을 의미하며, 이를 개선하기 위해서는 더 많은 데이터 증강을 하거나 정규화 기법이 필요하다 생각한다.

Dataset	AUROC	AUPRC	F1 score	Accuracy	Weighted Accuracy	Cost
Training	0.664	0.623	0.581	0.599	0.751	10615
Validation	0.553	0.536	0.468	0.477	0.666	9903
Test	0.605	0.579	0.532	0.562	0.722	11943

표9. 심장병 환자 분류 모델에 대해 제출된 최종 모델의 학습, 검증, 평가 데이터에 대한 결과

Dataset	AUROC	AUPRC	F1 score	Accuracy	Weighted Accuracy	Cost
Training	0.946	0.922	0.934	0.961	0.936	16377
Validation	0.705	0.554	0.617	0.805	0.671	13553
Test	0.717	0.569	0.571	0.769	0.69	17145

표10. 심잡음 분류 모델에 대해 제출된 최종 모델의 학습, 검증, 평가 데이터에 대한 결과

제 5 장 논의

제 1 절 대회 결과에 대한 고찰

본 논문에서는 기존의 심음도 분류 논문들에서 사용하는 방법들에 대해 소개하였으며, 심음도 분류에 처음으로 ViT를 사용하기 위한 전략을 제시하였다. 기존 딥러닝 기반 연구들은 성능이 좋지만 모델의 결과를 볼 수 없는 단점이 있어 본 논문에서는 성능이 뛰어난 모델이자 모델의 결과에 대한 시각화가 가능해 전문가에 해석에 도움을 줄 가능성이 있는 ViT 모델을 심음도 분류에 적용하기 위한 전략을 제시하였다. PhysioNet 2022 challenge에서 전문가의 도움 없이도 심장병 탐지 부분에서 6위를 기록했으며, 이를 통해 본 논문에서 제안하는 전략이 심음도에서의 ViT 모델의 사용하기에 적합하다는 것을 입증할 수 있다. 또한 절제연구를 통해 훈련데이터에서의 신호 분할, 데이터 증강, 인구통계학적 정보, 사전 학습된 가중치의 효과를 검증할 수 있었으며, 제한된 제출을 통해 공개되지 않은 검증 데이터셋에서의 각 기법들의 효과가 모델의 성능을 훈련데이터에 비해 향상시키는 기법임을 확인할 수 있었다. 대회에 제출된 다른 모델들은 대부분 딥러닝 모델을 사용했으며, 그 중에서도 1d-ResNet 기반의 방법이 가장 많이 사용되었다. 다음으로는 대회에서 심장병 환자 분류와 심잡음 분류 모델 모두에 대해 높은 결과를 기록한 방법을 살펴보고자 한다. RNN과 HSMM 기반의 방법[38]이 각 태스크에서 1위와 2위로 평균

1등을 기록했으며, 대조학습(Contrastive learning)을 적용한 방법[39]이 평균 2등을 기록했다.

[38]에서는 전처리 기법으로 심음도를 정규화 한 뒤 로그 스펙트로그램을 계산한 뒤 0~800Hz 구간을 제외한 부분은 제거한다. 잡음 예측을 계산할 때 전체 신호를 고려하기 위해 HSMM을 적용하여 RNN 관측값인 S1, S2, 이완기, 수축기, 심잡음 정보와 HSMM 매개변수(상태 기간 및 전이 행렬)가 주어지면 Springer 기간 종속 Viterbi 알고리즘을 사용하여 심장 소리 신호의 분할을 계산해서 세 가지 종류의 심잡음과 정상 총 4가지로 분류하여 심잡음 유무를 판단했다. 또한 환자 검출 모델에서는 그래디언트 부스팅 의사결정나무(Gradient Boosted Decision Tree)로 환자의 인구통계학적 정보를 자동으로 결합하여 임상 결과를 예측했다.

[39]에서는 심장 소리 신호를 8초단위로 사용했고, 1000Hz의 리샘플링을 진행하여 그대로 사용했다. PCG에 대해 백본망(backbone network)으로 ResNet50을 사용해 지도 대조 학습(supervised contrastive learning)을 적용했다. 이를 위해 심잡음을 잡음 유무로 이진분류를 진행하고, 심잡음 탐지 위치에 대한 정보를 추가로 사용했다. 그렇게 2개의 label로 분리된 결과에서 각각 판단하기 어려운 데이터가 유무에 대한 이진 분류를 진행하여 결과를 얻었다.

제 2 절 어텐션 마스크

ViT를 사용한 심음도 분류 모델의 경우 모델이 이미지의 어느 부분을 참고했는지에 대한 시각화인 어텐션 마스크와 어텐션 지도를 출력할 수 있다. 어텐션 점수가 높은 영역은 어텐션 마스크에서 밝게 출력되며, 원본 스펙트로그램에서 색상으로 세기의 정보를 표현하므로 주의 지도보다 주의 마스크를 원본 스펙트로그램과 함께 보면서 모델의 결과를 해석할 수 있을 것이라 판단했다. 그림19, 그림20, 그림21은 심잡음 모델에서의 스펙트로그램과 어텐션 시각화의 예시이다.

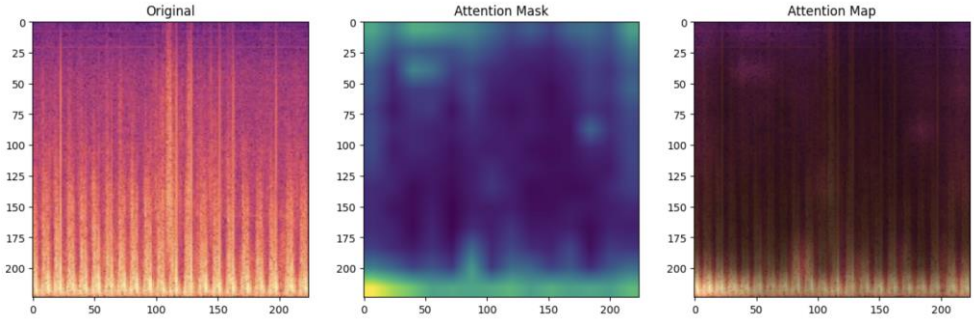


그림 19. 심잡음이 없는 스펙트로그램 및 어텐션 시각화

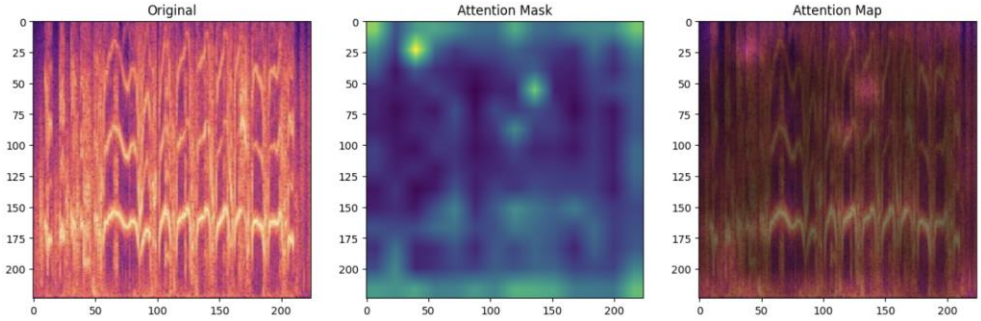


그림 20. 심잡음 유무를 판단할 수 없는 스펙트로그램 및 어텐션 시각화

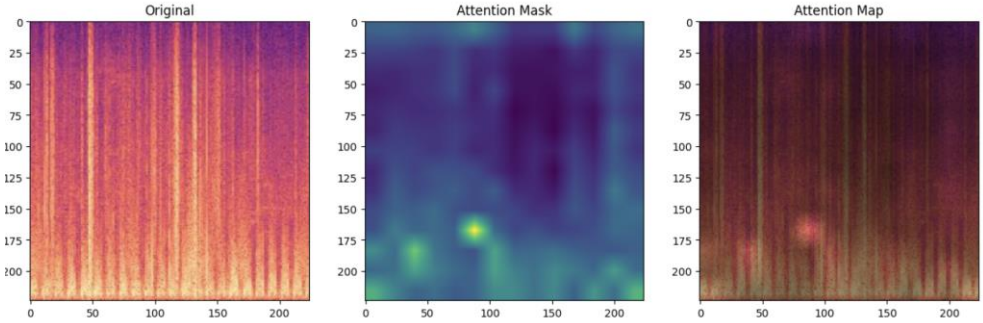


그림 21. 심잡음이 있는 스펙트로그램 및 어텐션 시각화

그리고 그림 22과 그림 23는 심장병 환자를 분류하는 모델의 스펙트로그램 및 어텐션의 시각화이다.

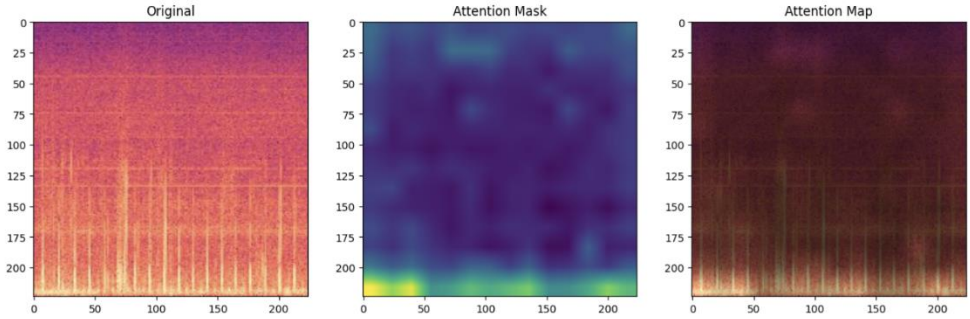


그림 22. 심장병이 없는 사람의 스펙트로그램 및 어텐션 시각화

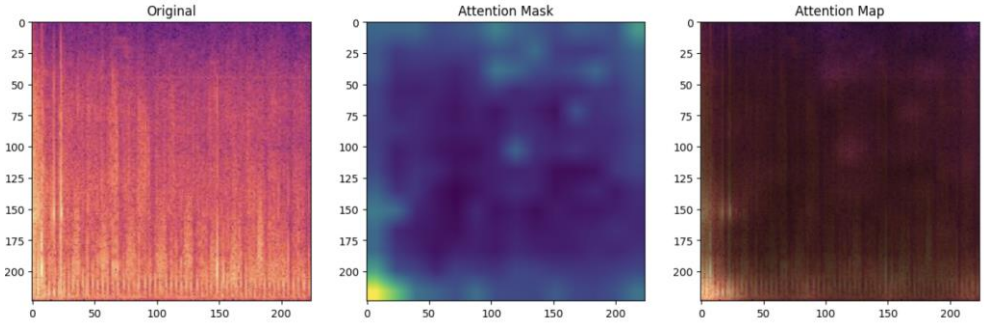


그림 23. 심장병이 있는 사람의 스펙트로그램 및 어텐션 시각화

어텐션 마스크를 통해 모델이 심장 소리에서 어떤 시간에 어텐션을 했는지, 그리고 어떤 주파수에 어텐션을 했는지를 알 수 있으며, 이를 소리데이터를 통해 확인할 수 있다. 심장 소리는 400Hz 이하에 존재하여 스펙트로그램의 y축 상단부의 어텐션은 심장 소리가 아닌 다른 소리에 어텐션을 했음을 알 수 있다. 하지만 심음도에는 청진 시 심장소리 외에도 폐에서 나오는 소리가 섞이며, 폐에서 나오는 소리는 20~1600Hz까지의 대역폭을 갖는다.[40] 따라서 ViT 모델이 심장 소리가 아닌 주파수를 어텐션 한 것이 신체의 다른 소리를 어텐션했을 가능성이 있어 모델에서의 오류라고 단정지을 수 없다. 아래는 전문가에게 해석하는데 도움을 줄 수 있다 생각되는 방법 중 한가지 예시를 제안한다.

1) 모델을 통해 환자의 유무 및 심잡음 유무를 판단한다.
2) 심잡음 혹은 심장병 환자로 의심된다면, 어텐션 시각화를 출력한다.
3) 주의 점수가 높은 구간들을 표시하고, 자동으로 그 시간대의 심음들을 전문가에게 제공한다. 만약 다양한 주파수의 신호가 포함되어 있다면, 어텐션한 주파수 근처로 대역 필터링을 거친 후 제공한다.

이로써 전문가가 모델이 어떻게 판단했는지를 해석하는데 있어 도움을 줄 수 있기를 기대한다. 이러한 어텐션 결과를 기반으로 시간을 기준으로 데이터를 잘라서 소리 및 시각화 된 데이터로 제공할 수도 있고, 특정 주파수들을 기준으로 자를 수도 있을 것이다. 이처럼 시각화와 소리를 제공하는 여러 방법이 있을 것이며, 제시한 방법은 그

중 하나가 될 것이다. 데이터에 대한 전문가인 심혈관 질환 전문의와의 논의 및 협업까지는 시행하지 못했다. 이렇듯, 본 전략을 통하면 모델의 해석에 도움을 줄 수 있는 방법들이 있을 것으로 기대하며, 이에 대한 검증은 추가 연구가 필요하다.

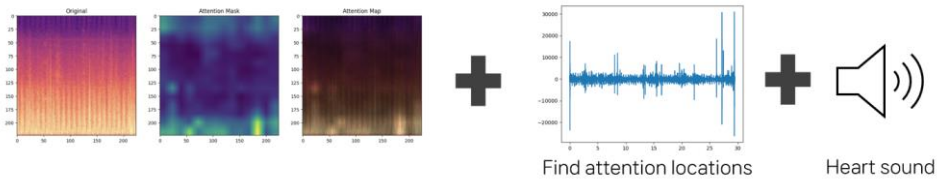


그림 24. 전문가에게 모델의 결과를 해석하는데 도움을 주는 방법 제안

제 3 절 향후 계획

이번 절에서는 본 연구의 향후 계획인 1) 전문가와의 협업을 통한 모델의 해석가능성에 관한 연구, 2) 다른 의료데이터로의 확장 연구, 3) 의료데이터 스펙트로그램 이미지를 위한 데이터 증강 기법 연구에 대해서 설명한다.

앞서 제안한 전문가가 모델의 결과 해석에 도움이 될 수 있는 방법을 검증하기위해 전문가와의 협업이 필요하다. 전문가의 협업이 있다면 사용한 데이터에 대한 특성을 이해하여 더 최적화된 전처리 기법을 찾을 수 있을 것으로 기대하며, 잘못된 레이블 및 정보 그리고 추가적인 데이터를 확보함으로써 모델의 성능을 향상시키는 것을 기대할 수 있다. 의료정보의 특성상 데이터가 적어, 특정 데이터에 대해 과적합 되는 경우가 많으며, 이 때 협업을 통해 의료 데이터를 확보하게 되면 이를 해결할 수 있을 것으로 기대한다.

본 논문에서 제안한 전략은 비슷한 성질을 가진 의료 데이터인 ECG에서의 심혈관 질환 환자의 검출에도 어렵지 않게 적용이 가능하다. 신호 형태의 ECG에서 선택적 전처리인 리샘플링, 필터링을 거친 뒤 신호 분할, 데이터 증강 및 대규모 데이터 셋인 ImageNet으로 사전 학습된 가중치를 사용해 심혈관 질환 환자를 분류할 수 있음을 확인했으며 연구를 하고 있었다. 또한 의료 영상 데이터인 CT 및 MRI 이미지의 경우에도 따로 이미지로 변환하는 과정 없이 데이터에 대한 이해를 토대로 ViT 구조를 적용할 수 있으며, 각 데이터에서 모델이

이미지의 어느 부분을 참고했는지 어텐션 마스크와 어텐션 지도의 출력을 통해 확인할 수 있을 것으로 기대한다.

본 연구에서는 데이터 증강 기법으로 스펙트로그램에 사용되는 시간 마스킹 기법 하나만을 사용했는데, 소리의 성질을 갖지만, 의료 정보를 담는 데이터들에 대한 새로운 데이터 증강 기법들에 대해 연구한다면 더 좋은 성능의 전략을 제안할 수 있을 것으로 기대한다.

제 4 절 한계

이번 절에서는 본 논문에서 제안한 모델 및 실험의 한계에 대해 논의한다. 본 연구에서 사용된 데이터는 대회에서 제공하는 데이터가 유일하여 다른 PCG 데이터에서의 검증이 부족하다. 하지만 공개된 심장 소리 데이터가 부족하며, 각 데이터별 성질이 상이하여 전문가 없이 사용하기에는 어려움이 있다. 또한 대회 데이터셋으로 평가받았지만, 학습에 사용된 공개된 훈련데이터와, 공개되지 않은 검증 및 평가 데이터 사이의 특성이 다르고 더이상 데이터와 결과를 확인할 수 없어 로컬에서의 실험을 검증할 수 없다.

다음으로는 전문가의 부재로 인한 한계를 논의하고자 한다. 전문가와의 협업이 있었더라면 심장병과 심장 소리에 대해 더 최적화된 전처리 기술을 발견했을 수 있다. 또한 PCG를 듣더라도 데이터에 대한 병리학적 이해가 부족하여 자세한 정보를 얻을 수 없었고, 어떤 심장 소리가 환자의 심음 인지, 그리고 심잡음이 있는지 판단하기는 어려웠으며, 인구통계학적 정보에서 잠재적인 이상을 감지하지 못하고 적절한 처리를 하지 못했다. 데이터에 대한 전문가를 통해 제외 기준이나 재라벨링(relabeling)을 수행하고 추가 데이터를 활용했다면 모델의 성능이 더 좋았을 것이다. 제안된 방법의 잠재적인 이점 중 하나는 PCG의 어떤 부분이 각 병리학적 상태에 대해 정보를 제공하는 어텐션 마스크와 어텐션 지도를 출력할 수 있다는 것인데, 이를 검증하고 보완하기 위해서는 전문가의 도움이 필요하다.

제 6 장 결론

본 논문에서는 ViT 구조를 사용한 시각적인 접근을 통해 심음도에서 심장병 환자와 심잡음을 분류할 수 있는 방법을 연구했고 PhysioNet challenge 2022의 결과를 통해 이러한 방법의 유효성을 검증했다. 뿐만 아니라 ViT 모델을 사용했을 때 출력가능한 어텐션 마스크와 지도를 시각화하고 전문가의 해석에 도움을 줄 수 있는 방법을 제안했으며, 본 연구의 향후 계획과 한계에 대해서 논의하였다.

심장병 환자 분류 모델에서는 공개된 학습 데이터를 통한 학습 전략이 공개되지 않는 검증 및 평가데이터에서도 유효했지만, 심잡음 분류 모델에서는 숨겨진 검증 및 평가 데이터의 여러 점수가 훈련 데이터보다 크게 낮은 결과를 보였다. 이는 공개된 데이터로만 학습할 때는 신호 분할과 시간 마스킹 데이터 증강 기법 하나만으로도 충분한 일반화가 되었으나, 공개되지 않은 검증 및 평가 데이터로 추론할 때는 적절하지 못했음을 알 수 있다. 이를 통해 심잡음 분류 모델이 훈련 데이터 세트에 과적합된 것으로 해석할 수 있으며, 정규화 기법들을 적용하여 더 나은 결과를 얻을 수 있다.

본 논문에서 제안된 방법을 사용하면, 심장 소리 및 심장에 대한 전문가 없이도 좋은 성능을 내는 모델을 만들 수 있었으며, 심장 소리 외에도 이를 스펙트로그램, 어텐션 마스크를 출력할 수 있다. 어텐션 점수가 높은 부분들을 추출하고 필터링해서 제공한 결과를 추후 전문가들과의 협업을 통해 모델 해석에 도움을 기대해 볼 수 있다. 이를

통해 본 논문에서 제안하는 방식이 전문가들이 모델의 결과를 이해, 해석 및 개선할 수 있는 기회를 열 수 있다고 믿는다. 이러한 전략이 의료데이터에서의 ViT 기반 연구들에 도움이 될 수 있을 것으로 기대한다.

참고 문헌

[1] Lee, Y., Lee, G., Na, S. D., Seong, K. W., Cho, J. H., & Kim, M. N. (2016). Heart Murmur Detection Algorithm based on Spectral Flatness. Journal of Korea Multimedia society, 19(3), 557-566.

[2] Andreoli, T., Behram, R., Bhattacharya, B., Borer, W., Canellos, G., & Flye, M. (2007). Dorland’s Medical Dictionary for Healthcare Consumers.

[3] “The top 10 causes of death”, World Health Organization, last modified 9 December 2020, accessed 24 November 2022, https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[4] PhysioBank, P. (2000). PhysioNet: components of a new research resource for complex physiologic signals. Circulation, 101(23), e215-e220.

[5] Reyna, M. A., Kiarashi, Y., Elola, A., Oliveira, J., Renna, F., Gu, A., ... & Clifford, G. D. (2022). Heart murmur detection from phonocardiogram recordings: The george b. moody PhysioNet challenge 2022. medRxiv.

[6] Springer D.B., Tarassenko L., Clifford G.D. Logistic Regression-HSMM-Based Heart Sound Segmentation. IEEE

Trans. Biomed. Eng. 2016;63:822. doi: 10.1109/TBME.2015.2475278.

[7] Chen, W., Sun, Q., Chen, X., Xie, G., Wu, H., & Xu, C. (2021). Deep learning methods for heart sounds classification: a systematic review. Entropy, 23(6), 667.

[8] Li, S., Li, F., Tang, S., & Xiong, W. (2020). A review of computer-aided heart sound detection techniques. BioMed research international, 2020.

[9] Zeinali, Y., & Niaki, S. T. A. (2022). Heart sound classification using signal processing and machine learning algorithms. Machine Learning with Applications, 7, 100206.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[11] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[12] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[13] Yoganathan, A. P., Gupta, R., Corcoran, W. H., Udwadia, F. E.,

Sarma, R., & Bing, R. J. (1976). Use of the fast Fourier transform in the frequency analysis of the second heart sound in normal man. Medical and biological engineering, 14(4), 455-460.

[14] Liu, C., Springer, D., Moody, B., Silva, I., Johnson, A., Samieinasab, M., ... & Clifford, G. D. (2016). Classification of Heart Sound Recordings-The PhysioNet Computing in Cardiology Challenge 2016. PhysioNet.

[15] Antink, C. H., Becker, J., Leonhardt, S., & Walter, M. (2016, September). Nonnegative matrix factorization and random forest for classification of heart sound recordings in the spectral domain. In 2016 Computing in Cardiology Conference (CinC) (pp. 809-812). IEEE.

[16] Li, J., Ke, L., & Du, Q. (2019). Classification of heart sounds based on the wavelet fractal and twin support vector machine. Entropy, 21(5), 472.

[17] Arora, V., Leekha, R., Singh, R., & Chana, I. (2019). Heart sound classification using machine learning and phonocardiogram. Modern Physics Letters B, 33(26), 1950321.

[18] Deng, M., Meng, T., Cao, J., Wang, S., Zhang, J., & Fan, H. (2020). Heart sound classification based on improved MFCC

features and convolutional recurrent neural networks. Neural Networks, 130, 22–32.

[19] Latif, S., Usman, M., Rana, R., & Qadir, J. (2018). Phonocardiographic sensing using deep learning for abnormal heartbeat detection. IEEE Sensors Journal, 18(22), 9393–9400.

[20] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90.

[21] Singh, S. A., Majumder, S., & Mishra, M. (2019, May). Classification of short unsegmented heart sound based on deep learning. In 2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) (pp. 1–6). IEEE.

[22] Flanagan, J. L. (1972). Speech synthesis. In Speech Analysis Synthesis and Perception (pp. 204–276). Springer, Berlin, Heidelberg.

[23] Krishnan, S. (2021). Biomedical signal analysis for connected healthcare. Academic Press.

[24] Zgallai, W. A. (Ed.). (2020). Biomedical Signal Processing and Artificial Intelligence in Healthcare. Academic Press.

[25] Diniz, P. S., Da Silva, E. A., & Netto, S. L. (2010). Digital

signal processing: system analysis and design. Cambridge University Press.

[26] Sejdic, E., Djurovic, I., & Stankovic, L. (2008). Quantitative performance analysis of scalogram as instantaneous frequency estimator. IEEE Transactions on Signal Processing, 56(8), 3837-3845.

[27] Arora, V., Verma, K., Leekha, R. S., Lee, K., Choi, C., Gupta, T., & Bhatia, K. (2021). Transfer learning model to indicate heart health status using phonocardiogram.

[28] Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. arXiv preprint arXiv:2005.00928.

[29] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48.

[30] Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

[31] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144.

[32] Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger,

J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 321, 321-331.

[33] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., & Beyer, L. (2021). How to train your vit? data, augmentation, and regularization in vision transformers. arXiv preprint arXiv:2106.10270.

[34] Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779.

[35] Oliveira, J., Renna, F., Costa, P. D., Nogueira, M., Oliveira, C., Ferreira, C., ... & Coimbra, M. T. (2021). The CirCor DigiScope dataset: from murmur detection to murmur classification. IEEE journal of biomedical and health informatics, 26(6), 2524-2535.

[36] Raza, A., Mehmood, A., Ullah, S., Ahmad, M., Choi, G. S., & On, B. W. (2019). Heartbeat sound signal classification using deep learning. Sensors, 19(21), 4819.

[37] Ho, Y., & Wookey, S. (2019). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. IEEE Access, 8, 4806-4813.

[38] (In press) McDonald, A., Gales, M., & Agarwal, A. Detection of Heart Murmurs in Phonocardiograms with Parallel Hidden Semi-Markov Models.

[39] (In press) L' ubomır Antoni, E. B., Bugata, P., Bugata Jr, P., Gajdoš, D., Hudák, D., Kmecová, V., ... & Vozáriková, G. Murmur Identification Using Supervised Contrastive Learning.

[40] Nersisson, R., & Noel, M. M. (2017). Heart sound and lung sound separation algorithms: a review. Journal of medical engineering & technology, 41(1), 13-21.

[41] (In press) Kim, J., Park, G., & Suh, B. Classification of Phonocardiogram Recordings using Vision Transformer Architecture.

Abstract

Classification of Phonocardiogram Recordings Using Vision Transformer Architecture

Joonyeob Kim
Department of Intelligence and Information
The Graduate School
Seoul National University

Diagnosis of cardiac disease starts with measuring heart sound. It provides information about the cardiac condition and function, which is important medical information. Diagnosis by listening to a heart sound through a stethoscope is less accurate than a detailed examination, but it is essential for diagnosis because it is easy and inexpensive. A healthy person can hear a regular and clear heartbeat, but in an unhealthy person, a murmur can be heard along with the heartbeat. Such a noise is called a heart murmur, and if the characteristics of the heart murmur and the location at which the noise is heard are used, heart disease can be judged by this alone. Phonocardiograms (PCG) data created by recording heart sound data can detect whether a person has a heart disease, detect abnormal heart sounds, etc. Developing models is an important research topic

for people suffering from heart disease in countries with limited expertise and capital, such as developing countries.

The method of modeling medical data can be divided into a method using a classical machine learning-based model and a method using a deep learning-based model. Since machine learning-based medical data modeling requires the direct extraction of features, the researcher's prior knowledge and pre-processing methods have a great influence on the data used. On the other hand, in deep learning-based models, the influence of the researcher's prior knowledge and preprocessing method on the data used is relatively low because the model learns and extracts even these characteristics. Traditionally, many machine learning models have been used for medical data modeling, but recently, in the medical data field, studies are underway to improve the performance of existing medical data modeling by using a model with good performance in the deep learning field. shows better performance than However, most deep learning-based models cannot interpret the results, so it is difficult to help experts in diagnosis.

In the case of the Vision Transformer structure used in this paper, self-attention is included, which calculates an attention score that helps to understand the result and converts it into an attention mask, and it is a model with the advantage of showing high performance in

the image classification task in the computer vision field.

In this paper, proposing a model that detects the presence or absence of murmurs from multiple heart sound recordings from multiple auscultation locations, as well as detecting the clinical outcomes from phonocardiogram well. A visual approach was used to aid in the interpretation of results and higher classification performance. The proposed model converts heart sound signals into spectrograms without requiring resampling or signal filtering, and infers cardiac noise and clinical outcomes by receiving the image with the patient's demographic information. For the clinical outcome identification task on the test data, it shows a Challenge cost score of 11943. The result shows a weighted accuracy score of 0.69 for the murmur detection classification on the test data. In addition, the model includes an attention mask that allows visualization of which part was viewed and judged.

Keywords : Phonocardiogram, Heart murmur, Vision Transformer, Spectrogram, Attention mask
Student Number : 2021-29046