공학석사 학위논문

심전도 분석을 위한 딥러닝 기반 다중 클래스 분류와 이상치 탐지: 데이터 불균형과 임상적 해석

Deep Learning-Based Multi-Class Classification and Anomaly detection for ECG analysis: Addressing Data Imbalance and Clinical interpretation

울 산 대 학 교 대 학 원

의 과 학 과

최 상 훈

심전도 분석을 위한 딥러닝 기반 다중 클래스 분류와 이상치 탐지: 데이터 불균형과 임상적 해석

지 도 교 수
주 세 경

이 논문을 공학석사학위 논문으로 제출함

2024년 2월

울 산 대 학 교 대 학 원
의 과 학 과
최 상 훈

최상훈의 공학석사학위 논문을 인준함

심사위원 남기병 인

심사위원 주세경 인

심사위원 김준기 인

울 산 대 학 교 대 학 원

2024년 2월

목 차

국문 요약

4차 산업혁명 발전에 따라, 최근 의료 데이터의 빅데이터 형성에 따른 다양한 연구들이 진행되고 있다. 특히, 의료영상분야에서 Magenetic Resonance Imgage(MRI), Computed Tomography(CT), Ultrasonud image등 영상 데이터를 활용하여 각종 질환을 빠르고 정확한 진단을 위해 다양한 분석 방법의 연구가 진행되었다. 의료영상 데이터를 시작으로 Electrocardiogram(ECG), Photoplenthysmography(PPG) 등 시계열 데이터를 활용한 연구들도 활발히 진행되고 있다.

최근에는 딥러닝(Deep learning) 기술의 적용에 따른 많은 발전이 이루어지고 있다. 딥러닝 기술에는 대표적으로 형태적인 특징을 추출하는 Convolutional Neural Network(CNN) 와 시간적인 특징을 추출하는 Recurrent Neural Network(RNN) 이 있다. 다양한 신경망을 통해 학습 방법에 따라 분류, 생성과 같은 연구들이 진행되고 있다. 학습 방법에는 지도학습, 비지도학습이 있다. 지도학습은 학습을 진행 할 때 입력에 대한 정답을 알려주는 방식으로 학습을 진행하는 방법이고, 비지도 학습은 정답을 알려주지 않고 데이터의 특징을 군집화하고 새로운 데이터에 대해서 결과를 예측하는 방법이다.

의료 인공지능의 기술이 발전됨에 따라 한계점이 존재하는데 대표적인 한계점으로 데이터 불균형과 임상적 해석에 따른 문제점있다. 먼저 데이터 불균형은 의료데이터에서 흔히 발생하는 현상이다. 실제 의료데이터에서 대부분은 정상인 데이터가 많고, 특정 질환이나 질병에 대한 데이터는 적기 때문에 데이터의 불균형은 불가피한 현상이다. 이러한 데이터를 가지고 딥러닝 학습을 진행하게 되면, 데이터의 지배적인 클래스에 대해서 편중화된 모델로 학습이 되고 이는 소수 클래스에 대한 예측 및 진단의 성능이 현저하게 떨어지게 된다. 다음으로 임상적 해석에 따른 문제점은 딥러닝 분석 방법에 대한 불확실성에서 발생한다. 딥러닝 모델의 알고리즘의 복잡한 계산과정을 이해할 수 없기 때문에 흔히 “블랙박스”라는 특징을 가지고있고, 이로 인해 분석에 대한 임상적인 의미가 불분명하다.

따라서 본 연구는 위의 문제점을 해결하기 위해, 심전도 데이터를 가지고 딥러닝 분석을 진행하고, 분석에 따른 데이터 불균형을 해소하기 위해 4가지 실험을 통해 성능을 비교하여 최적의 성능 및 방법을 찾고, 임상적

해석에 따른 문제점은 정상 심전도 데이터의 주요 세가지부분을 비지도 학습 방법을 사용하여 학습을 진행하고, 심방세동 심전도 데이터의 이상치를 계산하여 실제 의료진의 진단 방법 메커니즘을 모방하여 딥러닝으로 임상적 해석을 진행하여 해결한다.

데이터불균형 파트에서는 실제 병원 데이터 7355명의 다중리드 심전도를 사용한다. 이 파트에서의 분류기는 레즈넷(Resnet) 모델을 사용하여 지도 학습 방법으로 학습을 진행하여 8가지의 질환을 분류한다.

임상적 해석파트에서는 PTB-XL 9042명, 중국 7199명의 정상과 심방세동환자 단일리드 심전도를 사용한다. 이 파트에서는 정상 심전도의 Q파이전(preQ), QRS복합체(QRS), S파 이후(PostS) 로 나누어 오토인코더 모델을 사용하여 학습을 진행하고, 심방세동 환자의 심전도로 모델 평가를 진행하여 정상 심전도와 심방세동 심전도의 이상치를 계산하여 정상과 심방세동을 분류한다.

연구 결과는 데이터 불균형 파트에서는 정확도와 F1 점수로 모델평가를 진행하였다. 결과는 각각 0.96 으로 Focal 손실함수를 사용한 실험이 데이터 불균형 환경에서 가장 좋은 성능 보였다.

임상적 해석 파트에서는 각 주요부분인 PreQ,QRS,PostS 각 모델에 대한 이상치 점수는 Mean Squared Error(MSE)로 계산을 진행했고, 심방세동과 정상 환자의 분류는 각 그룹의 이상치점수를 가지고 정확도와 F1 점수로 평가를 진행했다. 결과는 PreQ에서 정상과 심방세동 이상치점수 각각 0.00126, 0.0182로 가장 큰 차이를 보였고, 분류 결과도 F1 점수와 정확도 각각 0.92로 가장 좋은 성능을 보였다.

그림 목차

표 목차

식 목차

서론

1. 연구배경

빅데이터의 형성으로 4차 산업혁명의 발전함에 따라 의료계에서도 인공지능기술의 연구가 세계적으로 진행되고 있다. 의료 인공지능을 적용한 의료기기들의 규모가 시간이 지남에 따라 증가율이 크게 증기하고 있다.(그림1).[1] 이와 더불어, 인공지능 관련 연구도 2017년 이후로 증기율이 급격하게 증가하고 있다.[2]

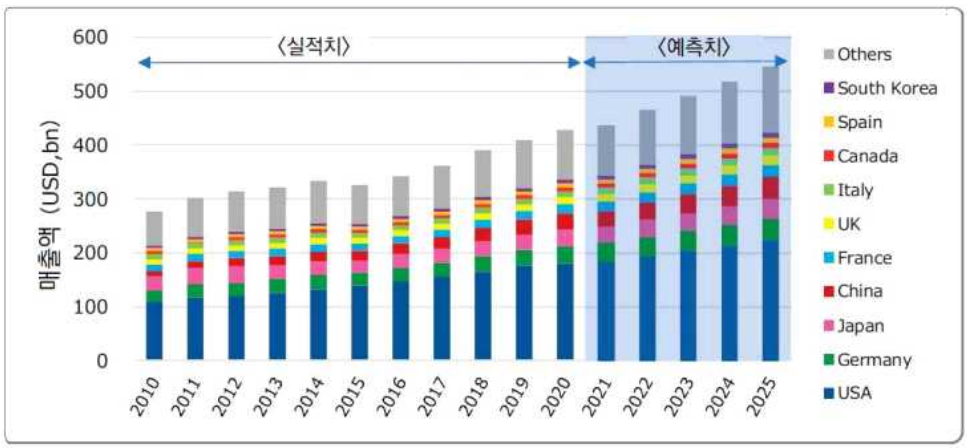

그림 1 세계 인공지능 의료산업 규모

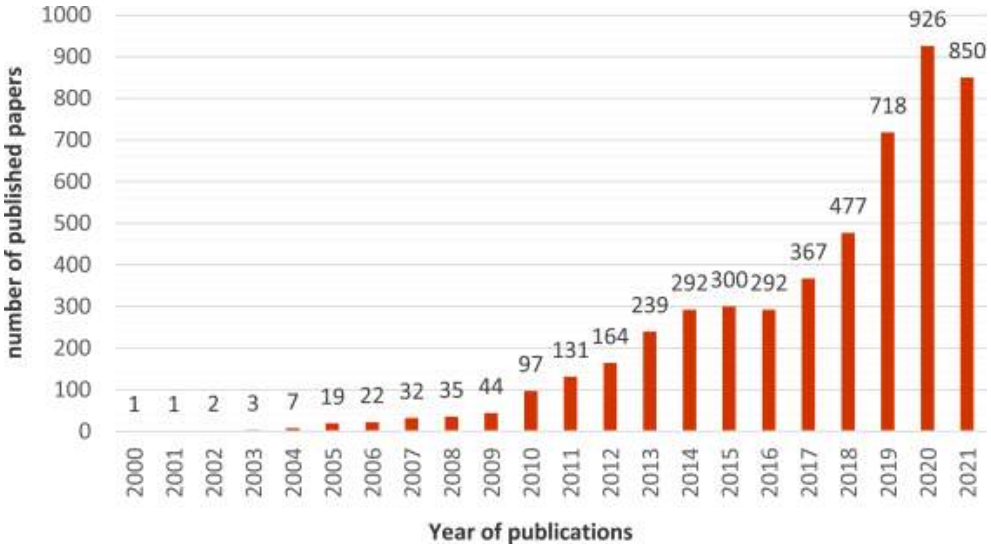

그림 2 헬스케어 관련 논문 출간 추이

모바일 시대 이전, 의료기술은 주로 보조적인 기술로 스탠트, 임플란트와 같은 기술로 알려져 왔지만, 최근에는 스마트폰과 웨어러블 기기의 발전과 함께 인공지능을 활용하여 많은 영역에서 복잡한 과제를 수행하고 있다. 질병의 진단을 넘어 질병 발생 시점 이전의 데이터를 활용하여 질병을 예측하여 미리 예방하는 기술까지 발전되어 왔다. 최근 심전도 측정 기기에 인공지능을 결합한 기기들이 등장하고 있는데, 심전도를 측정하면 측정된 데이터를 기반으로 질병을 진단할 수 있는 기기이다.(그림3)

그림3 휴대용 심전도 기기를 활용한 AI 기술

최근에 이러한 기기들의 적용된 인공지능 모델들이 다양하게 개발되고 있다. Convolutional neural network (CNN, 합성곱 신경망), Recurrent neural network(RNN, 순환신경망) 모델 중 하나인 Long Short Term Memory(LSTM)으로 심전도를 포함한 생체신호의 특징을 추출하여 인공지능 학습을 진행하여 지도 학습을 통해 특정 질환을 분류하는 모델이 개발되어 왔다.[3] 데이터의 라벨링 작업이 시간소요가 많이 들기 때문에 최근에는 비지도 학습도 연구가 많이 되고 있습니다. 특히 LSTM을 기반으로 오토인코더 모델을 학습하여 입력 데이터의 시간적 정보를 바탕으로 신호를 재구성하는 모델을 활용하여 정상 데이터를 학습시켜 비정상 데이터를 탐지하는 이상치 탐지에 대한 연구가 한 예가 될 수 있다.[4]

의료인공지능이 다양한 분야에서 적용됨에 따라 발생하는 문제점이 대표적으로 데이터 불균형과 인공지능 모델의 판단 결과에 따른 임상적인 해석의 모호함이다. 데이터 불균형 문제는 특히 의료데이터에서 특정 카테고리에 데이터가 몰려있어 다수 클래스와 소수 클래스가 발생할 수 있다. 정상환자의 데이터의 경우가 다수 클래스에 속하고 상대적으로 적게 발생하는 질병에 대한 소수클래스가 이에 속할 수 있다. 의료 데이터의 불균형은 인공지능 모델 학습에 치명적인 문제가 발생한다.(그림4) 인공지능 모델이 다수 클래스에 편향적으로 학습이 진행되기 때문에 소수클래스에 대한 성능이 낮아질 수 있다.

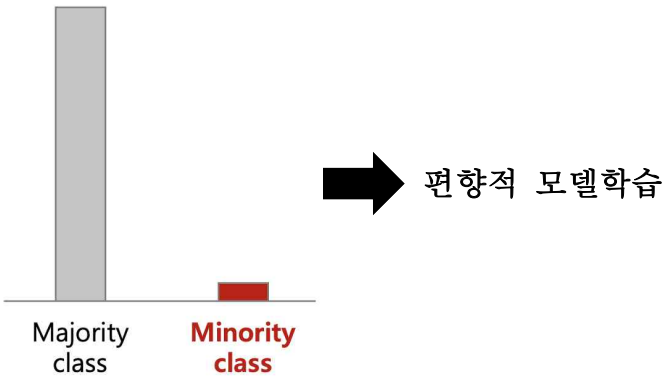

그림 4 데이터 불균형 문제

따라서 최근 많은 데이터 불균형 문제를 해결하는 연구들이 활발히 진행되고 있다. 대표적인 방법에는 소수클래스에 대한 데이터 증강 기법을 사용하여 인위적인 데이터를 생성하여 데이터의 균형을 맞춰 학습을 진행하는 방법이다. 생성 모델로는 Generative adversarial network(GAN, 적대적 신경망), SMOTE 등 인위적인 데이터를 생성하는 모델이 주로 사용된다.[5]
하지만 이렇게 인위적인 데이터를 생성하는 것이 인공지능 모델의 성능을 크게 향상시켰음에도 의료계에서 실제 환자에서 측정된 데이터가 아니기 때문에 학습된 모델에 대한 신뢰성 및 데이터의 규제에 대한 한계점을 지적하고 있다.[6]

임상적 해석의 모호함은 인공지능 모델의 예측하는 과정에 대한 설명이 부족하여 임상적인 판단의 근거가 부족한것을 말한다.(그림5) 인공지능 모델은 학습되는 과정이 모델의 깊이가 깊어짐에 따라 학습 파라미터(Parameter)가 많아지면서, 모델의 알고리즘이 굉장히 복잡해지고 이로인해 “블랙박스(Black Box)”라는 한계점을 지닌다.

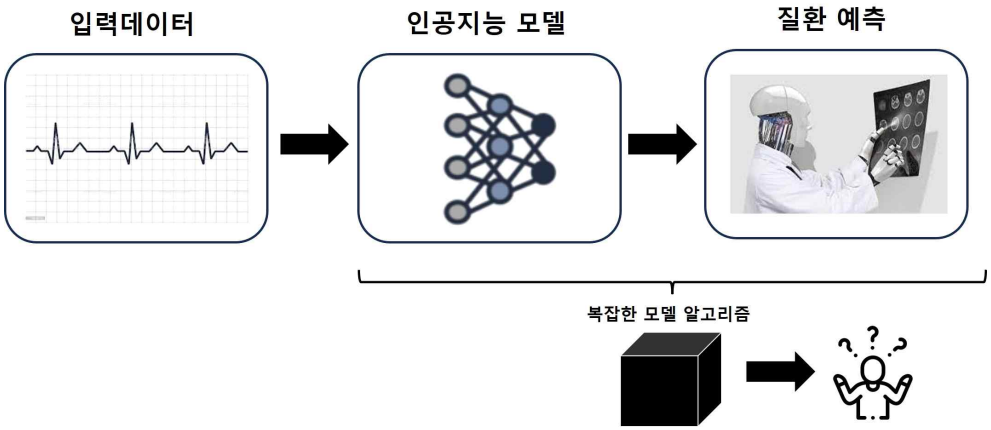

그림 5 인공지능 모델의 임상적 해석의 모호성

이러한 한계점을 극복하기 위해, 최근 설명 가능한 인공지능 모델 개발이 활발히 이루어지고 있다. 특히, Class activation map (CAM) 방법을 활용하여 모델이 입력 데이터의 어느 부분을 보고 예측을 했는지를 시각적으로 표현하는 방법과 Attention 기반 모델을 학습시켜 입력데이터의 특성 중 가장 중요한 부분을 선택적으로 초점을 맞춰 학습을 진행하여 설명 가능한 모델을 학습하는 방법이 있다.

하지만 이러한 방법들은 학습데이터에 의존적인 경우가 많아 모델의 판단 기준이 무작위성의 특성을 띌 수 있어 이러한 비일관적인 판단은 의료계에서는 치명적으로 작용할 수 있다.

본 연구는 의료 인공지능 적용에 따라 발생할 수 있는 문제점을 다루고 이를 해결하는 방법에 대해 고안한다. 연구에서 사용되는 의료데이터는 심전도 데이터를 사용한다. 연구는 데이터 불균형 해소 연구와 임상적 해석이 가능한 딥러닝 모델 연구로 나뉘어 진행한다.

데이터 불균형 해소 파트(Part)에서는 데이터 증강기법을 사용하지 않고, 제한된 데이터 내에서 모델의 손실함수에 가중치를 주는 클래스 가중치 방법, 포칼로스 손실함수방법과 데이터셋의 균형을 맞춘 데이터셋 형성, 하위클래스 형성 총 4가지 방법으로 다중 리드 심전도 10초 데이터를 가지고 딥러닝 모델 학습을 통해 성능을 비교하고 최적의 방법을 확인한다.

임상적인 해석 파트에서는 단일리드 심전도의 단일비트에서 임상적으로 의미가 있는 Q파 이전신호 (PreQ), QRS 복합체(QRS), S파 이후신호 (PostS) 로 나누어 정상 심전도의 세 부분을 오토인코더 모델에 학습시켜 심방세동의 세부분의 이상치를 계산하여 두 클래스의 각 세그먼트별 이상치의 차이를 기준으로 두 그룹을 분류하고, 심전도의 세그먼트별 모델의 이상치점수를 기반으로 임상적으로 설명가능한 모델을 제시한다.

2. 인공지능과 딥러닝 (Artificial Intelligence and Deep Learning)

인공지능(Artificial Intelligence)이란 인간이 지닌 다른 동물들과는 구별되는 지적인 능력인 생각, 다른 말로 표현하면 사고나 학습을 통한 능력을 컴퓨터를 통해 구현되는 기술이다. 인간은 학습을 통해 자연과 물체를 인식하고, 다양한 현상에 대해 판단하고 추론하고 이에 따른 문제 해결하는 과정을 이해하고 분석하는 학문 분야이고 최종적으로 인간의 지적능력의 과정을 기계로 실현하는 것이 인공지능의 목표이다. 이 개념이 최초로 등장한 시점은 1943년 워렌 맥클론(Warren McCulloch)과 월터 피츠(Walter Pitts)에 의해 발표된 논문이다. 인간의 지적 능력을 발현하기 위해 뇌의 현상에 대한 복잡한 네트워크를 논리적인 어느 하나의 네트워크로 표현이 가능하다고 주장했으며 이 주장의 시작으로 인공신경망(artificial neural network) 이라는 단어가 탄생하게 되었다.

인공지능 발전의 시작은 퍼셉트론(perceptron) 이란 알고리즘이다. 이 알고리즘은 1957년에 연구자 프랑크 로젠블라트(Frank Rosenblatt)가 처음 제시한 개념이다. 퍼셉트론은 다수의 입력을 받아 각 입력에 대한 가중치를 주어 하나의 입력을 도출한다.(그림6) 가중치는 다수의 입력 신호의 중요도에 따라 바뀔 수 있는 값이고 이 가중치(W)에 따라 입력이 앞으로 얼마나 전달될지가 결정된다.

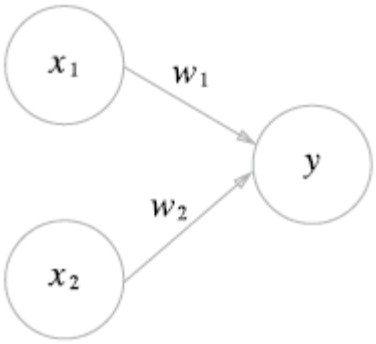

그림 6 퍼셉트론 구조

$$y = \begin{cases} 0 (w_1 x_1 + w_1 x_1 \leq \theta) \\ 1 (w_1 x_1 + w_1 x_1 > \theta) \end{cases}$$

식 1 퍼셉트론 공식

이 이론은 단순 선형적인 문제에 대한 분류는 가능하지만 복잡한 문제 즉, 비선형적인 문제에 대한 분류에서 한계점이 들어나게 된다. 한계점을 해결하기 위해 1986년 퍼셉트론을 다중으로 쌓은 다층 퍼셉트론 (Multi Layer Perceptron, MLP) 의 등장으로 비선형 문제를 분류할 수 있게 되었다. 등장 이전에는 컴퓨터의 계산 능력 및 데이터의 부족으로 인공지능 발전의 중지와 함께 연구가 더 이상 진행되지 않았다. 하지만 다층 퍼셉트론 계산이 가능해지면서 이와 같은 문제가 해결되었지만, 또다른 문제인 단순 다중 신호가 아닌 더욱 입력 데이터가 음성인식, 이미지 분류와 같은 이 이상 복잡한 문제에 대한 문제가 발생 했고, 이 문제를 해결하기 위해 나온 개념이 우리가 흔히 알고 있는 머신 러닝 (Machine Learning) 이다.

머신 러닝은 다중 퍼셉트론 입력 전에 복잡한 신호를 대표하는 특징을 추출하는 단계가 추가된다. 추출된 특징을 기반으로 분류, 예측을 하는 알고리즘이고, 궁극적으로 데이터의 특징 자체를 특정 알고리즘을 통해 학습을 진행하고, 데이터의 가장 대표하는 특징을 추출하는 것이 목표이다. 1990년 대량의 데이터를 다룰수 있는 컴퓨터가 개발되면서 머신러닝의 발전은 눈부시게 이루었다. 대량의 데이터의 특징을 좌표를 변환하거나, 이동, 주파수 영역 해석 등 대표하는 특징을 추출하고 이를 학습 데이터로 사용한다. 이를 기반으로 랜덤 포레스트 (random forest), 서포트 벡터 머신(support vector machine) 등 대표적인 모델이 있다. 하지만 머신러닝의 가장 큰 문제점은 데이터의 특징추출을 사람이 직접 수행해야 하고 추출 방법도 또한 천차만별이기 때문에 전부 적용하기에는 시간이 너무 오래걸린다는 점이다. 그래서 나온 개념이 딥러닝 이다(그림7).

그림 7 인공지능 관계도

딥러닝은 머신러닝에 포함되는 개념인데, 특징 추출 부분을 자동적으로 해주어 머신러닝의 문제점을 보완한 알고리즘이다. 퍼셉트론으로 이루어진 레이어(Layer)를 연속으로 쌓아 데이터의 특징을 자동적으로 추출하고 추출될 때 적용되는 가중치값을 모델의 정확한 출력을 위해 조절한다(그림 8).

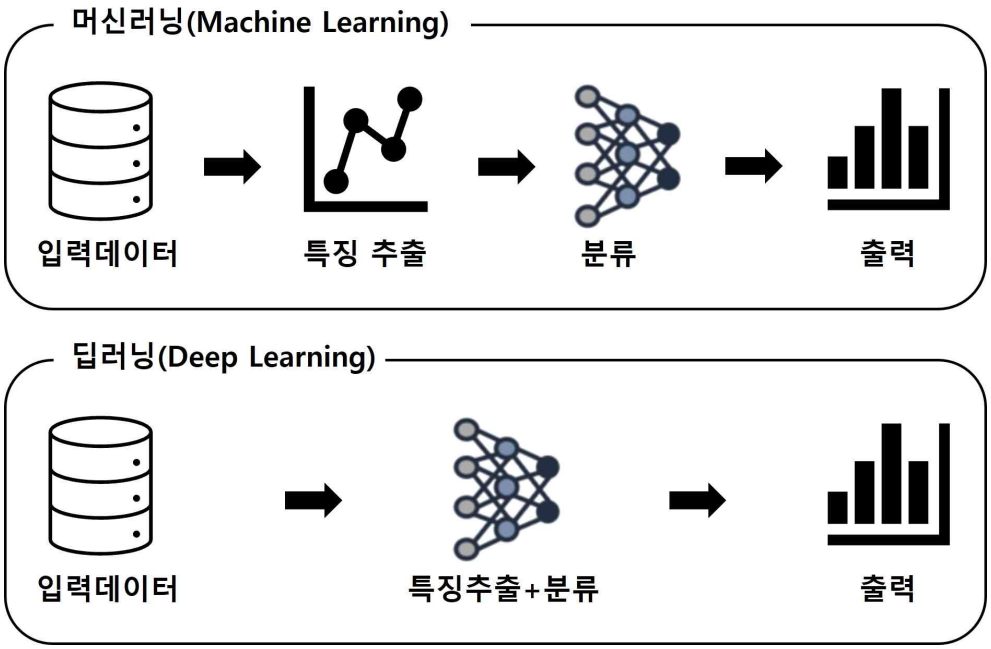

그림 8 머신러닝과 딥러닝 차이점

따라서 딥러닝 모델은 머신러닝과 다르게 직접 데이터의 특징을 추출할 필요가 없다. 딥러닝 모델은 2012년에 이미지 인식 대회인 Image Large Scale Visual Recognition Challenge(ILSVRC)에서 가장 높은 성능을 보인 AlexNet 모델이 최초 딥러닝 모델로 선보이게 된다. 모델에서는 1980년대에 제안된 합성곱 신경망 (Convolutional Neural Network, CNN)의 개념을 기반으로 하여 이미지의 특징을 자동적으로 추출하고 이를 통해 분류를 해주는 모델이다. 이 모델을 기점으로 모델의 깊이 즉, 레이어의 숫자가 늘어나고 이에 따른 계산량도 폭발적으로 늘어가게 된다. 데이터의 수가 많이지고 모델의 깊이에 따른 연산량이 증가함에 따라 이를 처리할 수 있는 컴퓨팅 기술의 발전과 함께 딥러닝은 이미지 인식 뿐만 아니라 의료계, 자율주행, 게임 등 다양한 산업에서 엄청난 영향력을 준 기술이다.(그림9)

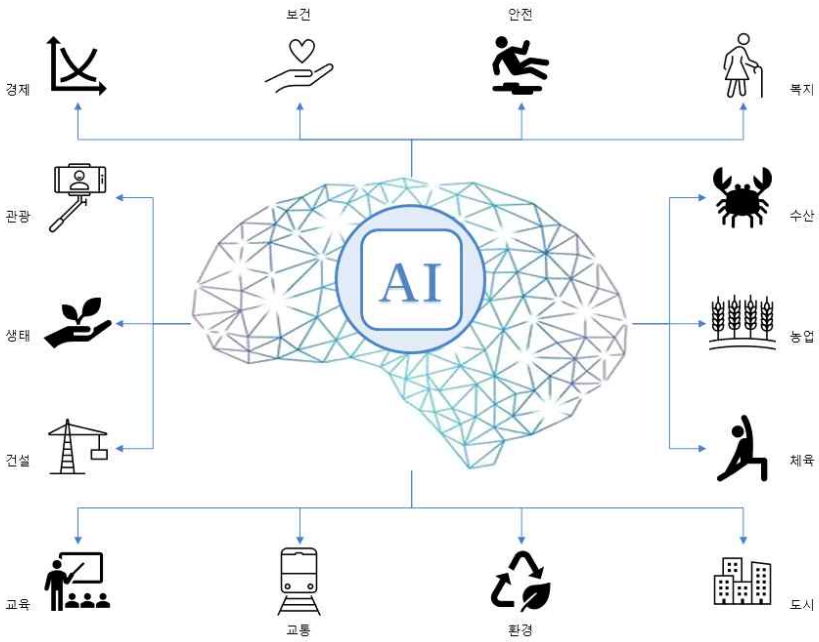

그림 9 여러 산업에서의 인공지능 활용
[출처: 서울대학교 아시아연구소]

3. 딥러닝 모델의 손실함수와 활성화 함수

딥러닝 모델 학습의 기반은 손실함수(Loss function)에 있다. 손실 함수는 경제 분야에서 자주 사용되는 단어인데, 실제 모델이 예측한 값과 실제 값의 차이가 얼마나 나는지를 숫자로 표현한 함수이다. 즉, 이 손실함수 값이 크다는 것은 그 만큼 모델의 레이어 값의 오차가 크다는 것을 의미하고 반대로 값이 작으면 오차오 작다는 것을 의미한다. 따라서 딥러닝 모델 학습의 궁극적인 목표는 손실함수의 값을 최소로 만드는것이고 이는 모델의 최적의 가중치(w) 와 편향(b)을 찾는 것이다.

손실함수 딥러닝 모델의 목표가 분류(Classification)모델이냐 회귀(Regression)이냐에 따라 다른 종류의 함수가 쓰인다.(표1,표2)

이름	식
평균 제곱 오차 (Mean Squared Error, MSE)	$MSE = \frac{1}{n}\sum_{k=1}^{n}(y_k - \widehat{y_k})^2$
평균 절대 오차 (Mean Absolute Error, MAE)	$MAE = \frac{1}{n}\sum_{k=1}^{n}(y_k - \widehat{y_k})$
평균 제곱근 오차 (Root Mean Square Error, RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - \widehat{y_k})}$
결정 계수 (R square, R^2)	$R^2 = 1 - \frac{\sum_{k=1}^{n}(y_k - \hat{y}_k)^2}{\sum_{k=1}^{n}(y_k - y'_k)^2}$

표 1 회귀 모델의 손실함수

이름	식
이진 크로스 엔트로피 (Binary Cross Entropy)	$BCE = -\sum_{i=1}^{2} t_i \log(s_i)$
단정적 크로스 엔트로피 (Categorical Cross Entropy)	$CCE = -\sum_{j=0}^{M}\sum_{i=0}^{N} (y_{ij} * \log(y'_{ij}))$
힌지 로스 (Hinge Loss)	$HingeLoss = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$

표 2 분류 손실함수

우선 분류 모델에서는 대표적으로 크로스 엔트로피 (Cross Entropy) 를 사용하고 회귀모델에서는 평균 제곱 오차 (Mean Squared Error)를 사용한다(표1,표2). 손실함수는 모델에 학습되는 데이터의 형태에 따라 가장 적합한 함수를 결정하여 학습을 진행해야한다. 딥러닝 모델의 가중치는 학습 초기에는 임의의 값으로 지정되기 때문에 선택한 손실함수를 기반으로 학습이 진행될 때마다 가중치가 업데이트 되어 최적의 값을 얻게 된다. 활성화 함수는 추출된 특징의 분포를 함수에 따라 수치의 범위를 정하는 함수이다.

활성화 함수는 각 레이어의 특징을 추출하는 모듈에서는 추출된 특징에서 필요없는 노이즈 특증이 있을 수 있다. 따라서 이러한 데이터를 줄이는 목적으로 사용된다. 이러한 활성화 함수의 특징을 통해 추출된 특징들을 분류기에 넣기 위해 0과 1사이의 확률값으로 표현하기 위해 사용된다. 데이터의 특징에 따라 활성화 함수를 결정할 수 있고 종류에는 시그모이드 함수, ReLU 함수, 소프트매스 함수 등이 사용된다(표3).

이름	식
이진 계단 함수	$\sigma(x) = \begin{cases} 0, x \leq 0 \\ 1, x > 0 \end{cases}$
시그모이드 함수	$\sigma(x) = \frac{1}{1+\exp(-x)}$
Tanh 함수	$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$
소프트맥스 함수	$y_k = \frac{\exp(a_k)}{\sum_{j=1}^{n} \exp(a_i)}$
ReLU 함수	$f(x) = \max(0, x)$
ELU 함수	$f(x) = \begin{cases} \alpha(e^{x} - 1), x \leq 0 \\ x \qquad\qquad , x > 0 \end{cases}$
LeakyReLU 함수	$f(x) = \max(0.01x, x)$

표 3 대표적인 활성화 함수의 식

Chapter 1. 제한된 의료데이터셋을 이용한 딥러닝 모델 기반 심전도 분류 모델 성능 향상

1. 배경

딥러닝을 활용한 심전도 분류 다양한 방법으로 연구되고 있다.[7-9] 기존에는 특징 기반 모델이나 규칙기반 알고리즘에 중점을 둔 심전도 분류 방법이 주를 이뤘으나, 최근에는 심전도 데이터를 그대로 활용하거나 최소한의 수정만을 가한 모델이 등장하였다. 이로 인해 정확도가 크게 향상된 심전도 분류 모델들이 제안되었다. 하지만 심전도 데이터는 시계열 데이터의 일종으로 1차원 신호에 속한다. 2차원 신호인 이미지와는 상대적으로 추출되는 특징이 적기 때문에 딥러닝 모델 학습에 필요한 데이터양이 많이 요구된다. 그럼에도 이러한 한계점을 극복하기 위해 심전도의 측정값, 푸리에 변환(Fourier transform), 스펙토그램(Spectogram) 이미지등 모델의 정확도를 향상시키기 위해 입력데이터로 쓰인다.[10-12]

한편, 의료 데이터는 종종 데이터 불균형 문제가 발생하는데 이는 분류모델에 상당한 문제를 야기한다.[13] 데이터 불균형이란, 데이터가 특정 카테고리에 치중되어 다수 클래스와 소수클래스로 나뉘게 되는 현상을 말한다. 데이터 불균형이 있는 데이터셋을 딥러닝 모델 학습을 진행할 경우 다수 클래스에 편향되도록 학습이 진행되기 때문에 소수 클래스에 대한 분류 성능이 현저히 떨어질 수 있다. 이를 해결하기 위해 신호 기반의 인공지능 생성 모델이나 수학적 모델링을 통해 소수클래스의 데이터 특징을 기반으로 인위적인 데이터를 생성하여 데이터 균형을 맞춰 학습을 진행하는 연구들이 많이 진행되고 있다.[14] 이러한 방법들이 모델의 성능을 상당히 향상되었음에도 불구하고 의료데이터 특성상 실제 측정되지 않은 환자의 데이터를 생성하여 모델에 학습을 진행했다는 사실 즉, 인위적으로 생성된 데이터는 실제 데이터가 아니라는 기조가 의료계에서는 논란이 되고 있다.[6] 따라서 본 연구 파트에서는 데이터 증강 기법을 사용하지 않고 제한된 의료데이터 셋에서 모델의 손실함수에 가중치를 주는 방법 2가지와 균형 데이터셋과 하위 클래스 방법 총 4가지의 방법을 비교하여 최적의 성능을 확인한다.

2. 연구방법

(1) 데이터셋

데이터는 서울 아산병원 뮤즈(Muse, GE Health care, USA) 시스템으로부터 7355명의 심전도 데이터를 사용한다. 데이터는 윤리심의의원회에서 승인(IRB 2021-1259)을 받은 데이터이다. 데이터는 18세 이상 환자에서 12 리드(Lead1, Lead2, Lead3, aVL, aVR, aVF, V1, V2, V3, V4, V5, V6)심전도 10초로 구성된 데이터이고, 샘플링속도는 500Hz만 포함시켰다.(그림10) 데이터의 클래스는 총 8개로 atrial flutter, a first or second degree AV block (Mobits type 1), supraventricular tachycardia(PSVT), a high degree or complete AV block, sinus node dysfunction, sinus tachycardia, ventricular premature contraction(VPC) 포함했다. 데이터셋의 비율은 최대 32.3% 부터 3.8% 까지 데이터 불균형이 존재하는 데이터셋이다. 데이터 셋은 총 3가지로 불균형 데이터셋, 균형 데이터셋, 하위클래스 데이터셋 이렇게 구성된다. 불균형 데이터셋은 데이터 불균형이 있는 데이터셋이고, 균형 데이터셋은 가장 소수클래스의 개수로 모든 클래스의 개수를 동일하게 맞춰준 데이터셋이고, 하위 클래스 데이터셋은 소수클래스를 하나의 클래스로 합친 데이터 셋이다.(표 4)

Class	Name of class	Imbalance dataset (%)	Balance dataset (%)	Subclass dataset (%)
1	Atrial Flutter	655 (8.9)	279(12.5)	1540(20.9)
2	1st degree AV block or 2nd degree AV block (Mobits type 1)	279 (3.8)	279(12.5)	
3	PSVT	307(4.2)	279(12.5)	
4	High degree or complete AV block	299(4.0)	279(12.5)	
5	Irregular narrow QRS tachyarrhythmia	1451(19.7)	279(12.5)	1451(19.7)
6	Sinus node dysfunction	850(11.6)	279(12.5)	850(11.6)
7	Sinus tachyarrhythmia	1132(15.4)	279(12.5)	1132(15.4)
8	VPCs	2382(32.4)	279(12.5)	2382(32.4)
Total		7355(100)	2232(100)	7355(100)

표 4 실험별 데이터 분포

모든 데이터셋은 딥러닝 모델에 맞게 동일한 방법의 데이터 전처리를 진행한다. 먼저 심전도 측정할 때 호흡 또는 가슴의 움직임으로 발생하는 기저선 잡음을 제거하기 위해 샤비츠키 골레이(Savitzky-Golay) 필터를 사용하여 저주파 노이즈를 제거한다.[15] 특정 데이터의 크기로 인해 편향되게 학습되는 것을 방지하기위해 정규화 과정을 거친다. 본 연구는 최대-최소 정규화 (Min-Max normalization)을 진행하여 전체 데이터의 분포를 -1에서 1사이의 값으로 정규화를 진행한다(그림11).

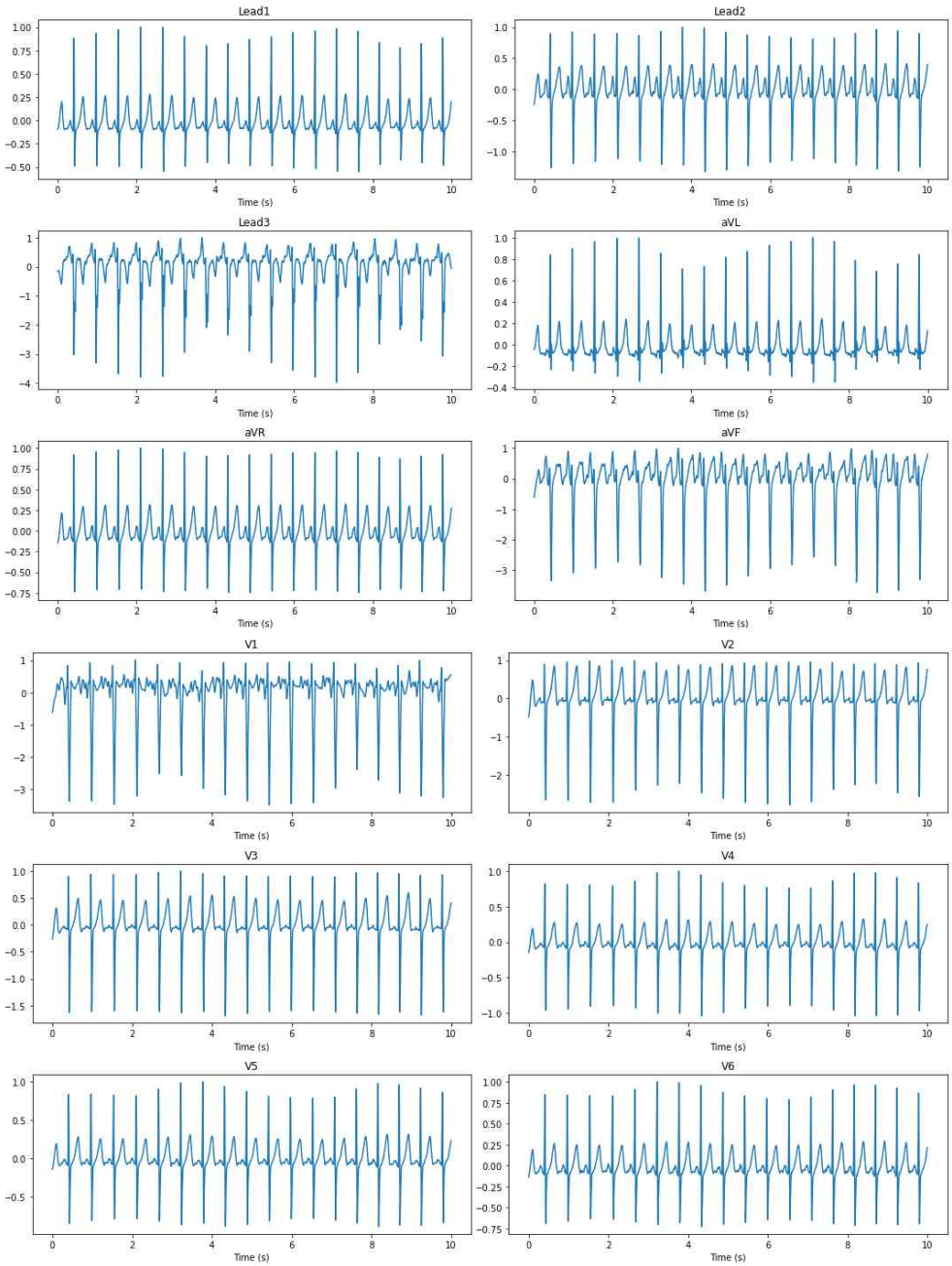

그림10. 12 리드 심전도

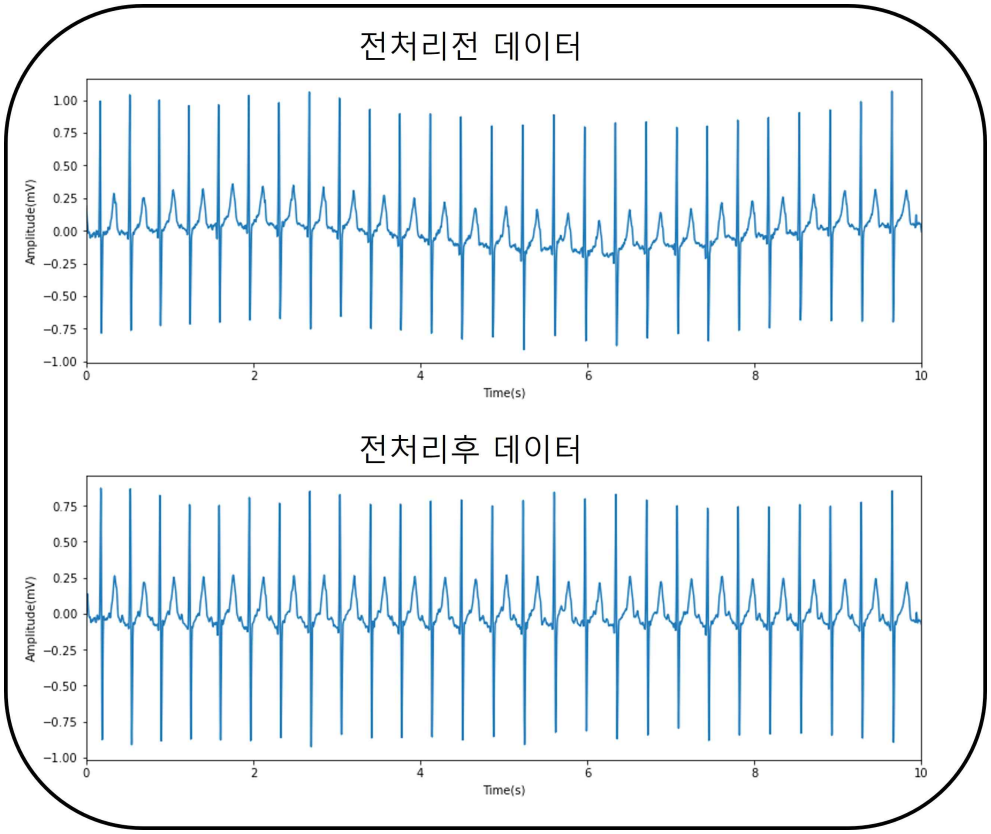

그림11. 데이터 전처리 결과

(2) 딥러닝 모델

심전도 분류를 위한 딥러닝 모델은 2016년에 Google에서 발표한 인셉션 넷 V3 모델을 채택한다.[16] 인셉션 넷 V3 모델은 이전에 개발된 모델인 AlexNet, VGGNet의 문제점을 개선한 모델이다. 이전 모델에서는 레이어를 많이 쌓게 되면 제한된 컴퓨터 메모리 환경에서 모델 학습에 대한 문제점이 발생한다. 그래서 인셉션 넷 V3 모델은 입력데이터에 대한 많은 특징을 다양한 필터의 크기를 이용하여 추출하고, 이를 풀링(Pooling) 하는 과정에서 생기는 정보손실도 최소화해주는 모델이다(그림12). 모델에서 사용된 기법은 합성곱 분해(Factorizing convolution) 및 비대칭 합성곱(asymmetric convolution) 기법과 그리드 축소(Grid reduction) 이다. (그림13) 합성곱 분해는 연산량이 큰 필터를 작은 필터로 개수를 늘려 연산량을 상대적으로 줄이는 기법이다. 이 방법은 모델이 점점 깊어 짐에 따라 발생하는 연산량 증가를 줄일 수 있다. 비대칭 합성곱은 NxN 크기의 필터를 1xN and Nx1 크기의 필터로 필터를 분해하는 기법이다. 이 방법은 기존 크기의 필터 연산량 보다 적은 연산량으로 특징을 추출할 수 있다. 일반적으로 합성곱 신경망에서 특징 맵(Feature map) 의 크기를 줄이는 방법으로 풀링(Pooling) 연산을 하게 된다. 하지만 이러한 풀링 연산은 연산량을 감소 시켜주는 대신 특징의 손실이 올 수 있다. 따라서 그리드 축소 방법을 사용한다. 특징을 추출하는 부분과 풀링 연산을 기존의 직렬적인 계산이 아닌 병렬적 계산으로 변경하여 특징의 손실을 방지하는 방법이다. 인셉션 넷V3 모델은 위의 방법으로 총 3가지의 인셉션 모듈 (Inception module) 과 그 사이에 그리드 축소 모듈 2개를 넣어 입력 데이터에 대한 특징을 다양한 필터를 사용하고, 레이어가 깊어짐에 따라 발생하는 연산량을 줄일 수 있고, 연산량을 줄이는 과정에서 발생하는 정보손실 또한 보완한 모델이다. 각 인센셥 모듈은 각기 다른 필터 크기의 합성곱 연산을 통해 특징을 추출하게 되고, 추출된 특징과 풀링 연산의 병렬적인 계산이 진행되어 특징을 추출한다. 마지막으로 추출된 특징을 기반으로 분류기(Classifier) 레이어에 들어간다. 분류기에는 추출된 특징의 해당 클래스에 속할 확률적 표현을 위해 0과 1사이의 값으로 표현되는 시그모이드 함수가 사용된다. 그리고나서 특징들의 차원을 축소시키는 Flatten 레이어를 통과하면 8개의 클래스에 대한 확률값들이 산출된다.

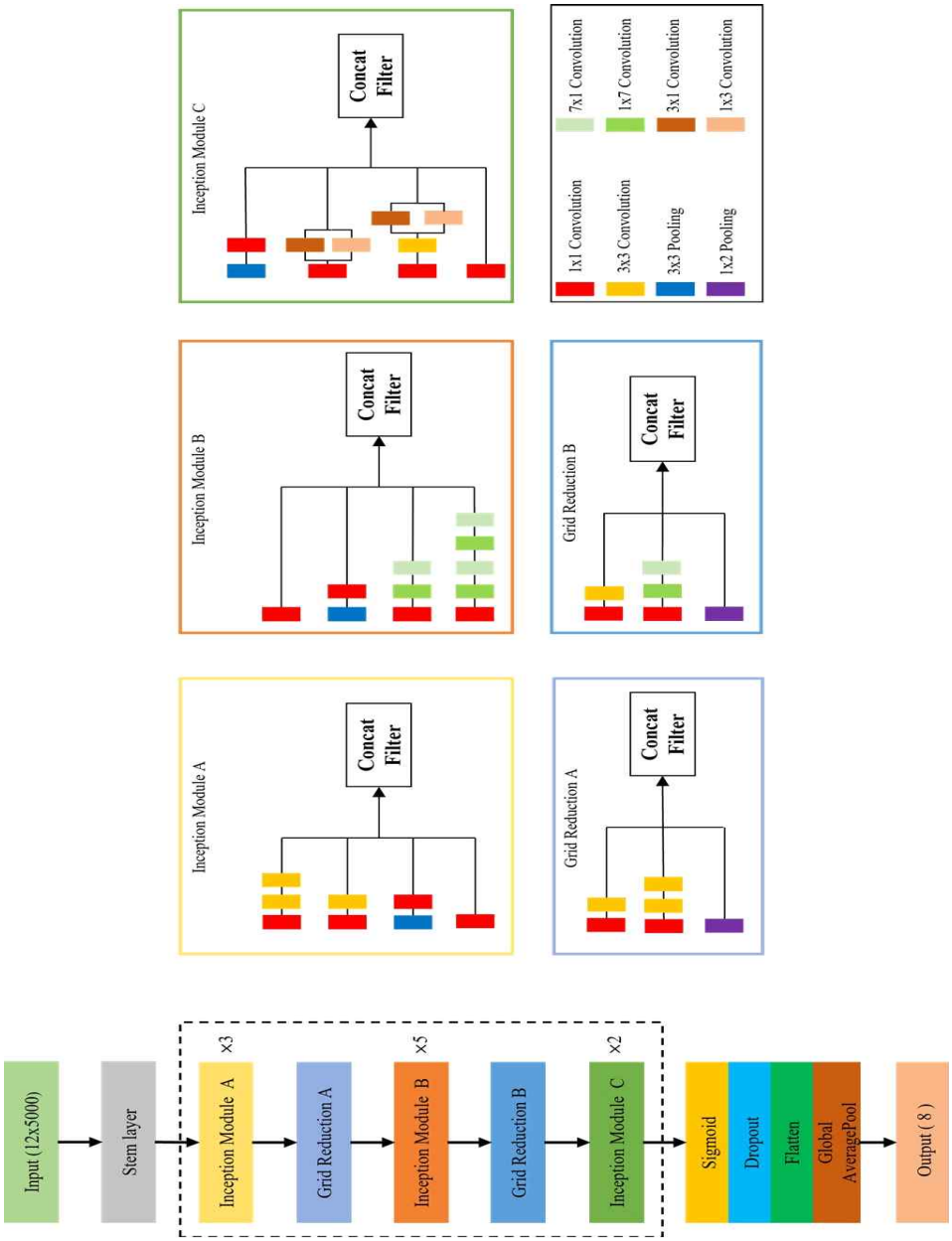

그림12. 인셉션넷 V3 모델 구조

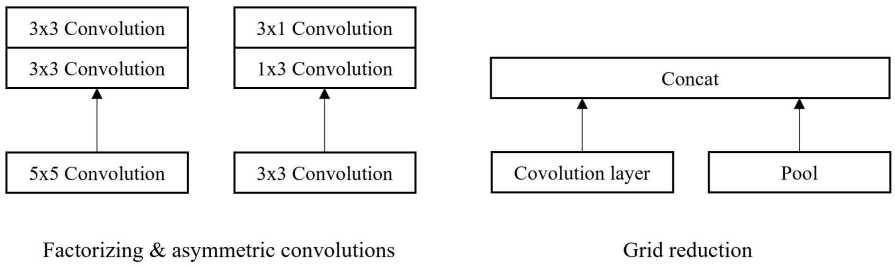

그림13. 합성곱 분해 및 비대칭연산과 그리드축소 기법

3. 연구평가 및 결과

연구모델의 평가는 분류모델에서 사용되는 평가지표를 활용하여 각 실험에 대한 평가를 진행한다. 평가 요소는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수를 사용한다.(표5) 식에서 True positive(TP)는 모델이 예측한 양성 케이스 중 실제 양성 케이스에 해당하는 것을 나타내고, True negative(TN)은 모델이 예측한 음성 케이스 중 실제 음성 케이스(Case)에 해당하는 것을 나타내고, False positive(FP)는 모델이 양성으로 예측했는데 실제는 음성인 케이스를 나타내고, False Negative는 모델이 음성으로 예측했는데 실제는 양성인 케이스를 나타낸다.

평가지표	식
정확도(Accuracy)	$\frac{TP+\ TN}{TP+\ FN+FP+TN}$
정밀도(Precision)	$\frac{TP}{TP+FP}$
재현율(Recall)	$\frac{TP}{TP+FN}$
F1 점수(F1 score)	$\frac{2\times(\mathrm{Pr}ecision\times\ Recall)}{\mathrm{Pr}ecision+\ Recall}$

표 5 모델 평가 지표

결과는 각 클래스에 해당하는 평가 결과와 각 실험에 대한 평가를 Micro average 지표를 사용하여 비교하였다.(표6,표7) 실험 B에서는 사용한 균형 데이터셋 모델의 결과는 0.85와 0.86 사이에 결과가 나왔다. 이는 데이터를 균형적으로 맞추기 위해 데이터를 소수클래스의 개수로 제한하는 것은 오히려 성능이 떨어졌음을 알 수 있다. 이와 대조적으로 실험 A에서 클래스 가중치 모델과 포칼로스 함수 모델의 결과가 기존의 데이터 불균형 모델에 비해 둘다 성능이 개선되었다. 그 중, 포칼로스 함수 모델이 가장 성능이 F1 점수 0.96으로 가장 높은 성능을 보였다. 하위 클래스 모델 성능은 첫번째 분류결과는 F1 점수가 0.97로 모든 실험에 대해서 가장 높은 성능 보였지만, 결국 소수클래스들간 분류를 진행한 두번째 분류결과는 F1 점수 0.84로 오히려 성능이 떨어졌다. 다음 그림은 각 실험에 대한 혼동행렬이다. 데이터 불균형 모델 성능에서 1번부터 4번 까지의 소수클래스에 대한 성능이 떨어졌음을 확인할 수 있다. 반대로 포칼로스 함수 모델에서 1번을 제외한 모든 소수 클래스에 대한 분류 성능이 크게 향상되었다.

Type		Accuracy	Precision	Recall	F1 Score
Imbalance		0.92	0.92	0.92	0.92
Focal loss		0.96	0.96	0.96	0.96
Class weight		0.95	0.94	0.95	0.95
Balance		0.85	0.86	0.86	0.86
Subclass	1st	0.97	0.96	0.97	0.97
	2nd	0.85	0.84	0.85	0.84

표 6 각 실험별 모델 평가 결과(Micro Average)

Class	Imbalance dataset	Focal loss	Class weight	Balance dataset	Subclass dataset	
					First classification	Second classification
Atrial Flutter	0.77	0.87	0.86	0.85	0.93	0.85
1st degree AV block or 2nd degree AV block(Mobits type 1)	0.85	0.94	0.91	0.68		0.86
PSVT	0.78	0.83	0.81	0.92		0.82
High degree or complete AV block	0.75	0.88	0.88	0.85		0.79
Irregular narrow QRS tachyarrhythmia	0.92	0.98	0.97	0.65	0.98	-
Sinus node dysfunction	0.97	0.98	0.98	0.98	0.99	
Sinus tachyarrhythmia	0.93	0.95	0.96	0.96	0.95	
VPCS	0.99	0.99	0.99	0.99	0.99	

표 7 각 클래스별 F1 점수 결과

4. 고찰

본 연구의 결과는 제한된 의료데이터의 최적의 12리드 심전도 분류 모델을 제안한다. 데이터 불균형이 있는 데이터셋을 학습할 때에는 각 클래스의 데이터 개수를 임의로 최소로 맞추거나 하위클래스를 구성하는 것 보다 데이터 불균형인 상태에서 모델의 손실함수에 가중치는 주는 방법이 가장 데이터 불균형으로 인한 성능 감소를 완화 해주는 것으로 나타난다. 우리는 또한 3가지의 이전 연구에 대한 성능을 비교한다. 각 연구는 같은 데이터셋으로 다른 모델을 적용하고, 손실함수는 포칼로스를 사용하여 모델의 성능을 비교한다(표8). Sakli 외[18] 와 Zhou외[19] 는 공통적으로 레즈넷(Resnet) 기반의 모델을 사용한 연구이다. 각 연구의 F1 점수는 0.94 와 0.92로 본 연구의 모델의 성능보다 낮음을 알 수 있다. 따라서 심전도와 같이 1차원 데이터의 같은 경우 이미지 기반 2차원 데이터보다 특징이 적기 때문에 여러 크기의 필터를 사용하여 다양한 특징을 추출을 기반한 인셉션넷 V3 모델이 심전도 데이터의 더욱 적합한 모델임을 알 수 있었다. 하지만 본연구는 몇가지 한계점이 존재한다. 먼저 연구에 사용된 데이터셋은 실제 의료데이터를 완전히 대표할 수 없다는 점이다. 따라서 다른 심장질환을 추가하여 모델의 성능을 다양한 모델과의 성능 평가가 필요하다. 그리고 데이터셋은 환자 중심이 아닌 심장질환 중심으로 데이터의 클래스를 나눴기 때문에 이점은 다른 요소에 대한 분류 결과 분석이 부족하다는 점이다. 마지막으로 하위 클래스를 구성할 때, 데이터 불균형 모델의 성능을 기반으로 소수클래스를 정의하여 하나의 하위클래스를 구성하였는데, 질환의 특성, 각 환자에 대한 특성을 고려한 다른 요소를 반영하여 하위클래스를 구성하여 다양한 분석이 이루어져야 한다는 점이다. 위의 한계점은 추후 딥러닝 분류 모델에 영향을 줄 수 있는 다양한 요소를 고려하여 연구되어야 한다.

Study	Method	Class	Precision	Recall	F1 score	Support
Ribeiro et. al. (2020) [17]	DNN	1	0.87	0.64	0.74	136
		2	0.78	0.56	0.65	67
		3	0.67	0.92	0.77	57
		4	0.75	0.84	0.79	58
		5	0.89	0.88	0.88	284
		6	0.98	0.99	0.99	178
		7	0.91	0.96	0.94	216
		8	0.97	1	0.98	475
		Micro Average	0.91	0.91	0.91	
Sakli et.al. (2022) [18]	Resnet50	Class	Precision	Recall	F1 score	Support
		1	0.86	0.77	0.81	136
		2	0.95	0.72	0.82	67
		3	0.8	0.93	0.86	57
		4	0.7	0.92	0.79	58
		5	0.97	0.89	0.93	284
		6	0.99	0.98	0.99	178
		7	0.93	0.98	0.95	216
		8	0.98	1	0.99	475
		Micro Average	0.94	0.94	0.94	
Zhou et.al. (2020) [19]	Resnet50 + LSTM	Class	Precision	Recall	F1 score	Support
		1	0.85	0.66	0.74	136
		2	0.77	0.72	0.75	67
		3	0.73	0.92	0.81	57
		4	0.73	0.76	0.75	58
		5	0.9	0.92	0.91	284
		6	0.99	0.98	0.99	178
		7	0.93	0.97	0.95	216
		8	0.98	0.99	0.99	475
		Micro Average	0.92	0.92	0.92	
Our study	Inception Net	Class	Precision	Recall	F1 score	Support
		1	0.9	0.83	0.87	136
		2	0.95	0.93	0.84	67
		3	0.79	0.88	0.83	57
		4	0.87	0.9	0.88	58

		5	0.98	0.99	0.98	284
		6	0.99	0.98	0.98	178
		7	0.95	0.96	0.95	216
		8	0.99	1	0.99	475
		Micro Average	**0.96**	**0.96**	**0.96**	

표 8 이전 모델과의 성능 비교 결과

Chapter 2. 심전도 구간별 인공지능 이상치 탐지를 기반으로 한 심방세동 진단

1.연구배경

심방세동은 전세계적으로 가장 흔히 발생하는 치명적인 질환이다. 심방세동의 유병률은 3억 7천만 건에 이르며, 지난 20년간 33%가 증가했다.[20] 심방세동은 허혈성 뇌졸중과 심부전과 같은 심각한 건강 문제로 이어질 수 있다. 고령 인구가 증가함에 따라 심방세동의 유병률이 더욱 높아지고 이에 따라 조기진단에 대한 중요성이 커졌다. 또한 심방세동의 조기진단은 이에 따른 잠재적인 합병증의 위험을 줄이는데 중요하다. 심전도 신호는 심장 질환의 진단에 흔히 사용되고 신호의 형태와 리듬을 통해 심장 질환을 식별하는 도구이다. 심전도 형태는 크게 P파, QRS 복합체 및 T파로 구성되며 각 심박주기마다 순차적으로 기록된다.(그림14)

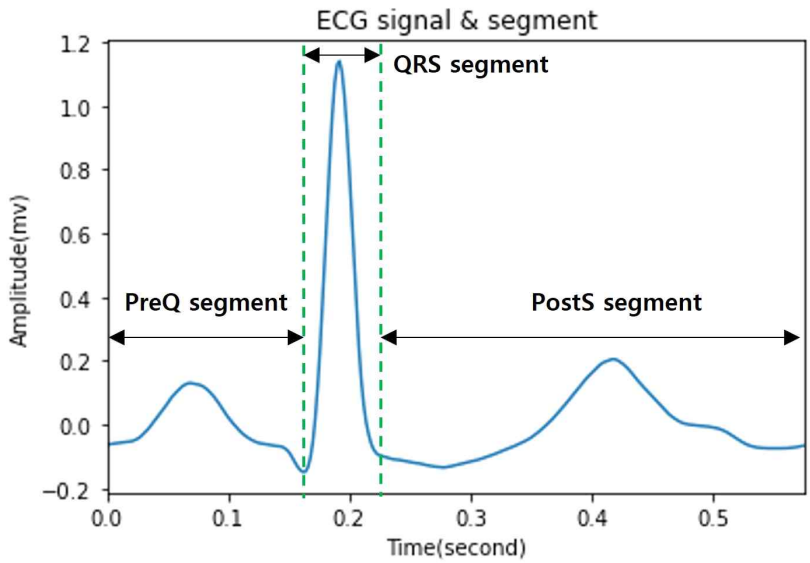

그림 14 심전도 신호 및 구간

최근 딥러닝 기법은 심방세동을 정확하게 진단하는데 뛰어난 성능을 보였으며 의사료진들의 빠르고 정확한 진단을 내리는 데 도움이 되고 있다. 대부분의 심방세동 진단 연구는 지도 학습 방법을 사용한다. 그러나 지도 학습은 라벨링 과정이 필요하기 때문에 비용이 많이 들고 시간이 많이 소요될 수 있다. 이러한 문제점을 해결하기 위해 최근에는 비지도 학습 방법도 사용된다. 비지도 학습의 한 유형인 이상치 탐지는 대량의 정상 데이터를 학습하고 실시간 신호에서 정상 특징과 다른 비정상 데이터가 들어오게 되면 이상치가 발생하여 이상치를 탐지하게 되는 알고리즘이다. 심전도 신호에서의 이상치는 정상 심전도와 다른 다양한 부정맥이 이에 포함될 수 있다. 오토인코더(Autoencoder) 및 GAN 과 같은 다양한 모델들이 심전도 신호의 이상치를 감지하거나 예측하는데 사용되어왔다. Thill외[21]는 MIT-BIH 부정맥 데이터셋에서 시간적과 공간적 특징을 추출할 수 있는 시간적 합성곱 오토 인코더 모델을 활용하여 F1 점수가 0.92에 도달한 연구가 있고, Jang 외[22]은 정상 심전도 신호를 기반으로 4가지 유형의 심장질환(이상치)를 감지하기위해 합성곱 변이형 오토인코더 모델을 사용하여 모든 이상치에 대해서 F1 점수 0.86, 심방세동에 대한 F1 점수는 0.76의 성능을 보였다. Hou외[23]은 RNN 기반 LSTM 오토인코더 모델을 활용하여 MIT-BIH 부정맥 데이터셋을 사용하여 정상과 비정상 신호를 구별하여 전체적인 평균 정확도는 0.994에 도달했다. 오토인코더 뿐만 아니라 GAN 기반 모델도 연구 되었으며 이상치 탐지 성능 향상에 많은 발전을 이루었다. Zhu 외[24]은 정상과 비정상 클래스를 구별하기 위해 정상과 비정상 심전도를 구별하기위해 LSTM 기반 GAN 모델을 사용하여 정확도 0.81의 성능을 보였다. Qin외[25] 은 정상 심전도만 학습한 ECG-ADGAN 모델을 개발하여 MIT-BIH 데이터셋를 사용하여 정상 비정상 클래스를 구별하고 비정상 클래스도 추가적인 분류를 진행했다. 해당연구의 성능은 F1점수 0.94의 성능을 보였다. 일반적으로 위의 언급된 대부분의 연구는 입력데이터의 형태가 10초 기록, 비트 또는 R피크 간격 에서 심전도 데이터를 사용했다.

이러한 연구들의 유의미한 성능을 보였주었지만, 현재 딥러닝 모델로 예측된 결과는 많은 의료진들에게 임상적 해석에 대한 근거가 부족하고 여전히 블랙박스로 남아있다. 이 한계를 극복하기 위해 Class Activation map(CAM) 및 Attention 기반 모델이 사용되어 예측한 결과에 대한 설명 모델들이 활발히 연구 되고 있다.[26-29] 하지만 이러한 접근 방법들은 학습한 데이터에 대한 의존성이 너무 강하고 결과가 일관적이지 않기 때문에 의료계의 진단 시스템에서는 큰 한계점을 가진다.

따라서 본연구 파트에서는 심전도의 임상적의미를 가지는 구간을 Q파 이전(PreQ), QRS 구간(QRS), S파 이후(PostS) 로 나누어 이를 포함하는 심전도 구간을 가지고 정상 심전도를 LSTM 기반 오토인코더 모델을 학습하여 각 클래스에 대한 이상치 탐지를 사용한 심방세동의 진단 시스템을 제안한다. 본연구의 기여는 다음과 같다.

-비지도 학습을 통해 데이터 라벨링에 소요되는 시간과 라벨된 데이터셋의 제한적인 부분을 완화를 위해 이상치 탐지를 사용하여 입증한다.

-딥러닝 모델의 예측된 결과를 임상적으로 해석하기 위해 심전도를 PreQ, QRS, PostS 부분으로 나누고 각 부분에서 이상치 점수를 계산하여 정상과 심방세동의 각 부분별 점수를 통해 구별한다.

-심전도 구간별 이상치 점수를 비교하여 심방세동을 진단하는 새로운 진단 시스템을 제안하고, 의료진에게 기존 딥러닝 모델과 비교하여 높은 신뢰성을 준다.

2.연구방법

본 연구는 데이터 전처리, 모델학습, 이상치 점수 계산 및 이상치탐지 과정으로 이루어져 있다.(그림15)

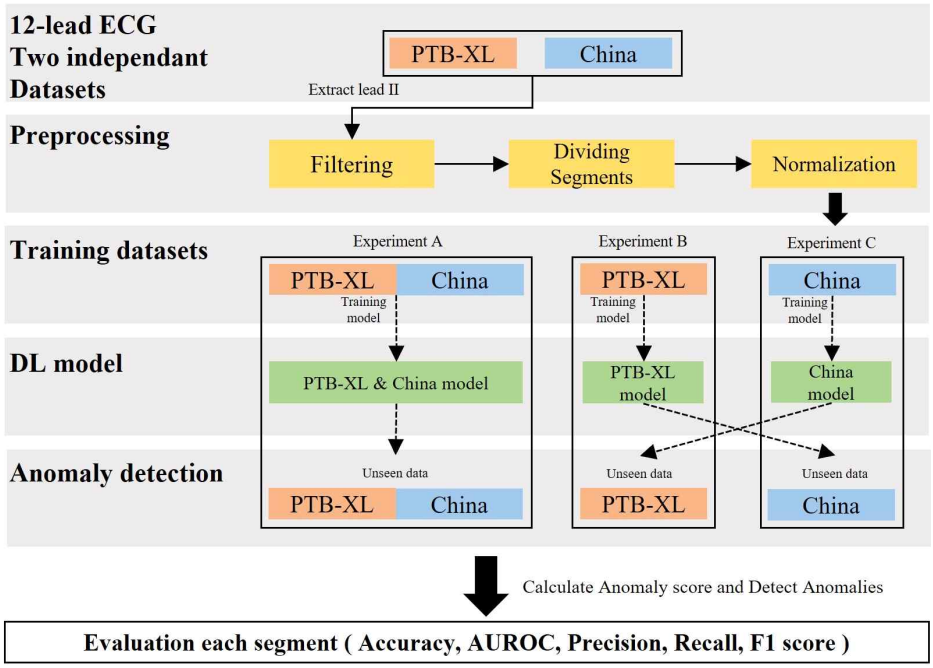

그림 15. 연구의 오버뷰(Overview)

(1) 데이터셋

연구의 데이터셋은 공공데이터셋 PTB-XL 독일 데이터와 PhsiyoNet 에서 제공한 중국데이터를 사용한다. PTB-XL 데이터셋은 1989년 에서 1996년까지 독일에서 측정된 12 리드 심전도 데이터이고, 총 18885명의 환자 데이터의 21837 데이터 기록을 사용한다.[30] 중국데이터는 GE MUSE ECG 시스템에 의해 Shaoxing People's Hospital of China에서 기록된 12리드 심전도 데이터이고, 총 10646명의 환자 기록을 사용한다.[31] 두 데이터셋 모두 10초의 기록과 샘플링 속도 500Hz 이다. 본 연구에 사용한 데이터는 PTB-XL 독일 데이터의 정상심전도 7528개, 심방세동 심전도 1514개와 중국데이터의 정상심전도 5419개, 심방세동 심전도 1780개를 포함하고, 12리드 중 리드 2 기록만 추출하여 단일리드의 심전도를 사용한다. 우리는 3개의 실험을 통해 모델의 일반성을 입증한다. 전체 데이터를 섞어 모델학습을 진행 및 평가를 하고 2개의 실험에서 교차검증을 통해 인종간 모델의 일반성을 확인한다(그림16). 데이터셋은 훈련셋과 테스트셋의 비율은 8:2로 나누어 실험을 진행한다.

12-lead PTB-XL & China Dataset
(Normal : 12947 , AF : 3294)
Preprocessing for model
Experiment A
Training
10357 Normal records with 50763 each segment
Test
2590 Normal records
3294 AF records
12-lead PTB-XL Dataset
(Normal : 7528 , AF : 1514)
12-lead China Dataset
(Normal : 5419, AF : 1780)
Preprocessing for model
Experiment B
Training (PTB-XL)
9528 Normal records with 30406 each segment
Test (China)
5419 Normal records
1780 AF records
Experiment C
Training (China)
5419 Normal records with 20357 each segment
Test (PTB-XL)
7528 Normal records
1514 AF records

그림 16. 데이터셋

(2) 전처리

전처리 과정은 3단계로 필터링, 심전도 분할, 정규화 과정을 포함한다 . 심전도의 저주파 노이즈(기저선 흔들림) 와 고주파 노이즈(전원선 노이즈, 근육떨림)를 제거하기 위해 주파수 범위를 0.5에서 50Hz로 설정한 4차 버터워스(Butterworth) 밴드패스 필터(Band-pass filter)를 이용한다. 필터링 적용후에는 심전도의 단일리드 단일 비트에서 의미있는 부분인 PreQ, QRS, PostS 부분을 나눈다. 먼저 단일 비트를 나누기 위해 10초 데이터에 Pan-Tomkin's 알고리즘을 이용하여 R peak를 검출한다.[32] 검출된 R 피크는 각 beat를 나누기 위해 3개의 R피크를 선택한 다음 가운데 R피크를 기준으로 앞 R 피크간 거리의 1/3 뒤 R 피크간 거리의 2/3로 비트를 나누어 하나의 단일 심전도 비트를 만든다. 그리고나서 PreQ, QRS, PostS 부분을 나누기 위해, 단일 비트에서 Q피크, R피크, S피크를 검출한다. R피크는 기존의 Pan-Tomkin's 알고리즘을 이용하여 검출한다. Q피크는 우선 R피크 이전 0.08초 부분의 데이터를 추출한다. 그리고나서 추출된 데이터 사이의 기울기들을 구한다. 그래서 기울기가 처음으로 양수에서 음수로 바뀌는 부분의 지점을 Q피크로 정한다. 동일한 방법으로 S피트는 R피크 이후 0.1초 데이터를 추출하고 R-S 부분의 기울기를 계산하여 처음으로 음수에서 양수로 바뀌는 지점을 S피크로 정한다. 단일비트에서 세개의 부분으로 PreQ, QRS, PostS로 나누게 되면, 각 부분에 해당하는 데이터의 길이가 전부 다르기 때문에, 너무 길게 잡힌 부분 또는 짧게 잡힌 부분은 제거한다. PreQ는 77 부터 174 샘플(sample), QRS는 28에서 62샘플, PostQ는 167에서 362샘플로 생물학적 범위내에 해당하는 데이터만 포함시킨다. 그리고나서 각 부분의 해당하는 심전도 데이터의 분포를 일정한 간격으로 맞추기 위해 최대-최소 정규화 방법을 사용하여 데이터의 폭을 -1에서 1로 정규화를 진행한다. (자세한 식은 Chapter 1-(1) 데이터셋을 참고) 각 심전도 부분에 대한 길이가 전부 다르기 때문에, 딥러닝 모델의 입력을 동일하게 하기위해, PreQ, QRS, PostS의 길이를 256,64,512 길이로 각 부분을 제로 패딩(Zero padding)을 한다.(그림17)

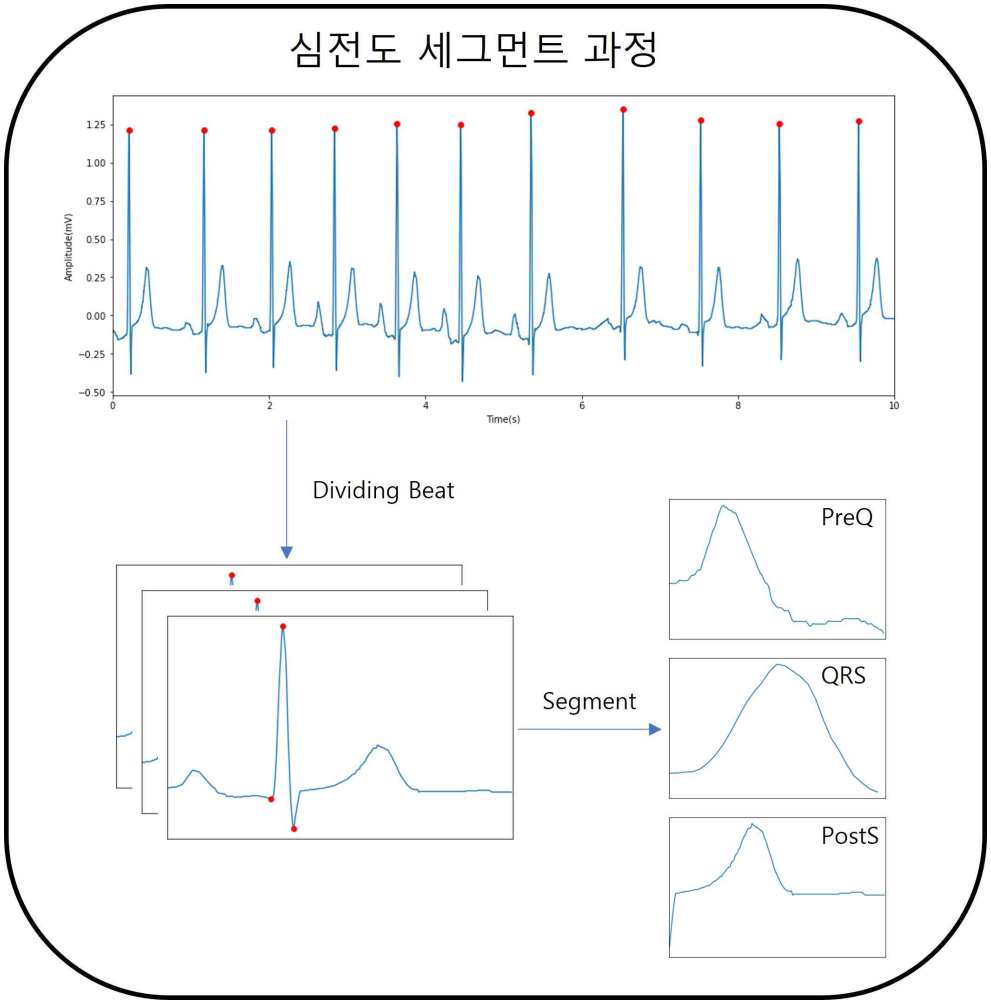

그림 17 심전도 세그멘트 과정

(3) 딥러닝 모델

본 연구에 사용된 모델은 RNN의 한 기법인 LSTM 기반 오토인코더 구조를 채택한다. RNN은 시계열 데이터와 같이 어떤 순서에 따른 데이터를 모델링 하기 위한 기법이다. RNN의 기본적인 원리는 시계열 데이터의 시간적인 정보 즉, 순서에 따른 정보의 특징을 추출하는것으로 은닉 상태(Hidden state)를 단계별로 시계열 전체 정보를 요약하도록 해준다. 하지만 RNN 기법은 입력데이터가 길어지면 앞 시간대에 있는 정보가 손실된다. 따라서 이를 해결하기위해 고안된 모델이 LSTM 모델이다(그림18). LSTM 모델은 과거 데이터의 정보 손실을 어느정도 완화해주고 과거 데이터를 잘 반영하여 미래 데이터를 잘 예측하도록 해주는 모델이다. 그림18에서 W는 단계별 가중치를 나타내고, 활성화 함수는 시그모이드 함수를 사용한다. 먼저 Forgetgate 단계를 통해 들어오는 입력과 이전에 받았던 데이터를 기반으로 현재 정보의 어느부분은 저장하고 버릴것인지 결정한다(표9-식,(1),(2),(3)). 그리고나서 식 (4)에서 알 수 있듯이 사라진 정보와 입력게이트를 통해 어떤 정보를 새로 저장할건지 결정한다(식(5),(6)) 이후에 추가할 정보를 결정하게되면 전체적으로 셀 상태가 최신화 되어 다음 출력값을 도출한다. 위의 언급한 LSTM 기법을 가지고 오토인코더 모델을 구성한다. 오토인코더는 입력데이터의 특징을 추출하여 데이터의 차원을 압축하는 인코더(Encoder)와 압축된 특징(Representation)을 디코더(Decoder)를 이용하여 입력과 같은 차원의 데이터로 다시 복원하는 모델로 구성되어 있다(그림19).

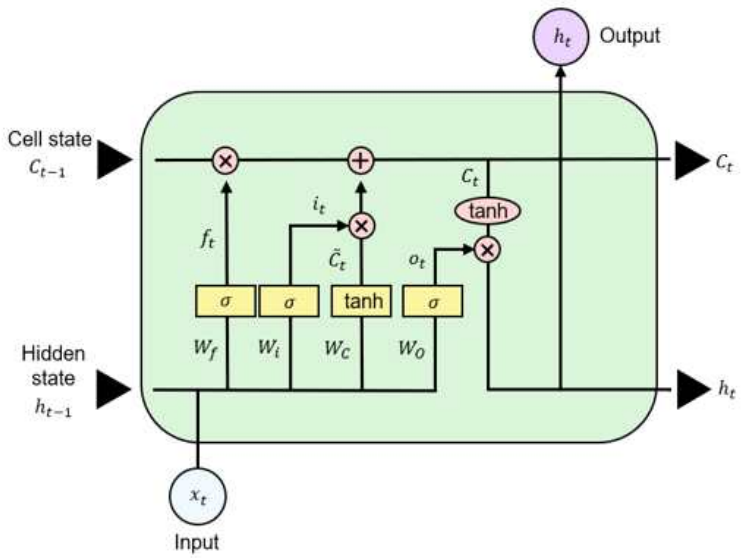

그림 18 LSTM 구조

	수식
(1)	$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$
(2)	$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$
(3)	$\widetilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$
(4)	$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t$
(5)	$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$
(6)	$h_t = o_t * \tanh(C_t)$

표 9 LSTM 식

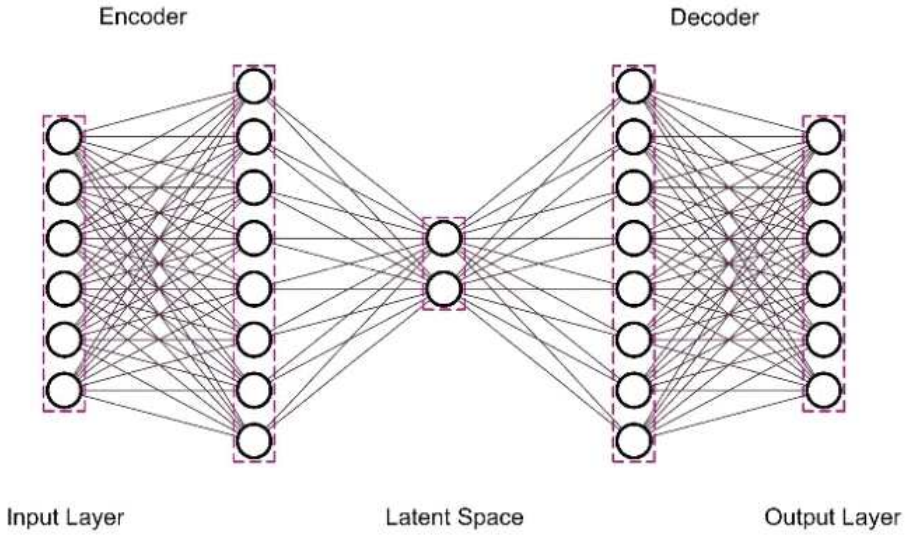

그림 19 오토인코더 기본구조

PreQ, QRS, PostS 모델 3개를 동일 구조의 모델로 학습을 진행한다.(그림20) 인코더는 입력데이터 x를 받아 가중치 W와 편향이 더해져 활성화 함수 f에 들어가고, 이때 활성화 함수는 탄젠트 하이퍼볼릭(hyperbolic tangent) 함수를 사용한다. 인코더의 출력 Z는 다시 디코더의 입력으로 들어가 입력데이터와 같은 차원의 데이터를 생성한다.(표10) 본 모델의 학습목표는 입력데이터의 특성을 잘 추출하고 추출된 특징을 기반으로 생성된 데이터가 입력데이터와 똑같이 재구성 하는 것이다. 따라서 모델의 reconstruction 손실함수는 MSE 값을 사용하여 입력데이터와 출력데이터간 오차가 최소화 되도록 학습을 진행한다(식). 모델의 전반적인 그림언급 및 넣기. 모델의 인코더의 3개의 LSTM 레이어로 구성하고 LSTM의 유닛(Unit)은 64개로 시작하여 절반씩 줄여 신호를 압축하고, 디코더는 반대로 16개로 시작하여 2배씩 늘려 신호를 재구성한다. 표11는 모델의 하이퍼파라미터(hyperparameter)를 나타낸다.

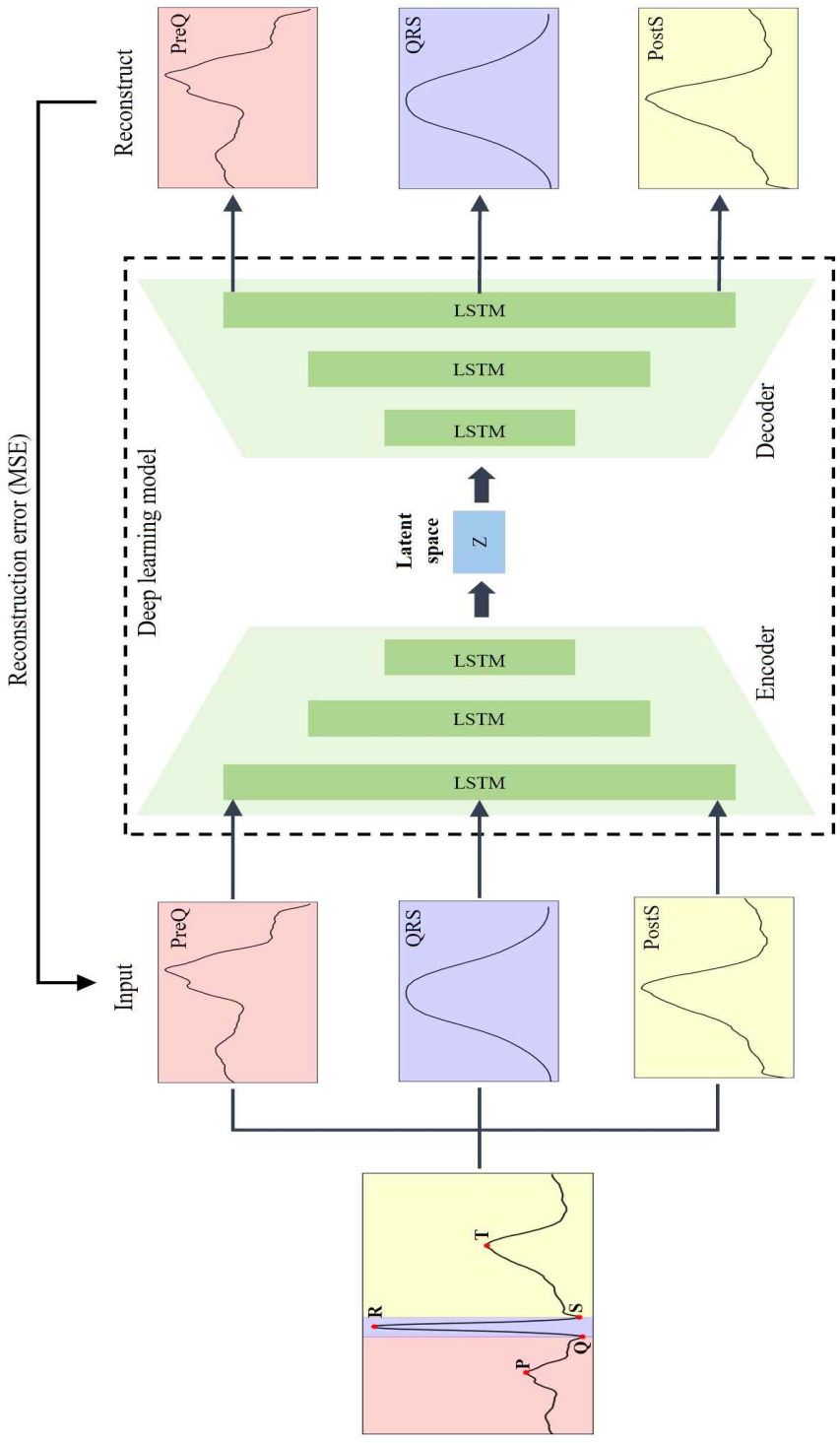

그림20 연구의 전반적인 딥러닝 모델 구조

	수식
인코터	$h_{output\ of\ encoder} = f\left(w_i x + b_i\ \right) = f(Z)$
디코더	$\hat{x} = f'\left(w_j h + b_j\right)$

표 10 오토인코더 식

Experiment	Activation function	Loss function	Batch size	Learning rate	Epoch
Experiment A	Tangent hyperbolic	MSE	32	0.0005	200
Experiment B					
Experiment C					

표 11 딥러닝 모델 하이퍼파라미터

3.연구평가 및 결과

모델 평가를 위해 우리는 각 세그먼트 PreQ, QRS, PostS의 이상치 점수를 계산한다. 이상치 점수는 각 세그먼트(segment)의 제로 패딩이 안된 신호부분만 MSE를 계산하여 각 10초 기록에 대한 이상치점수의 평균을 구한다.(표12) 이상치 점수에 대한 이상치를 탐지하기 위한 역치점(Threshold)을 결정하기 위해 정삼 심전도의 구간별 이상치 점수의 분포와 심방세동 이상치 분포의 Area under receiver of curve(AUROC)의 Youden index를 계산하여 역치점을 구한다.심방세동을 진단하기위해 각 이상치 점수에 대한 역치점을 기준으로 분류모델 평가요소인 정확도, 정밀도, 재현율, F1 점수를 사용하여 모델평가를 진행한다.(자세한 식은 Chapter1-3. 연구평가 및 결과 표6 참고)
실험 A에서 정상군과 심방세동군의 각 세그먼트별 차이의 비율은 각각 14.4 , 2.3, 4.6 으로 PreQ 부분에서 가장 크게 나타났다.

Experiment	Segment	Normal	AFIB	Threshold
Experiment A	PreQ	0.00126	0.0182	0.00284
	QRS	0.0247	0.056	0.0784
	PostS	0.0184	0.086	0.0251
Experiment B	PreQ	0.00393	0.0311	0.00774
	QRS	0.0208	0.123	0.0011
	PostS	0.0486	0.143	0.0543
Experiment C	PreQ	0.00423	0.0346	0.00863
	QRS	0.0279	0.167	0.000693
	PostS	0.036	0.149	0.0914

표 12 각 세그먼트별 이상치점수 및 역치점

이를 기반으로 표13과 그림21는 각 실험에 대한 분류모델 평가 점수와 혼동행렬 및 ROC 그림이다. 각 역치점을 기반으로 실험1에서 PreQ 모델이 AUROC 0.96 으로 가장 높게 나왔고, QRS 모델이 0.75로 가장 낮은 점수가 나왔다. 실험 B와 실험3에서도 마찬가지로 PreQ 에서 AUROC 0.9, 0.96으로 가장 좋은 성능을 보였다.(그림22,그림23)

Experiment	Segment	AUROC	Precision	Recall	F1 score	Accuracy
Experiment A	PreQ	0.96	0.92	0.92	0.92	0.92
	QRS	0.75	0.69	0.7	0.7	0.7
	PostS	0.95	0.9	0.89	0.9	0.9
Experiment B	PreQ	0.9	0.84	0.84	0.84	0.84
	QRS	0.76	0.7	0.7	0.7	0.7
	PostS	0.89	0.79	0.79	0.79	0.79
Experiment C	PreQ	0.96	0.9	0.9	0.9	0.9
	QRS	0.74	0.56	0.56	0.56	0.56
	PostS	0.95	0.87	0.87	0.87	0.87

표13 각 실험별 분류 평가 결과

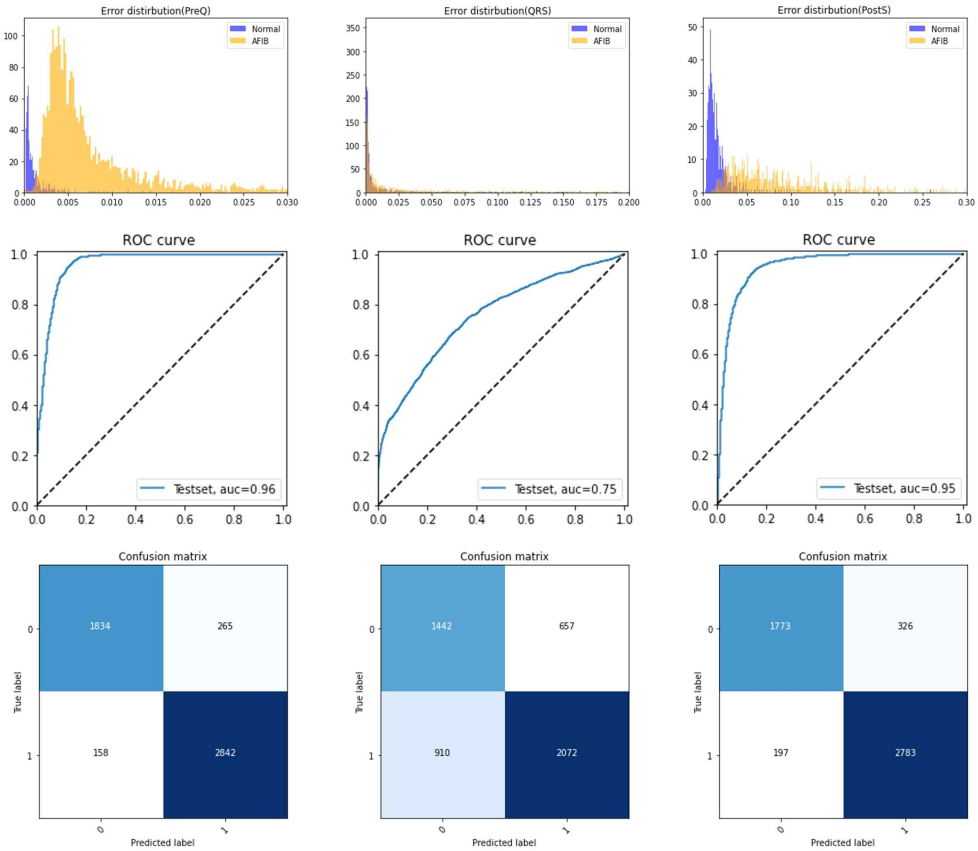

그림 21 실험A의 각 세그먼트별 결과

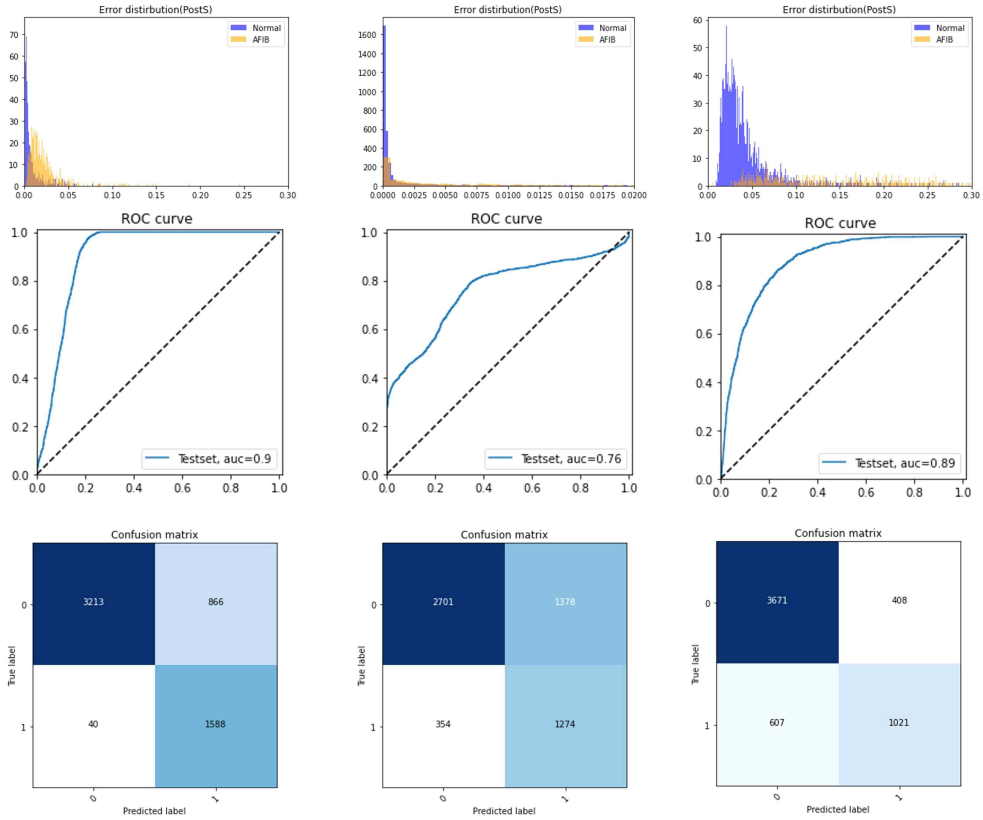

그림 22 실험B의 각 세그먼트별 결과

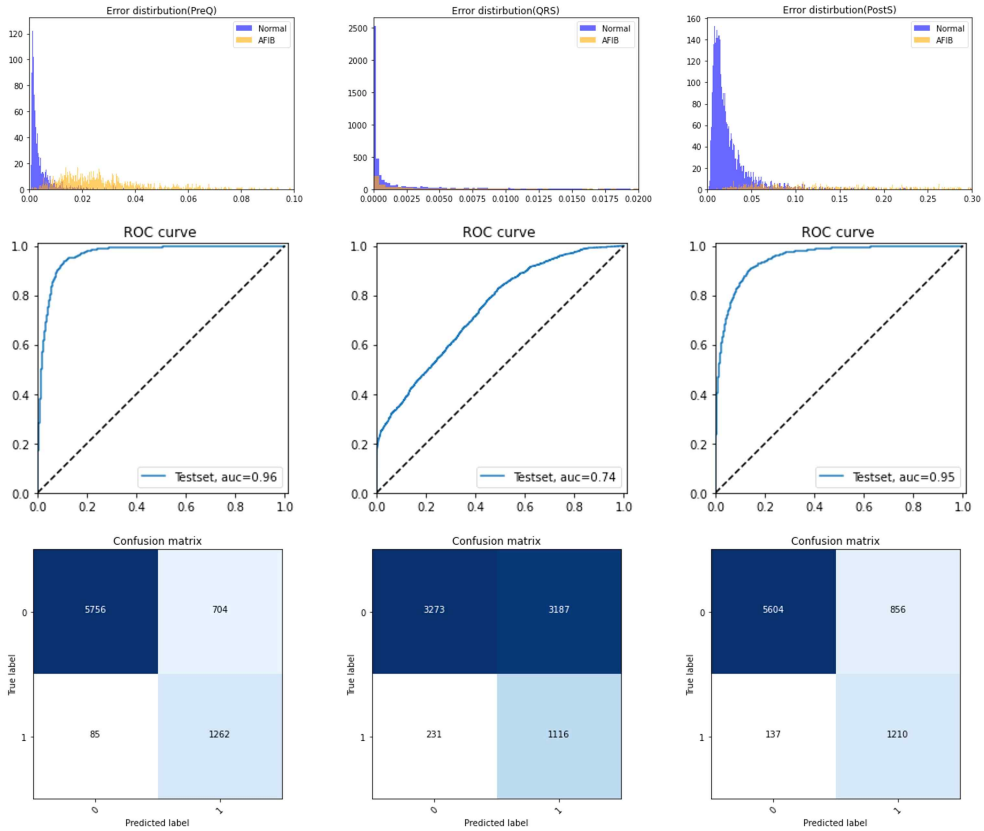

그림 23 실험C의 각 세그먼트별 결과

4.고찰

본 연구는 딥러닝 모델 블랙박스 특성의 한계를 해결하는 심방세동 진단 시스템을 새로운 접근 방식을 제안한다. 결과를 보면, 모든 실험에서 PreQ 모델에서 가장 좋은 성능을 보인 반면 QRS에서 상대적으로 낮은 성능을 보였다. 이 결과는 정상군과 심방세동군의 차이가 Q파 이전 신호가 QRS 부분보다 상대적으로 크게 나타났다는 것을 알 수 있다. 따라서 딥러닝 기반 이상치 점수를 통해 두 그룹간의 구별되는 가장 큰 요소는 Q파 이전 부분이라고 해석 할 수 있다. 실험 B와 실험3 의 교차실험에서도 같은 경향으로 PreQ 모델에서 가장 좋은 성능을 보였다. 이는 훈련셋의 특성에 의존하지 않고 인종간 데이터의 차이 상관없이 비슷한 경향을 보인다고 해석할 수 있다.

딥러닝 기반 이상치 점수가 유의미한 특징인지 확인하기위해 우리는 실험 A에서 계산된 이상치 점수를 머신러닝 모델인 XG-Boosted 모델에 학습 및 평가를 진행했다. 그림24와 표14를 보면, 결과는 AUROC 0.98 과 F1 점수 0.94로 모델의 성능이 매우 우수했다.

Class	Precision	Recall	F1 score	Support
Normal	0.93	0.9	0.91	412
AFIB	0.93	0.95	0.94	603
Accuracy			0.93	1015
AUROC			0.98	1015

표 14 XG-Boosted 모델 분류 평가결과

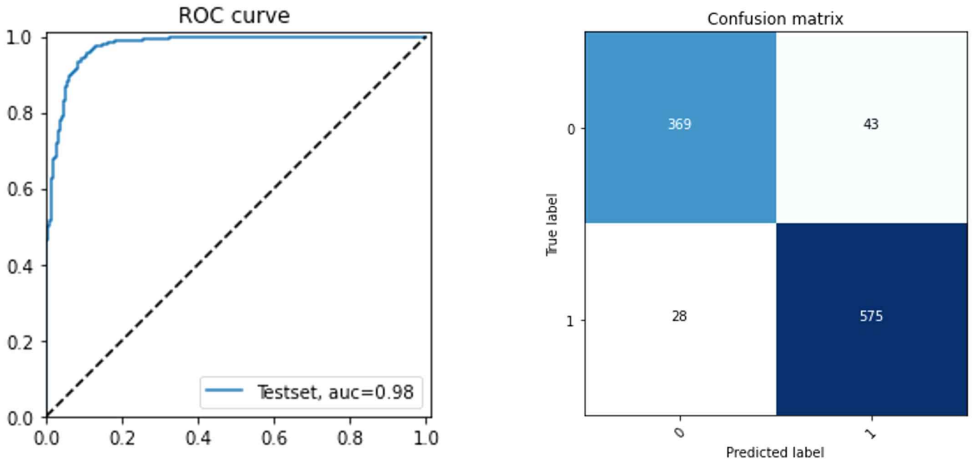

그림 24 XG-Boosted 모델 결과

심방세동을 진단한 이전 연구들과 성능을 비교하고 임상적인 해석이 가능한지에 대한 여부를 표15에서 알 수 있다. [37],[38]는 딥러닝 모델만을 사용한 방법에서 F1 점수 0.97로 우수한 성능을 보였고, [33] 또한 PTB-XL 데이터셋을 사용하여 AUROC 0.98로 성능이 매우 좋은 점수를 얻었다. 하지만 위 연구들에서 어떠한 임상적인 해석 또는 근거를 찾을 수 없어 여전히 한계점이 남아 있다. 반면, PreQ 모델만을 사용한 방법의 성능은 AUROC 0.96으로 이전 연구들과 비교해도 비슷한 성능을 보이고 있다. 게다가 이상치 점수를 가지고 XG-Boosted 모델의 성능은 AUROC 0.98로 심방세동 진단의 3가지 PreQ, QRS, PostS의 이상치 점수가 유의미한 특징임을 입증했고, 동시에 PreQ 모델과 PostS 모델이 비슷한 우수한 성능을 보인것으로 보아, 심방세동과 정상군의 차이가 QRS 부분보다는 Q파 이전 신호와 S파 이후 신호에 다름이 딥러닝 모델의 판단 근거로써 임상적 해석이 가능하다.

Study	Method	Dataset	AUROC	F1 score	Clinical explain
Kent et.al [33]	Feature extraction + DL	PTB-XL	0.98	-	x
Xu et.al [34]		MIT-BIH	0.95	-	x
Jo et.al [35]		PTB-XL	0.97	0.93	o
B Chen et.al [36]	DL	Own dataset	0.98	-	x
Anderson et.al [37]		MIT-BIH	0.94	0.97	x
Petmezas et.al [38]		MIT-BIH	-	0.97	x
Kropf et.al [39]	Feature extraction + ML	CINC (2017) dataset	-	0.81	x
Czabanski et al [40]		MIT-BIH	-	0.97	x
Our study		PTB-XL + China dataset	0.98	0.94	o
Our study	Anomaly score (PreQ)	PTB-XL + China dataset	0.96	0.92	o

표 15 이전연구 결과 및 임상적해석 가능여부

우리는 잘못 분류된 양성케이스를 분석하기 위해 한 예를 그림25에서 볼 수 있다. 그림25-(1)은 심방세동의 이상치 점수에서 가장 낮은 점수를 받아 음성으로 분류된 10초 기록을 나타냈고, 그림25-(2)는 기록의 각 세그먼트 별 이상치 점수와 역치값을 나타냈다. 대부분의 심방세동의 심전도에서는 P파의 모양이 정상과 다르게 나타내는데, 정상과 P파의 모양이 비슷하게 형성되어 이상치점수가 낮아졌다고 해석했고, 10초 기록의 R사이의 간격이 불규칙하기 때문에 앞 심전도 비트의 T파가 겹쳐졌을 경우가 생길 수 있기 때문에 이상치가 낮다고 해석될 수 있다.

하지만 본 연구는 몇가지 한계점이 있다. 첫번째로 본 연구는 단일리드만을 사용하여 심방세동 진단 시스템을 고안했다. 이는 심방세동이 아닌 다른 심장질환 진단에는 여러 리드가 요구 될 수 있기 때문에 한계점을 지닌다. 두번째는 이상치를 심방세동으로만 정의했다는 점이다. 따라서 추후 연구에는 심방세동뿐만아니라 다른 심장질환도 다뤄져야 할 필요가 있다. 세번째는 본 연구에서 임상적인 해석을 위해 심전도의 임상적으로 해석이 가능한 PreQ, QRS, PostS 만을 나눠 모델학습을 진행했지만, PR 간격이나 QT 부분 등 다양한 임상적인 부분의 모델도 고려될 필요가 있다. 마지막으로 우리는 연구에 사용된 심전도 데이터는 10초 기록으로 갑작스럽게 발생하는 심장질환에 대한 검출이 어려울수 있기 때문에 홀터(Holter) 데이터와 같이 긴 시간의 기록을 가지고 본 모델에 적용시켜 이상치를 탐지할 수 있는 모델도 고려되야 할 부분이다.

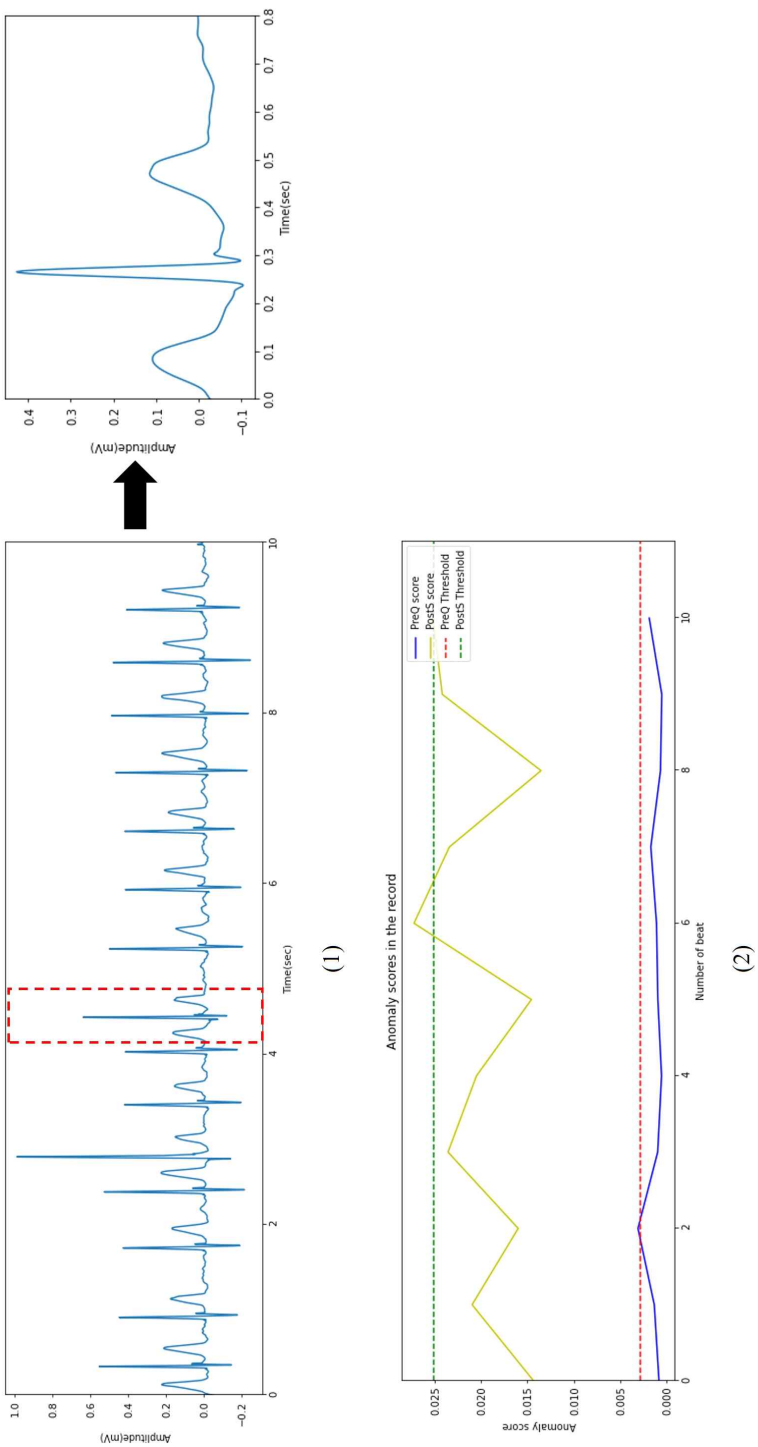

그림25 False Negative 예시 및 이상치 점수

고찰

Chapter 1에서는 제한된 의료데이터를 활용한 12리드 심전도 분류 모델을 제안하며, 데이터 불균형 문제에 대한 가중치 부여 방법을 도입하여 성능을 향상시켰다. 그러나 실제 의료데이터를 완전히 대표하지 못하고, 다양한 환자 특성을 고려한 분석이 부족하다는 한계가 존재한다. 추후 연구에서는 더 다양한 심장질환을 고려한 데이터셋과 환자 중심의 데이터 수집을 통해 일반화 성능을 향상시키는 방향으로 나아가야 할 것이다.

Chapter 2에서는 비지도 학습 방법으로 이상치 탐지 알고리즘을 활용한 심방세동(이상치) 진단 시스템을 제안하였으며, 임상적인 해석이 가능한 심전도의 주요 세 부분의 세그먼트의 모델을 학습시켜 각각의 성능 및 차이를 비교하여 임상적 해석을 할 수 있는 모델 및 알고리즘을 제안한다. 연구결과 PreQ 모델이 우수한 성능을 보였고,이는 심방세동과 정상 심전도에서 가장 크게 차이가 난다고 해석이 가능하다. 그러나 단일리드만을 사용한 점, 이상치를 심방세동으로만 정의한 점 등의 한계가 있다. 추후 연구에서는 다중 리드를 활용한 다양한 심장질환을 고려한 진단 시스템을 개발하고, 심전도의 다양한 부분을 고려하여 모델을 확장하는 방향으로 나아가야 할 것이다.

통합적으로 두 연구에서는 실제 의료환경에서의 적용 가능성을 높이기 위해 데이터 불균형과 딥러닝 임상적 해석에 대한 문제점을 해결해야하는 모델 및 알고리즘을 제안한다. 의료계에서 인공지능 모델이 더 나은 일반화 성능과 임상적 유용성에 대한 원활한 발전이 있어야 하고 이러한 통합된 접근으로 심전도 데이터를 활용한 진단 및 분류 모델의 향상된 성능이나 임상적인 실용성을 기대할 수 있을 것이다.

결론

본 연구는 Chapter1(데이터 불균형), Chapter2(임상적 해석)을 통해 의료 인공지능이 실제 의료계에 적용될 때 발생하는 문제점을 보완하는 방안을 제시한다. Chapter 1에서는 제한된 의료데이터에서 다중 리드 심전도를 활용한 8개의 클래스를 분류하는 최적의 모델을 제안한다. 데이터 불균형을 해결하기 위해 인위적인 심전도 데이터를 딥러닝 방법 또는 수학적 모델링을 통해 데이터 증강기법을 사용하지 않고 모델의 손실함수에 가중치를 주는 방법과 데이터 개수를 균형적 환경으로 만들어 분류하는 방법을 비교하였다. 가장 좋은 성능을 보인 포칼로스 함수 모델에서 F1 점수 0.96을 보였고, 데이터 개수를 최소로 맞춰 균형을 맞춘 균형 데이터 모델에서 F1 점수 0.86으로 낮은 성능을 보였다. 결과를 기반으로 모델의 손실함수에 적절한 가중치를 주어 클래스별 학습을 진행한 방법이 데이터의 개수를 임의로 바꾸는 방법보다 데이터 불균형 완화에 도움이 되었다는 것을 알 수 있다. Chapter 2에서는 이상치 탐지 기법을 활용하여 심전도 데이터를 이용한 심방세동을 진단하는 새로운 시스템을 제안한다. 이 시스템은 PreQ, QRS, PostS 세그먼트에서 이상치를 감사하고 이를 정상 심전도와 비교를 한다. PreQ 모델에서 가장 높은 AUROC 0.96, QRS 모델에서 0.75로 낮은 성능을 보였고, 이는 P파에서 Q파 까지의 부분이 심방세동을 감지하는데 중요한 특징을 가지고 있다는 것을 딥러닝 모델을 통해 알 수 있었다. 또한 PTB-XL 과 중국데이터 셋을 사용하여 모델을 훈련하고 테스트하여 교차검증을 수행한다. 이를 통해 인종에 관계없이 일반화된 모델임을 보이고, 기존 딥러닝 모델 문제에서 발생하는 데이터 종속성 문제를 해결할 수 있음을 서사한다. 의료 분야에서 딥러닝 모델기반 분석을 진행할 때, 진단의 근거가 의사 및 전문가들에게 설명 가능해야 하고 의학적으로 타당해야한다. 이점에서 이상치 점수를 기반으로 정상과 심방세동을 딥러닝 기반 모델에서의 임상적 근거를 명확히 제시했다는 점에서 다른 연구와 구별된다. 추후 연구에 다양한 심장 질환을 감지하는 시스템을 발전하여 임상적 근거를 제시하는 딥러닝 모델로 성장할 잠재력을 보여줄 수 있다.

참고문헌

[1] 한국보건산업진흥원(KKIDI),“인공지능 의료기기의 글로벌 동향” vol.441,pp5-10,2022.

[2] B. L. Jimma, "Artificial intelligence in healthcare: A bibliometric analysis," *Telematics and Informatics Reports,* p. 100041, 2023.

[3] F. Nabiyeva, S. Umarova, and S. Umirkulova, "Artificial intelligence in medicine," *Journal of new century innovations,* vol. 30, no. 3, pp. 153-155, 2023.

[4] S. Baker and W. Xiang, "Artificial Intelligence of Things for Smarter Healthcare: A Survey of Advancements, Challenges, and Opportunities," *IEEE Communications Surveys & Tutorials,* 2023.

[5] E. Strelcenia and S. Prakoonwit, "A Survey on GAN Techniques for Data Augmentation to Address the Imbalanced Data Issues in Credit Card Fraud Detection," *Machine Learning and Knowledge Extraction,* vol. 5, no. 1, pp. 304-329, 2023.

[6] A. Arora and A. Arora, "Synthetic patient data in health care: A widening legal loophole", Lancet, vol. 399, no. 10335, pp. 1601-1602, Apr. 2022.

[7] X. Yang, X. Zhang, M. Yang, and L. Zhang, "12-Lead ECG arrhythmia classification using cascaded convolutional neural network and expert feature," Journal of Electrocardiology, vol. 67, pp. 56-62, 2021.

[8] Ö. Yıldırım, P. Pławiak, R.-S. Tan, and U. R. Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ECG signals," Computers in biology and medicine, vol. 102, pp. 411-420, 2018

[9] S. Śmigiel, K. Pałczyński, and D. Ledziński, "ECG signal classification using deep learning techniques based on the PTB-XL dataset," Entropy,

vol. 23, no. 9, p. 1121, 2021.

[10] V. Gliner, N. Keidar, V. Makarov, A. I. Avetisyan, A. Schuster, and Y. Yaniv, "Automatic classification of healthy and disease conditions from images or digital standard 12-lead electrocardiograms," (in English), Sci Rep-Uk, vol. 10, no. 1, Oct 1 2020, doi: ARTN 1633110.1038/s41598-020-73060-w

[11] B. V. P. Prasad and V. Parthasarathy, "Detection and classification of cardiovascular abnormalities using FFT based multi-objective genetic algorithm," (in English), Biotechnol Biotec Eq, vol. 32, no. 1, pp. 183-193, 2018, doi: 10.1080/13102818.2017.1389303.

[12] O. Yildirim, "A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification," (in English), Computers in Biology and Medicine, vol. 96, pp. 189-202, May 1 2018, doi: 10.1016/j.compbiomed.2018.03.016.

[13] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," International Journal of Machine Learning and Computing, vol. 3, no. 2, p. 224, 2013.

[14] J. C. Shuai Ma, Weidong Xiao, and Lijuan Liu, "Deep Learning-Based Data Augmentation and Model Fusion for Automatic Arrhythmia Identification and Classification Algorithms," Computational Intelligence and Neuroscience, p. 17, 2022.

[15] N. Rastogi and R. Mehra, "Analysis of SavitzkyтGolay filter for baseline wander cancellation in ECG using wavelets," Int. J. Eng. Sci. Emerg. Technol, vol. 6, no. 1, pp. 2231-6604, 2013

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818-2826.

[17] A. H. Ribeiro et al., "Automatic diagnosis of the 12-lead ECG using a deep neural network," Nature communications, vol. 11, no. 1, pp. 1-9,

2020.

[18] N. Sakli et al., "ResNet-50 for 12-Lead Electrocardiogram Automated Diagnosis," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[19] Y. Zhou, H. Zhang, Y. Li, and G. Ning, "ECG heartbeat classification based on ResNet and BiטLSTM," in IOP Conference Series: Earth and Environmental Science, 2020, vol. 428, no. 1: IOP Publishing, p. 012014.

[20] S. S. Chugh et al., "Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study," Circulation, vol. 129, no. 8, pp. 837-847, 2014.

[21] M. Thill, W. Konen, H. Wang, and T. Bäck, "Temporal convolutional autoencoder for unsupervised anomaly detection in time series," Applied Soft Computing, vol. 112, p. 107751, 2021.

[22] J.-H. Jang, T. Y. Kim, H.-S. Lim, and D. Yoon, "Unsupervised feature learning for electrocardiogram data using the convolutional variational autoencoder,"

[23] B. Hou, J. Yang, P. Wang, and R. Yan, "LSTM-based auto-encoder model for ECG arrhythmias classification," IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 4, pp. 1232-1240, 2019.

[24] G. Zhu, H. Zhao, H. Liu, and H. Sun, "A novel LSTM-GAN algorithm for time series anomaly detection," in 2019 Prognostics and System Health Management Conference (PHM-Qingdao), 2019: IEEE, pp. 1-6.

[25] J. Qin et al., "A novel temporal generative adversarial network for electrocardiography anomaly detection," Artificial Intelligence in Medicine, p. 102489, 2023.

[26] D. Jin, E. Sergeeva, W. H. Weng, G. Chauhan, and P. Szolovits, "Explainable deep learning in healthcare: A methodological survey from an attribution view,"

WIREs Mechanisms of Disease, vol. 14, no. 3, p. e1548, 2022.

[27] N. Sobahi, O. Atila, E. Deniz, A. Sengur, and U. R. Acharya, "Explainable COVID-19 detection using fractal dimension and vision transformer with Grad-CAM on cough sounds," Biocybernetics and Biomedical Engineering, vol. 42, no. 3, pp. 1066-1080, 2022.

[28] S. Vijayarangan, B. Murugesan, R. Vignesh, S. Preejith, J. Joseph, and M. Sivaprakasam, "Interpreting deep neural networks for single-lead ECG arrhythmia classification," in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020: IEEE, pp. 300-303.

[29] P. Singh and A. Sharma, "Attention-based convolutional denoising autoencoder for two-lead ECG denoising and arrhythmia classification," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-10, 2022.

[30] P. Wagner et al., "PTB-XL, a large publicly available electrocardiography dataset," Scientific data, vol. 7, no. 1, p. 154, 2020.

[31] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," Scientific data, vol. 7, no. 1, p. 48, 2020.

[32] L. Sathyapriya, L. Murali, and T. Manigandan, "Analysis and detection R-peak detection using Modified Pan-Tompkins algorithm," in 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, 2014: IEEE, pp. 483-487.

[33] M. Kent, L. Vasconcelos, S. Ansari, H. Ghanbari, and I. Nenadic, "Fourier space approach for convolutional neural network (CNN) electrocardiogram (ECG) classification: A proof-of-concept study," Journal of Electrocardiology, vol. 80, pp. 24-33, 2023.

[34] X. Xu, S. Wei, C. Ma, K. Luo, L. Zhang, and C. Liu, "Atrial fibrillation beat identification using the combination of modified frequency slice wavelet transform and convolutional neural networks,"

Journal of healthcare engineering, vol. 2018, 2018.

[35] Y.-Y. Jo et al., "Detection and classification of arrhythmia using an explainable deep learning model," Journal of Electrocardiology, vol. 67, pp. 124-132, 2021.

[36] B. Chen et al., "A deep learning model for the classification of atrial fibrillation in critically ill patients," Intensive Care Medicine Experimental, vol. 11, no. 1, pp. 1-10, 2023.

[37] R. S. Andersen, A. Peimankar, and S. Puthusserypady, "A deep learning approach for real-time detection of atrial fibrillation," Expert Systems with Applications, vol. 115, pp. 465-473, 2019.

[38] G. Petmezas et al., "Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets," Biomedical Signal Processing and Control, vol. 63, p. 102194, 2021.

[39] M. Kropf, D. Hayn, and G. Schreier, "ECG classification based on time and frequency domain features using random forests," in 2017 Computing in Cardiology (CinC), 2017: IEEE, pp. 1-4.

[40] R. Czabanski et al., "Detection of atrial fibrillation episodes in long-term heart rhythm signals using a support vector machine," Sensors, vol. 20, no. 3, p. 765, 2020.

영문초록 (Abstract)

With the advancement of the Fourth Industrial Revolution, various studies have been underway due to the formation of big data in recent medical data.Studies utilizing time-series data such as Electrocardiogram (ECG) and Photoplethysmography (PPG), in addition to medical imaging data, have been actively conducted. Recently, the extensive accumulation of such data has led to significant advancements driven by the application of Deep Learning technology.With the advancement of medical artificial intelligence technology, there are inherent limitations. The prominent ones include data imbalance and challenges arising from clinical interpretation. Firstly, data imbalance is a common occurrence in medical data. It adversely affects the performance of minority classes in classification model training, which can have critical implications in medical data. Secondly, challenges stemming from clinical interpretation arise from the uncertainty in deep learning analysis methods. The term 'black box' is frequently used to describe deep learning interpretation, indicating that it's not always clear how the deep learning model arrived at its analysis given the input data.Therefore, this study aims to address these issues. It conducts deep learning analysis using electrocardiogram data. To mitigate data imbalance resulting from the analysis, four experiments are conducted to compare performance and identify the optimal approach. Moreover, to tackle challenges arising from clinical interpretation, the study utilizes unsupervised learning to train on three key segments of normal ECG data. Additionally, it calculates outliers in atrial fibrillation ECG data, emulating the diagnostic mechanism of actual medical professionals, and employs deep learning for clinical interpretation.

In the chapter one for addressing data imbalance, actual clinical data from 7,355 individuals with multi-lead electrocardiograms is utilized. In this phase, a ResNet model is employed as the

classifier, trained through supervised learning methods to classify eight different conditions.

In chapter two for the clinical interpretation, a dataset comprising 9,042 individuals from PTB-XL and 7,199 individuals from China is used. This dataset includes both normal cases and cases of atrial fibrillation. Within this section, the normal electrocardiograms are divided into three main segments: pre-Q wave (preQ), QRS complex (QRS), and post-S wave (postS). An autoencoder model is utilized for training. Furthermore, the model is evaluated using electrocardiograms from patients with atrial fibrillation. This evaluation involves calculating outliers to mimic the diagnostic mechanism of actual medical professionals, allowing for the classification of normal and atrial fibrillation cases.