

GANs를 사용한 X-Ray 이미지 질병 인식 성능 향상

2020-1 데이터분석캡스톤디자인 최종 발표

소프트웨어융합학과 유태원

2020/6/26

목차

1. 문제 인식
2. 해결 방안
3. 실험 방법
4. 실험 결과
5. 결론 및 향후 연구 계획

문제 인식

의료 데이터 분석 기술의 향상

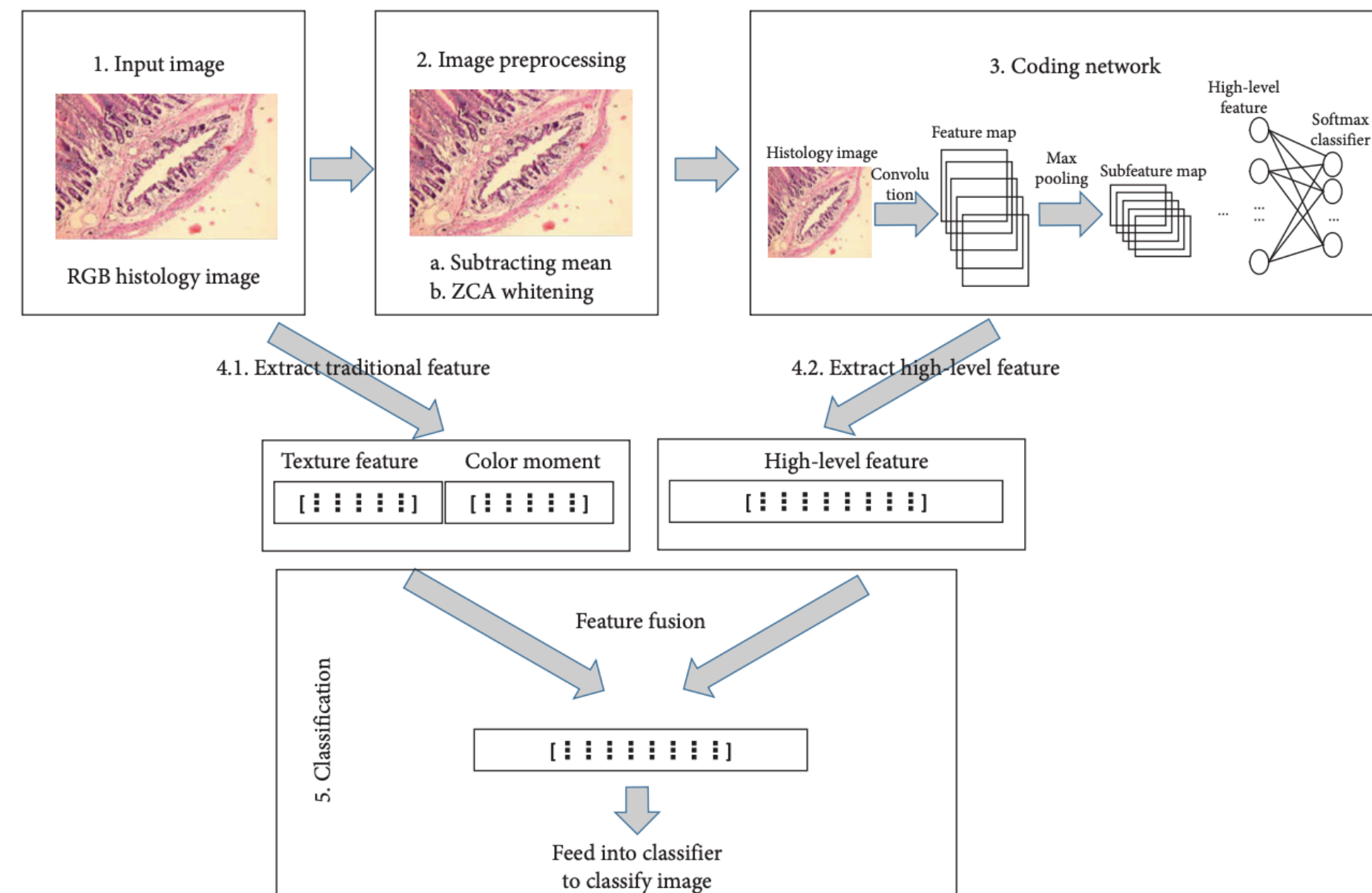


FIGURE 1: Framework of the approach.

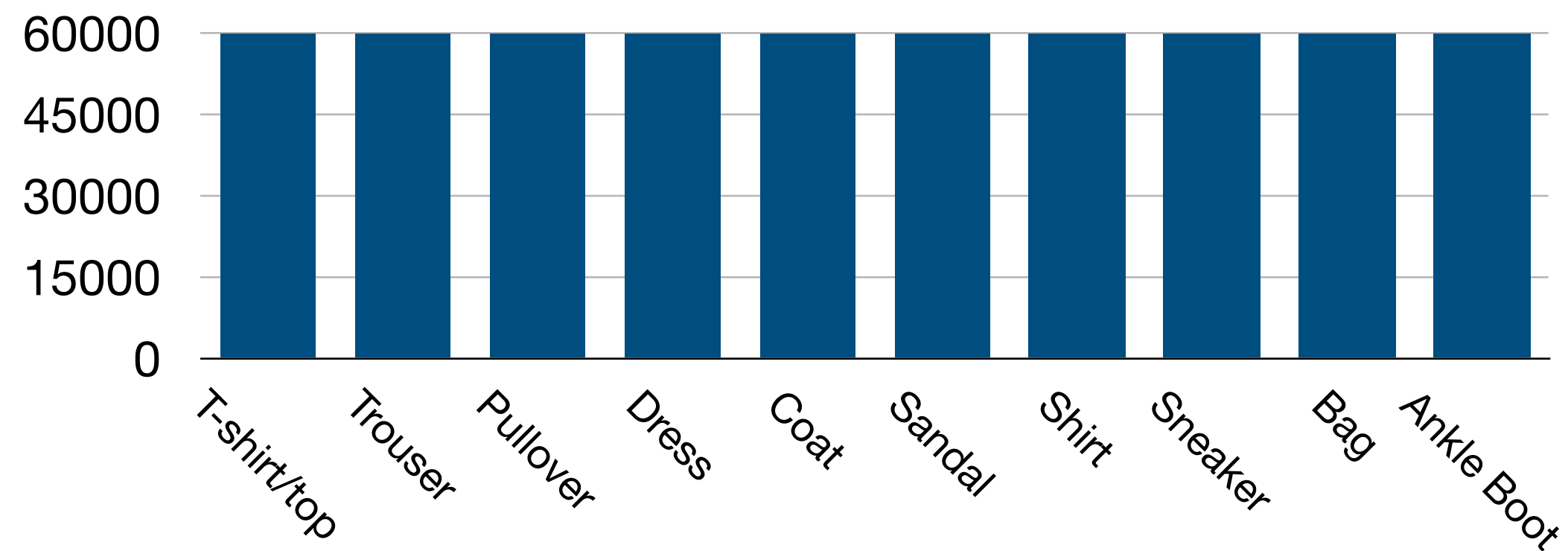
- 의료 이미지 데이터 분석은 환자의 질병을 미리 예측하고 예방하는 데 매우 중요하다.
- 최근에는 end-to-end 딥러닝모델을 사용한 의료 이미지 질병 분류 기술이 발전하고 있다.

의료 이미지를 사용한 분류 예시

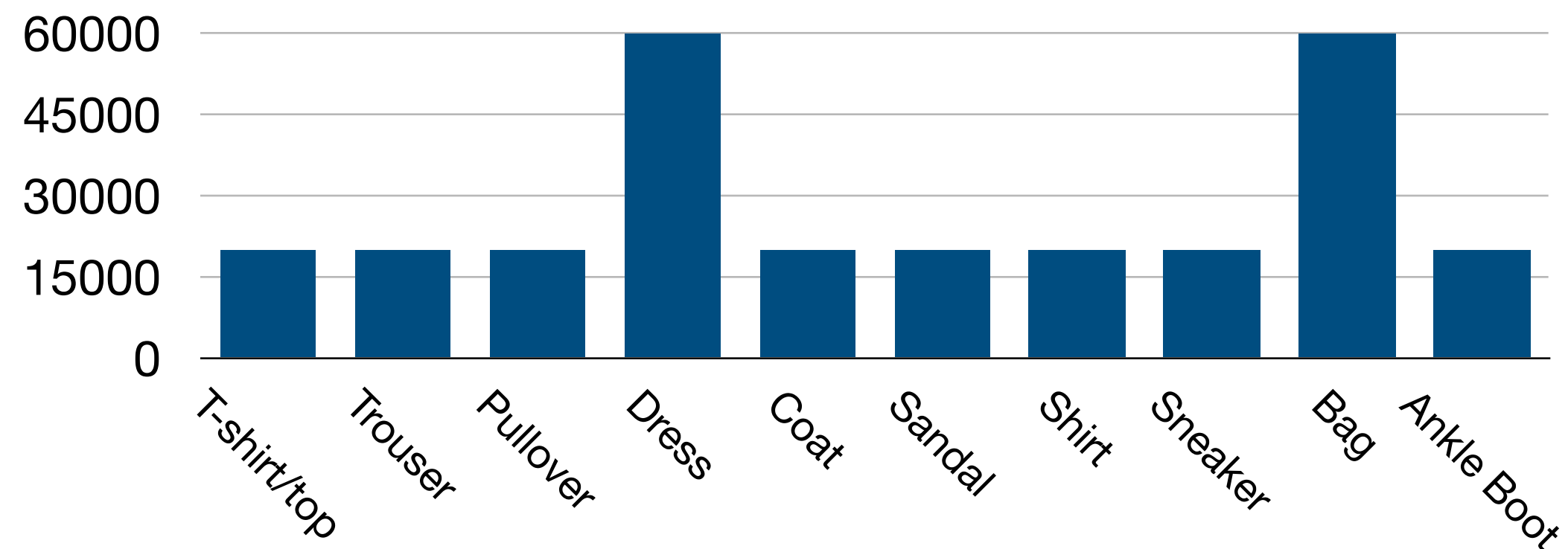
문제 인식

Data Imbalance 문제

Fashion MNIST Balanced Dataset

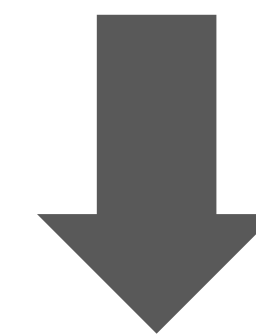


Fashion MNIST Imbalanced Dataset



Imbalance Ratio	Accuracy (%)
Balanced (1:1)	89.94
10:1	84.21
20:1	81.31
40:1	77.13

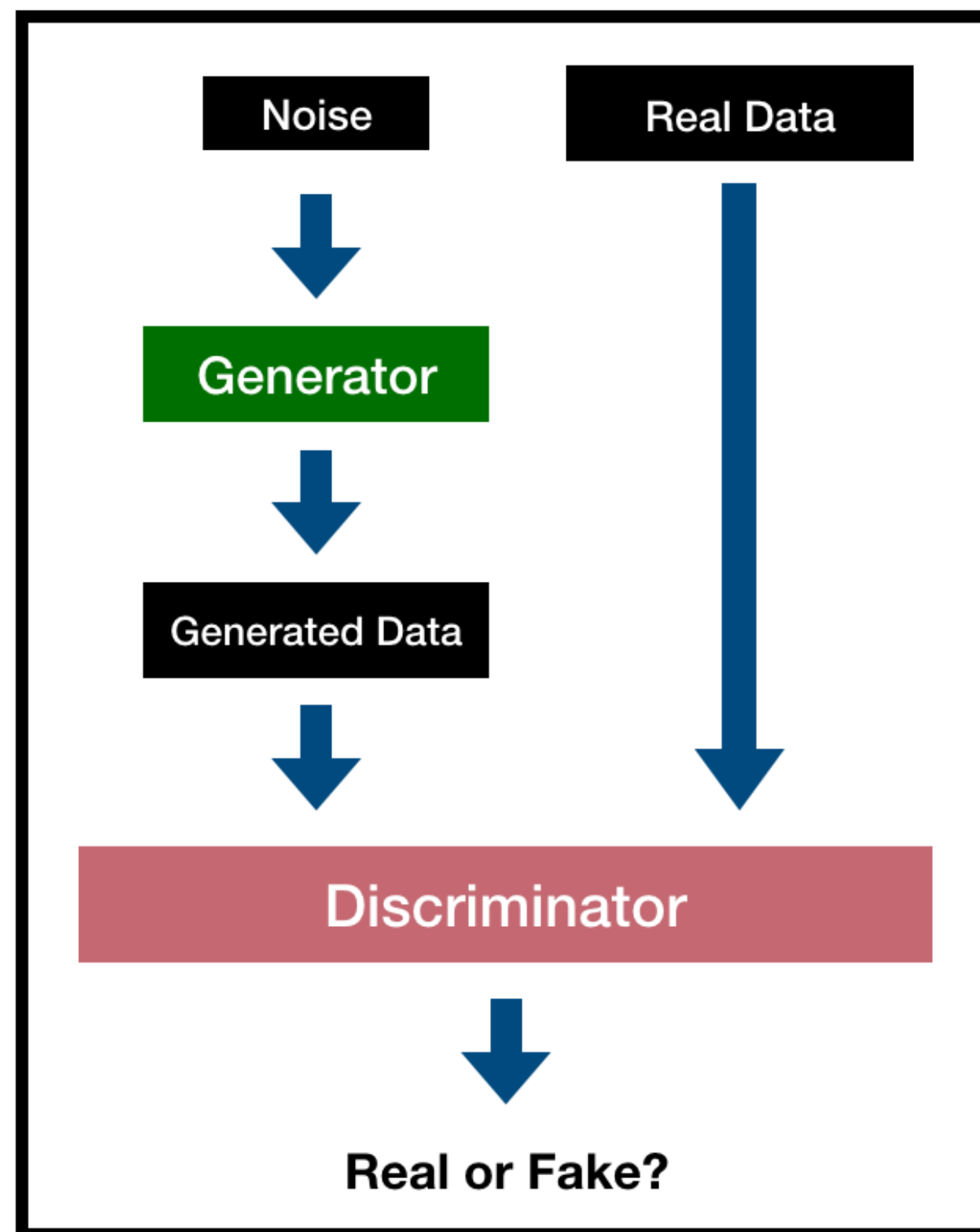
Fashion-MNIST 이미지를 LeNet 5 모델로 학습했을 때,
 imbalance ratio에 따른 분류 정확도
 10개의 클래스 중 2개가 major, 나머지가 minor
 major class 데이터 수: 6000



**Data imbalance problem을
 해결할 방법이 필요하다.**

해결 방안

Generative Adversarial Networks(GANs)를 사용한 Data Augmentation

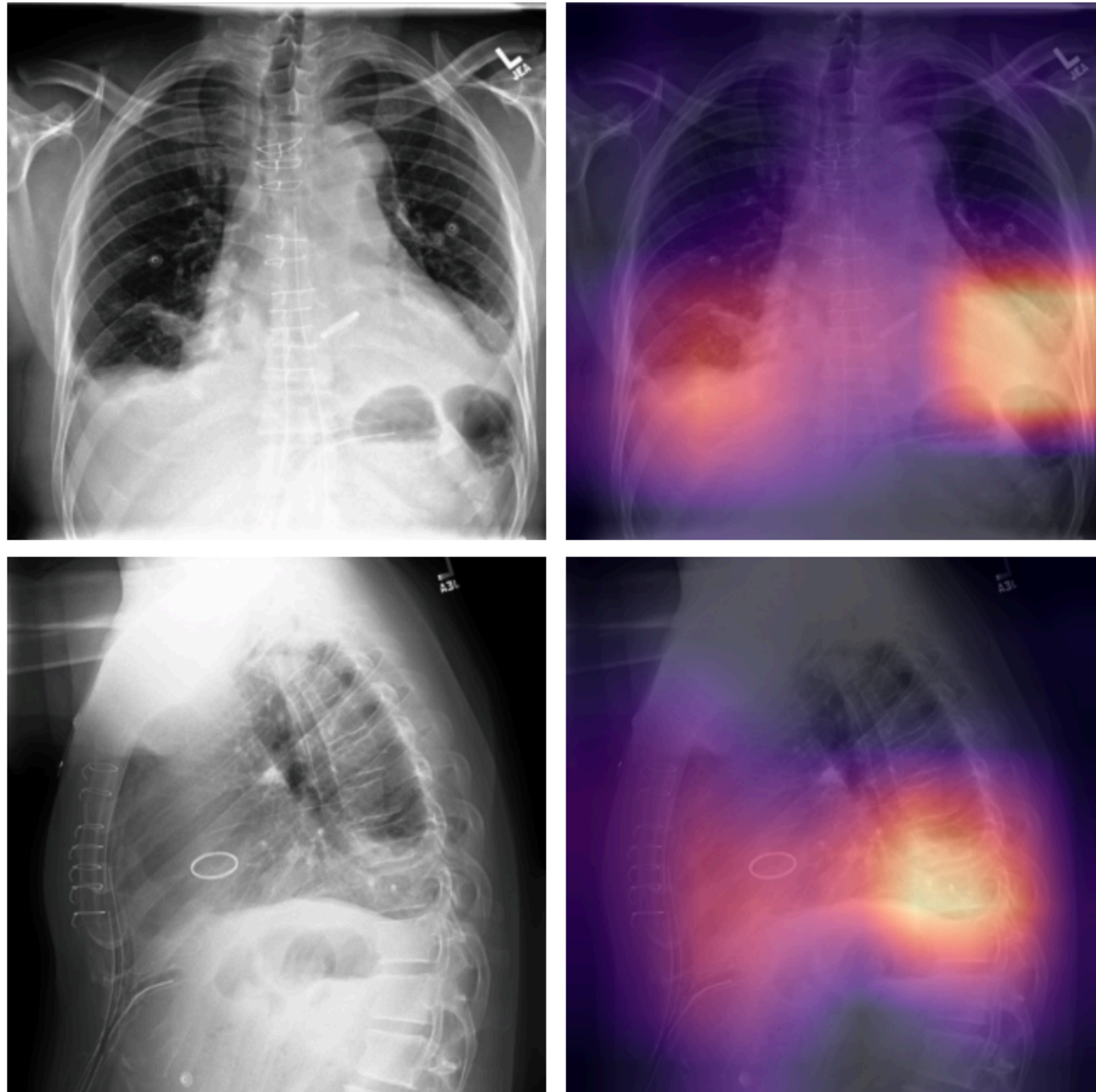


GAN 구조

- GANs이란, 가짜 데이터를 생성하는 Generator와 데이터가 진짜인지 가짜인지 구별하는 Discriminator가 적대적으로 학습하여, 궁극적으로 진짜 데이터와 비슷한 가짜 데이터를 생성하는 모델이다.
- GANs을 통해 부족한 데이터를 증가시켜 Data Imbalance 문제를 해결할 수 있다.
- GANs로 실제 데이터를 생성하기 보다, 각 클래스의 특징을 뽑은 Feature를 생성하면 분류 모델 학습에 효과가 더 좋을 것이다.

실험 방법

CheXpert Dataset



CheXpert 데이터셋 예시
좌(정면), 우(측면)

- CheXpert 데이터셋은 흉부 X-Ray 데이터 해석, 질병 레이블 분류 등을 목적으로 Stanford University ML Group에서 제공하는 데이터이다.
- 이 데이터셋에는 총 224,316개의 흉부 X-Ray 이미지가 있으며, “질병 없음”을 포함해 총 14개의 클래스로 구성되어있다.

실험 방법

CheXpert Dataset

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

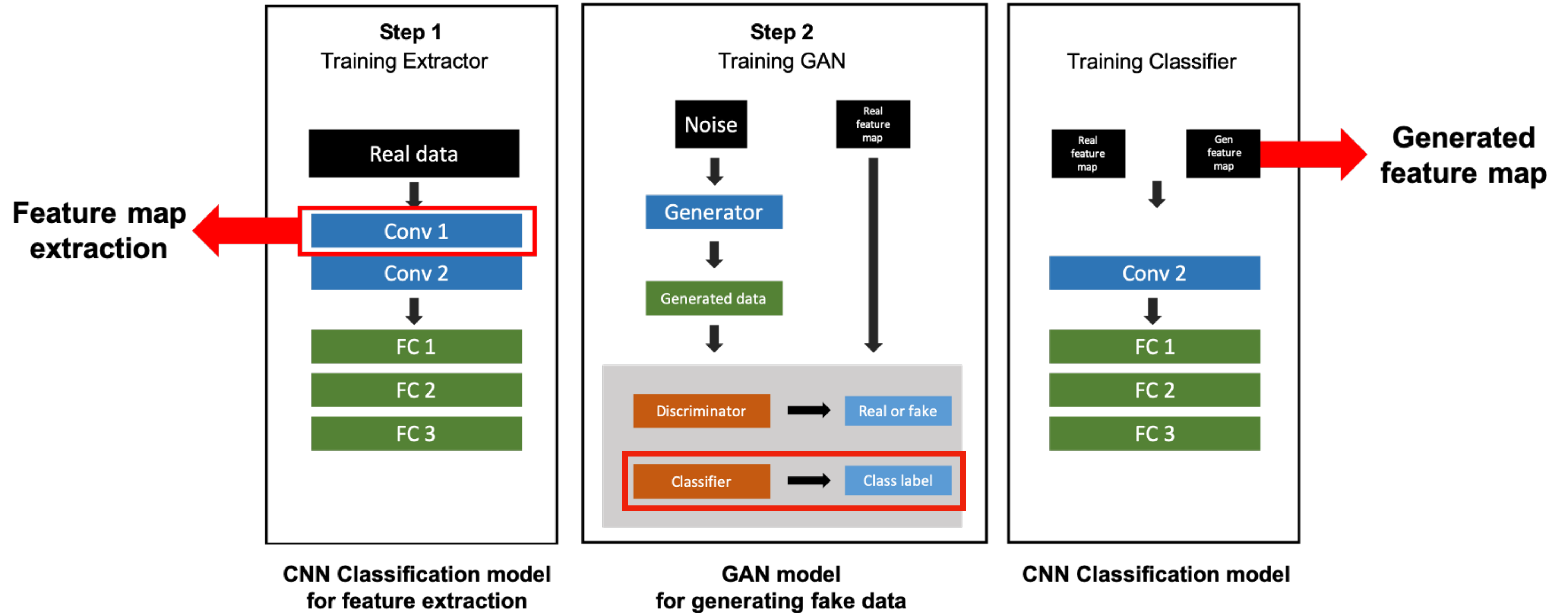
CheXpert 데이터셋의 클래스 별 데이터 수

- 표에서 볼 수 있듯이, CheXpert 데이터의 클래스 별 데이터가 Imbalanced 상태임을 알 수 있다.
- 이 중 “질병 없음” 클래스와 다섯 개의 질병 클래스를 선정하여, 6개 클래스의 전면 흉부 X-Ray 이미지 데이터를 imbalanced 상태로 만들어 실험 데이터셋을 구성하였다.

클래스 이름	데이터 수 (train + test)
No Finding	3,000 + 400
Atelectasis	5,400 + 400
Cardiomegaly	5,400 + 400
Consolidation	2,700 + 400
Edema	9,500 + 400
Pleural Effusion	13,500 + 400
합	39,500 + 2,400

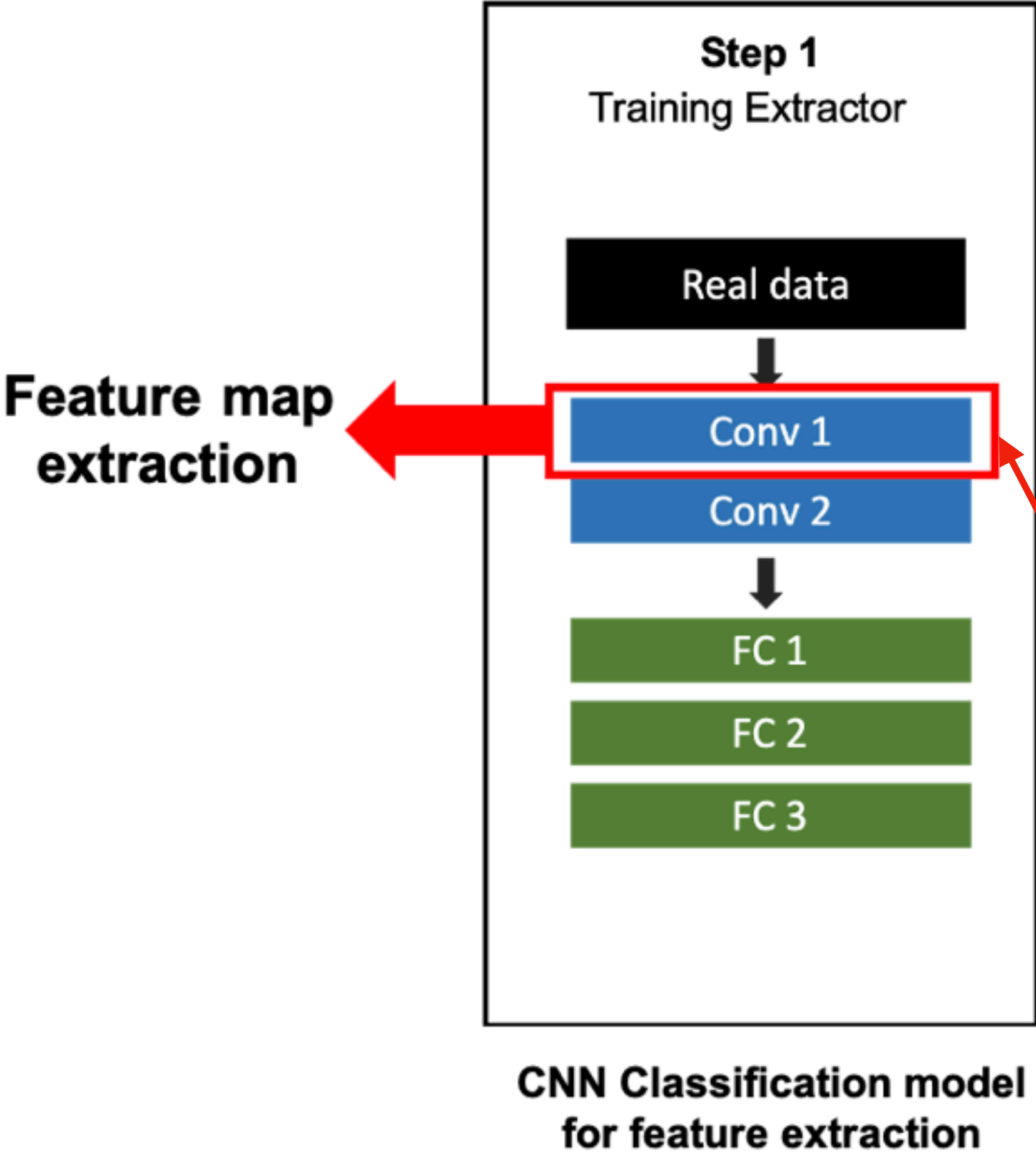
실험 방법

모델 학습 Framework



실험 방법

Step 1 Extractor 학습: GANs 모델의 input으로 들어갈 데이터의 feature를 생성하기 위해, VGGNET 분류기를 학습한다.



Type	Patch Size	Output Size
Input		160 x 160 x 3
Convolution	3 x 3 x 64	160 x 160 x 64
Convolution	3 x 3 x 64	160 x 160 x 64
Max Pool	2 x 2	80 x 80 x 64
Convolution	3 x 3 x 128	80 x 80 x 128
Convolution	3 x 3 x 128	80 x 80 x 128
Max Pool	2 x 2	40 x 40 x 128
Convolution	3 x 3 x 128	40 x 40 x 128
Convolution	3 x 3 x 128	40 x 40 x 128
Convolution	3 x 3 x 256	40 x 40 x 256
Max Pool	2 x 2	20 x 20 x 256
Convolution	3 x 3 x 256	20 x 20 x 256
Convolution	3 x 3 x 256	20 x 20 x 256
Convolution	3 x 3 x 512	20 x 20 x 512
Max Pool	2 x 2	10 x 10 x 512
Convolution	3 x 3 x 512	10 x 10 x 512
Convolution	3 x 3 x 512	10 x 10 x 512
Convolution	3 x 3 x 512	10 x 10 x 512
Max Pool	2 x 2	5 x 5 x 512
Flatten		1 x 1 x 12,800
Fully Connected	4,096	1 x 4,096
Fully Connected	512	1 x 512
Fully Connected	6	1 x 6

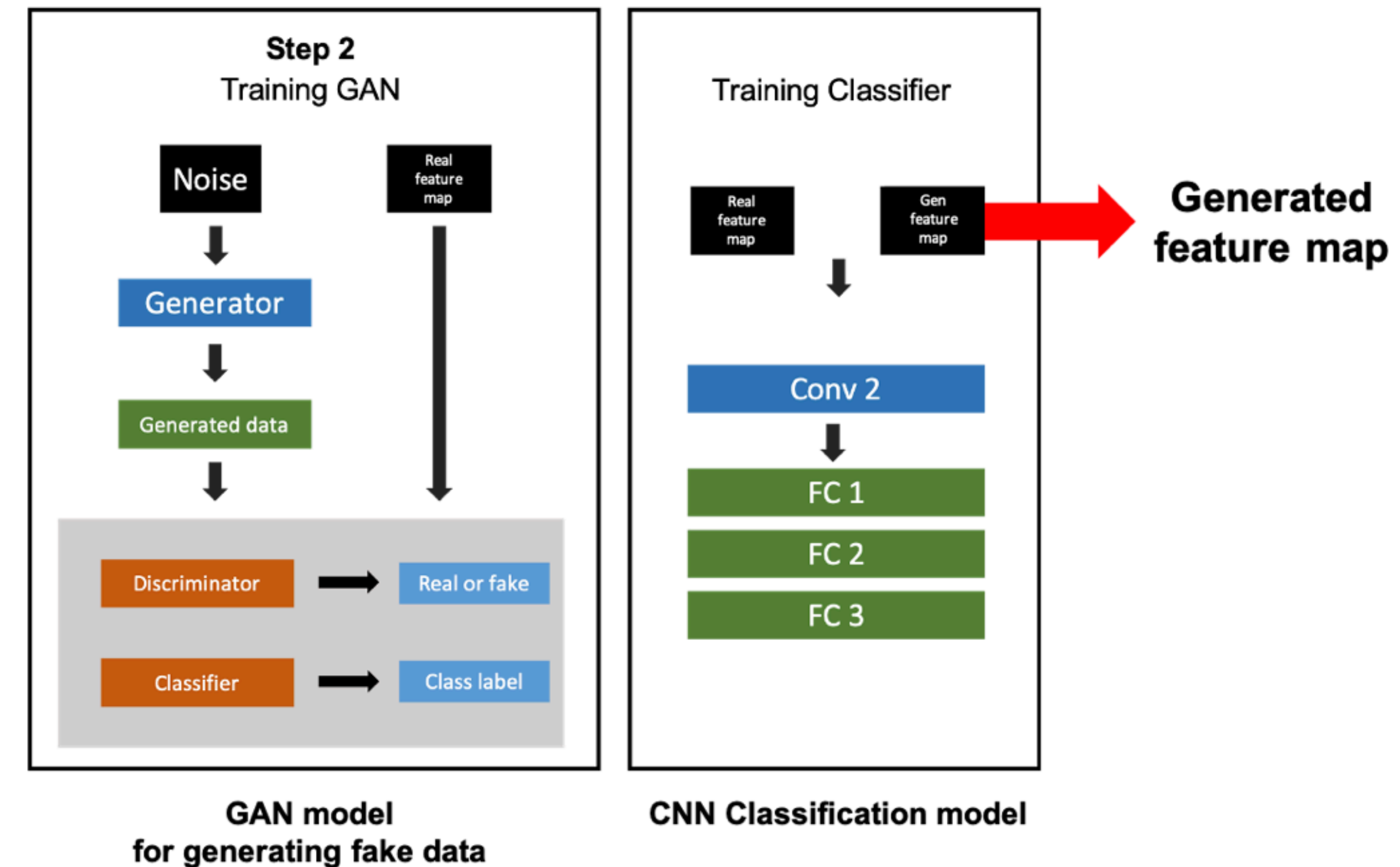
실험 방법

Step 2 GANs 학습: Step1에서 생성한 Feature로 GANs 모델을 학습한다.

Generator: 실제 데이터와 비슷한 데이터를 생성하도록 학습하는 모델

Discriminator: 실제 데이터 또는 가짜 데이터를 입력받아, 데이터가 진짜인지 가짜인지 판별하는 모델

Classifier: 실제 데이터 또는 가짜 데이터를 입력받아, 데이터의 클래스를 분류하는 모델 (Extractor 모델의 뒷 부분)



Step 2 GANs 학습: Step1에서 생성한 Feature로 GANs 모델을 학습한다.

Generator

Type	Patch Size	Output Size
Input		1 x 100
Flatten		1 x 1 x 2048
Deconvolution	1 x 1 x 2048	1 x 1 x 2048
Deconvolution	5 x 5 x 1024	5 x 5 x 1024
Deconvolution	10 x 10 x 512	10 x 10 x 512
Deconvolution	10 x 10 x 512	10 x 10 x 512

Discriminator

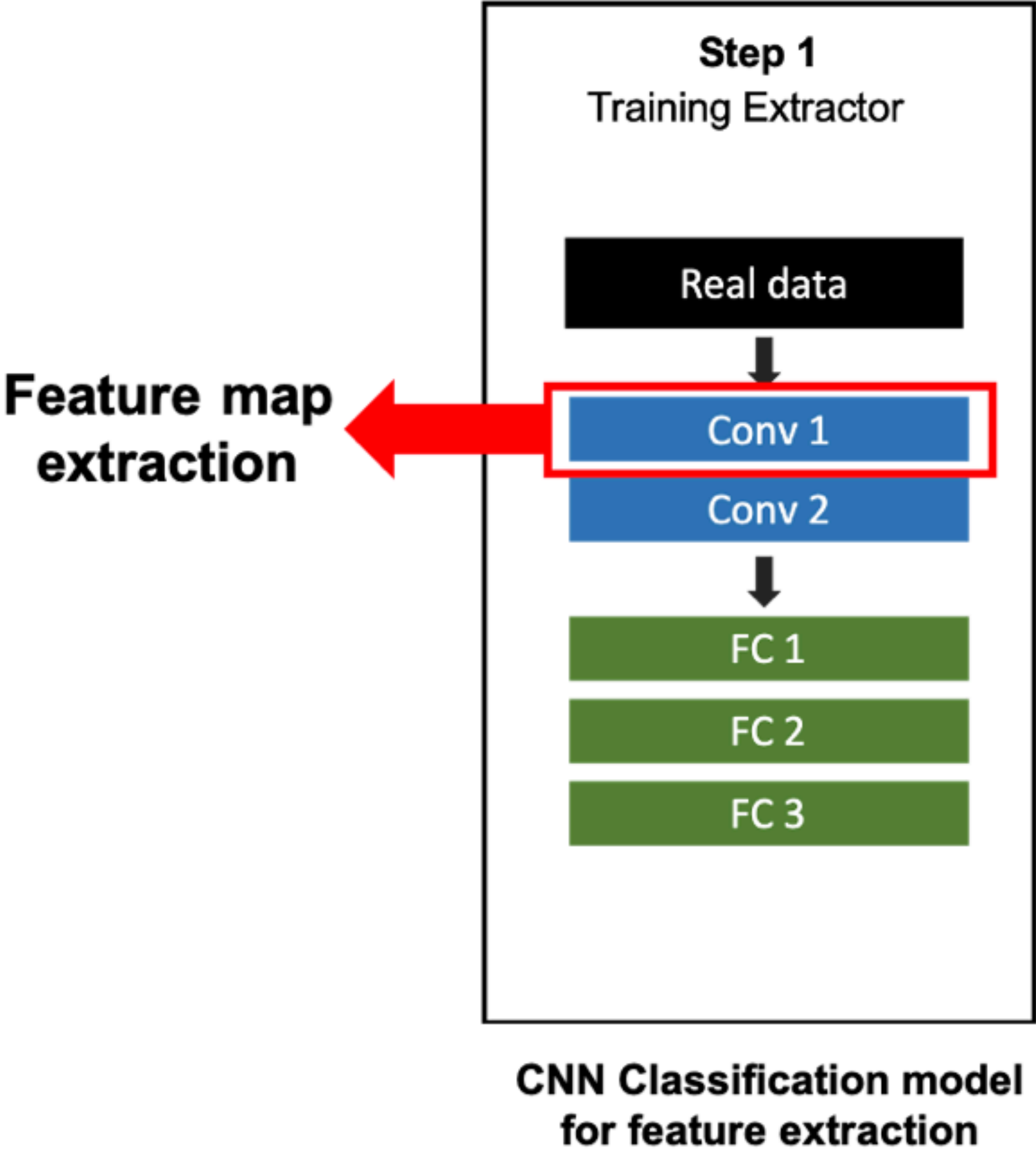
Type	Patch Size	Output Size
Input		10 x 10 x 512
Convolution	5 x 5 x 512	10 x 10 x 512
Convolution	5 x 5 x 1024	5 x 5 x 1024
Convolution	5 x 5 x 2048	2 x 2 x 2048
Flatten		1 x 1 x 8,192
Fully Connected	1,024	1 x 1,024
Fully Connected	64	1 x 64
Fully Connected	1	1 x 1

Classifier

Type	Patch Size	Output Size
Input		10 x 10 x 512
Convolution	3 x 3 x 512	10 x 10 x 512
Convolution	3 x 3 x 512	10 x 10 x 512
Convolution	3 x 3 x 512	10 x 10 x 512
Max Pool	2 x 2	5 x 5 x 512
Flatten		1 x 1 x 12,800
Fully Connected	4,096	1 x 4,096
Fully Connected	512	1 x 512
Fully Connected	6	1 x 6

실험 결과

Step 1 Extractor 학습



Parameter	Value
Learning Rate	0.0001
Batch Size	40
Epoch	50

➡ Classification Accuracy
56.8%
Baseline 정확도로 사용

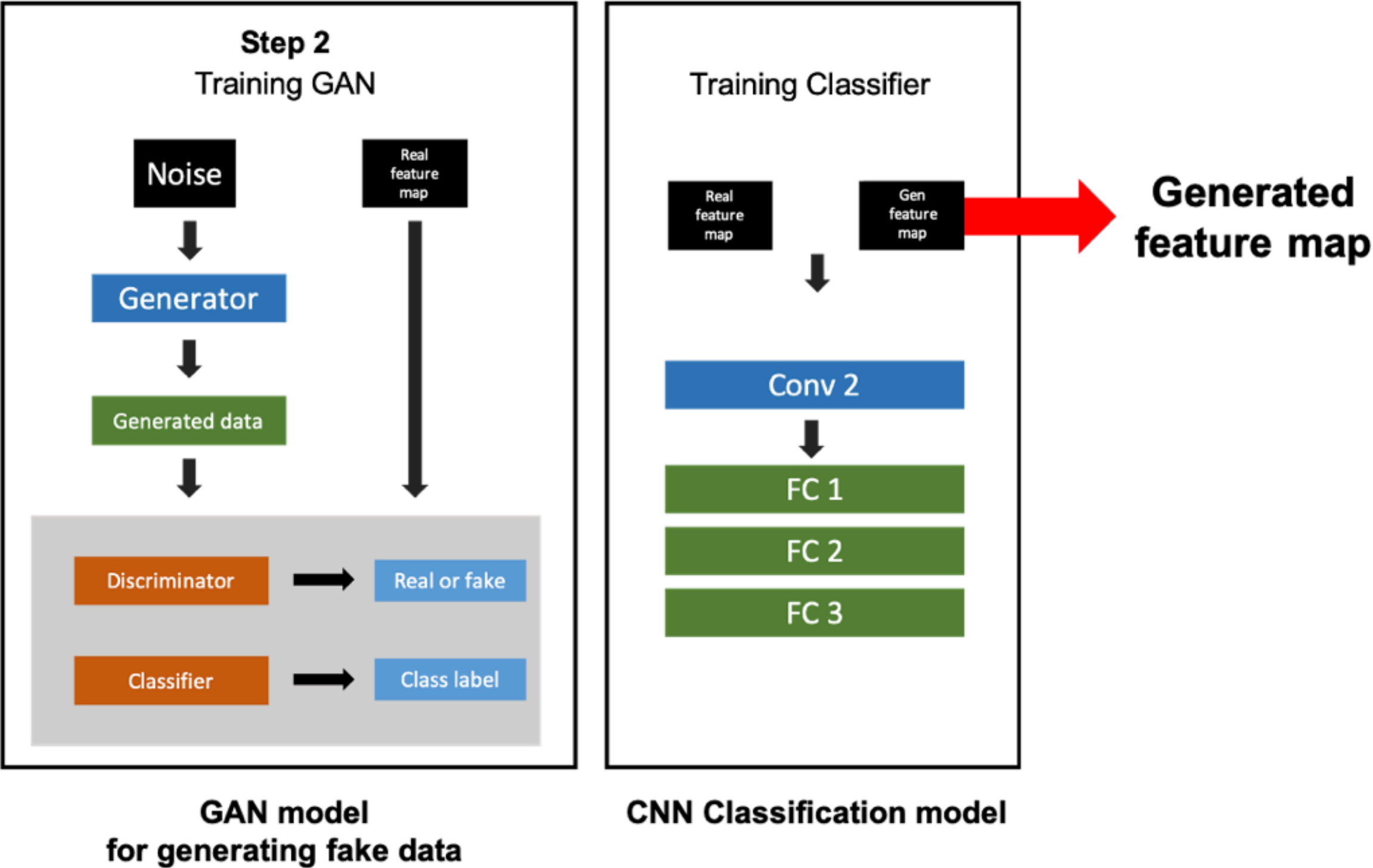
실험 결과

Step 2 GANs 학습

Classification Accuracy
56.8% -> **61.7%**

Parameter	Value
Generator Learning Rate	0.001
Discriminator Learning Rate	0.0001
Classifier Learning Rate	0.000001
Batch Size per Step	40

2 step 마다 가짜 데이터로 Classifier 학습
5 step 마다 진짜 데이터로 Classifier 학습



결론 및 향후 연구 계획

- GANs를 사용하여 Data Imbalance 상황에서 X-Ray 이미지 질병 분류의 정확도를 향상시킬 수 있다.
- Classifier를 사용하여 Multi Class GANs 모델 학습에 도움을 줄 수 있다.
- Step1에서 Classification Accuracy 향상을 위해 다른 모델과 학습방법을 적용시킬 수 있다.
- Step2에서 GANs 모델과 학습방법을 수정하면 Classification 성능 향상 가능성이 있다.
- 다른 의료 데이터 (MRI, CT 데이터 등)를 분석할 때 발생하는 Data Imbalance 문제를 해결할 수 있다.

감사합니다.