



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](#)

공학석사 학위논문

부비동염 엑스레이 진단 성능의  
최적화를 위한 생성적 적대 신경망  
이용 합성데이터 생성에 관한 연구

A Study on the Generation of Synthetic Data  
using Generative Adversarial Network for  
Optimization of Diagnostic Performance on  
Paranasal Sinusitis Radiography

2021년 8월

서울대학교 대학원

협동과정 바이오엔지니어링전공

문혜민

부비동염 엑스레이 진단 성능의  
최적화를 위한 생성적 적대 신경망  
이용 합성데이터 생성에 관한 연구

지도 교수 Sungwan Kim

이 논문을 공학석사 학위논문으로 제출함  
2021년 5월

서울대학교 대학원  
협동과정 바이오엔지니어링전공  
문혜민

문혜민의 공학석사 학위논문을 인준함  
2021년 6월

위원장 \_\_\_\_\_ 이정찬

부위원장 \_\_\_\_\_ Sungwan Kim

위원 \_\_\_\_\_ 공현중

# 국문초록

문혜민

서울대학교 대학원

협동과정 바이오엔지니어링 전공

딥러닝 기술은 최근들어 다양한 의료 분야에 적용되고 있다. 딥러닝 모델의 효과적인 학습을 위해서는 방대한 양의 데이터가 필요한데, 실제로 환자가 의료기관에서 받은 검사를 통해 얻어지는 의료 데이터의 특성상 데이터가 부족한 경우가 많으며, 이는 딥러닝 기술을 의료 분야에 적용함에 있어 큰 한계점이다. 이러한 학습 데이터 부족을 극복하기 위한 다양한 연구들이 현재까지도 활발히 진행되고 있다. 본 연구에서는 학습용 의료 데이터가 부족한 상황에서 우수한 판독 성능에 도달할 수 있도록 생성적 적대 신경망 (Generative Adversarial Network, GAN)을 이용하여 합성 데이터 증강 (data augmentation)을 수행하고 가장 우수한 성능에 최적화된 모델을 자동적으로 찾아 줄 수 있는 방법을 제안한다.

기존의 관련 연구에서는 GAN 모델을 이용하여 합성 데이터를 생성하고 데이터 증강 개수가 학습 모델의 성능에 미치는 영향에 대한 실험을 진행하였다. 앞선 연구는 데이터 증강 샘플을 일정한 간격없이 임의적으로 증가시켰기 때문에 정량적이지 않다는 점과 적절한 양을 탐색하기 위해 증강 개수를 매번 조정해야되는 수동 작업이 개입되므로 상당한 시간이 소요된다는 한계점이 있다.

본 연구에서는 GAN 기반의 합성 데이터 증강을 수행하는 과정에서 어느 시점에서 가장 최적화된 모델 성능이 나타나는지에

대하여 자동화 개념으로 접근한다. 연구 데이터는 2차 병원에서 수집한 부비동 (Paranasal Sinuses, PNS) X-ray 이미지와 일반성을 확보하기 위해 외부 검증 (external validation)용으로 3차 병원에서 수집한 PNS X-ray 이미지로 구성되어 있다. 실험에 사용될 모든 데이터는 전처리 과정에서 상악동 (maxillary sinus) 부위를 기준으로 잘라내어 (cropping) 영상 패치를 생성하였다.

데이터 증강 기법은 두 가지의 방법으로 수행하였다. 첫 번째 방법은 전통적 데이터 증강 기법 (Conventional data augmentation)에서 흔히 사용되는 회전 (-10도에서 10도 범위 내로)과 좌우 뒤집기 (horizontal flipping) 를 적용하였고, 두 번째 방법으로는 GAN을 이용하여 합성 데이터를 생성 후 원본 데이터와 결합시켜 증강을 하는 합성 데이터 증강 기법이다. GAN기반의 합성 데이터 증강을 수행하기 위해서는 Auxiliary Classifier GAN (ACGAN) 모델을 기존 학습용 데이터에 적용하여 새로운 데이터를 생성하였다. 본 과정을 통해 생성된 데이터를 일정한 비율로 증가시키며, 이를 바탕으로 학습된 딥러닝 모델이 가장 높은 성능을 보이는 최적의 비율 (optimal multiples)을 발견할 수 있도록 자동화 방법을 구축하였다.

이미지 분류 과정에서는 CheXNet 모델을 이용하여 전이 학습 (transfer learning)을 수행하였고 이후, 본 연구의 모델 성능을 평가하기 위해서 전통적 데이터 증강 기법을 이용하여 학습한 모델의 성능과 GAN 기반의 합성 데이터 증강 기법의 최적의 증강 비율로 학습한 모델의 성능을 비교하였다. 더불어 두 모델의 성능 간의 통계적 유의성을 제시하였다.

모델 성능 평가 결과, 본 연구에서 제안한 GAN 기반의 데이터 증강 기법 모델이 전통적 데이터 증강 기법보다 내부 및 외부 검증에서

모두 우수한 성능을 보였음을 입증할 수 있었다.

본 연구에서 제시한 GAN 기반의 데이터 증강 기법은 X-ray 영상 뿐만 아니라 다양한 의료 영상에서도 적용될 수 있으며, 증강시킬 데이터 양의 크기를 수동으로 설정해주는 이전 방식과는 달리 일정한 배율로 증가시키며 가장 우수한 성능에 최적화된 모델 및 증강 배율을 자동적으로 찾아주기 때문에 편리성과 정확성을 제공할 것이다. 추후 데이터 증강 기법 관련 연구에서 GAN 기반의 합성 데이터 증강 자동화 알고리즘이 더 활용될 것으로 기대한다.

**주요어** : 데이터 증강 기법, 보조 진단 시스템, 적대적 신경망 네트워크, 인공지능

**학 번** : 2018-20423

# 목 차

제 1 장 서론 .....	1
제 1 절 연구의 배경 .....	1
제 2 절 관련 연구의 동향 .....	3
제 3 절 관련 연구의 한계점 .....	5
제 4 절 연구의 목적 .....	6
제 2 장 연구 방법 .....	6
제 1 절 데이터 수집 및 구성 .....	6
제 2 절 데이터 전처리 과정 .....	10
제 3 절 ACGAN 모델을 이용한 데이터 생성 .....	13
제 4 절 데이터 증강 기법 .....	18
제 5 절 분류 모델 학습 과정 .....	23
제 6 절 모델 성능 평가 및 유의성 검증 방법 .....	24
제 3 장 연구 결과 .....	26
제 1 절 GAN 기반의 데이터 증강 기법의 최적 배율 .....	27
제 2 절 모델 성능 평가 .....	35
제 4 장 고찰 .....	40
제 5 장 결론 .....	45
참고문헌 .....	46
Abstract .....	49

## 표 목차

[표 2.1] 부비동 x-ray 데이터 구성.....	11
[표 3.1] 내부 테스트 세트를 이용하여 학습 데이터 개수의 증가에 따른 CheXNet 모델의 성능 평가.....	28
[표 3.2] CheXNet 모델의 내부 및 외부 검증 결과 .....	37

## 그림 목차

[그림 2.1] 연구 대상 선정 기준 및 제외 기준 .....	8
[그림 2.2] 부비동 x-ray 데이터 샘플 및 의사 소견 .....	9
[그림 2.3] 이미지 패치 생성 과정 .....	12
[그림 2.4] GAN, CGAN, ACGAN 모델 비교.....	15
[그림 2.5] ACGAN 아키텍처와 부비동 x-ray 분류 방법 .....	16
[그림 2.6] ACGAN 모델을 통해 생성된 이미지 샘플.....	17
[그림 2.7] 전통적 데이터 증강 기법을 적용한 이미지 샘플.....	19
[그림 2.8] 슬라이딩 윈도우를 이용한 GAN 기반의 데이터 증강 기법 순서도.....	22
[그림 2.9] 혼동 행렬 .....	25
[그림 3.1] 표 3.1에 대한 산점도 (scatter plot).....	31
[그림 3.2] 각 배율에 대한 CheXNet 모델의 ROC 곡선.....	32
[그림 3.3] 슬라이딩 윈도우 시각화.....	34
[그림 3.4] CheXNet 모델의 내부 및 외부 검증 결과에 대한 ROC 곡선 .....	38
[그림 3.5] 통계적 유의성 검증 결과 .....	39
[그림 4.1] MS-SSIM 결과.....	42



# 제 1 장 서론

## 제 1 절 연구의 배경

현재 인공지능의 빠른 발전은 다양한 산업분야로의 적용으로 확산되고 있으며, 경제, 자동차의 자율주행, 자연어 처리 (natural language processing) 및 컴퓨터 비전 (computer vision) 분야에서 상당한 발전과 주목을 이끌고 있다[1]. 인공지능의 여러 기술 중 최근 가장 각광받고 있는 딥러닝 (deep learning)은 인간 두뇌의 신경망에서 영감을 받은 신경 네트워크를 기반으로 한 기계학습 (machine learning)의 분야이다. 딥러닝은 얇은 층의 뉴런 모델들의 연결로 구성되어 있는 기존의 인공신경망 (Artificial Neural Network, ANN)에서 보다 깊은 뉴런 모델의 층을 쌓아 올림으로써 신경망의 학습 능력을 높이게 된다. 또한, 최근 대량의 데이터 수집 방법 및 기술과 그래픽스 처리장치 (Graphics Processing Units, GPU)를 이용한 풍부한 영상 지원의 자원으로 다양한 분야에 적용이 가능하게 되면서 사람의 능력을 뛰어넘는 성능을 보여주었다. 이러한 딥러닝의 급격한 발전은 현재 의료영상 분석에서도 활발히 적용되고 있다. 의료영상 분야에서는 딥러닝 기술을 통해 특정 장기를 분할 (segmentation)하거나, 병변의 위치를 찾고 (detection), 분류 (classification)를 통해 질병을 진단하는 목적으로 주로 활용되고 있다[2].

기계학습 및 딥러닝 기술은 데이터로부터 학습한다는 특징을 가지고 있다. 이에 데이터는 효과적인 딥러닝 알고리즘 개발에 있어

중요한 요소이다. 딥러닝 기술은 주어진 데이터로부터 패턴이나 특성을 전문적 지식이나 경험에 의거하여 수동 작업을 통해 추출해야되는 기존의 기계학습 방식과는 달리 데이터의 특징을 자동적으로 추출하여 학습할 수 있다는 장점이 있지만 앞서 언급한 바와 같이 인공지능의 층이 깊어지면서 학습 모델의 파라미터 (parameter) 수가 증가하고 뛰어난 모델 성능에 달성하기 위해서는 더 많은 양의 학습데이터와 이에 대한 정답 데이터, 그리고 컴퓨터 연산량이 요구된다. 하지만 의료 영상 데이터의 경우 환자의 개인 정보 보호와 관련된 사회적 우려를 유발하고 정답 데이터를 만들기 위해 의료진의 상당한 노력이 필요하기 때문에 대량의 데이터를 수집하는데 있어 시간과 비용, 그리고 제한된 접근성으로 인해 많은 제약을 받는다[3]. 특히, 1차 및 2차 병원의 경우 학습데이터가 부족한 경우가 많아 딥러닝을 이용하여 고성능의 모델을 개발하는데 어려움이 있다. 이는 자체적으로 딥러닝 학습 모델을 개발하여 연구를 진행하거나 상용화 하려는 관련 종사자들에게 큰 장애물이 될 수 있다. 따라서 소규모의 의료기관에서도 적은 양의 학습데이터로 우수한 성능의 딥러닝 모델을 개발 할 수 있도록 기회를 제공 할 필요가 있다.

소량의 학습데이터로 딥러닝 모델을 학습하는 경우 모델이 학습한 데이터에 대해서는 높은 성능을 보이지만, 새로 입력 받는 데이터에 대해서는 낮은 성능을 보이는 과적합 (overfitting) 현상이 일어날 수 있다. 또한, 데이터를 수집하는 과정에서 질환의 유병률 등에 따라 수집된 데이터의 클래스 간 균형이 맞지 않는 가능성도 존재한다. 데이터 클래스가 불균형한 경우 전체적인 데이터에 대해 제대로 학습을 하지 못하고 큰 비중을 차지하는 클래스에 편향되어 학습된 모델의 전체적인 성능은 높으나 정작 원하는 항목에 대해서는 분류를 해내지

못하는 현상이 발생하게 된다. 데이터 부족과 클래스 불균형 문제는 의료분야에서 가장 흔히 접하게 되는 한계점들이며 이를 극복하기 위한 연구들은 현재까지도 활발하게 진행되고 있다[4], [5].

과적합 문제를 방지하기 위해 현재까지 다양한 방법들이 제시되어 왔다. 전통적으로는 드롭 아웃 (dropout), 앙상블 (ensemble) 기법, label smoothing 와 같은 정규화 기법 (regularization technique)들을 통해 모델의 복잡성 (complexity)을 줄이는 모델 관점에서 접근하는 방법이 있고, 데이터의 양을 늘리는 데이터 관점에서의 접근 방법이 있다. 클래스 불균형 문제의 경우 under-sampling (다수 클래스 (majority class)) 데이터를 제거하는 방법 [6] 와 over-sampling (기존의 데이터에서 인위적으로 혹은 복제 방식으로 소수 클래스 (minority class)에 결합하는 방법) [7], [8], [9] 등의 re-sampling 방법들이 존재한다.

본 연구에서는 학습데이터가 충분하지 않은 상황에서도 효과적인 딥러닝 모델을 개발하기 위하여 데이터 관점에서 접근하였으며 가장 대중화된 데이터 증강 기법 (data augmentation)을 적용하여 최고의 진단 성능을 보이는 최적화된 모델을 찾는데 초점을 두었다.

## 제 2절 관련 연구의 동향

데이터 증강 기법은 데이터를 수집하기 어려운 상황에서 원본 데이터를 기반으로 새로운 데이터를 만드는 기법이다. 데이터 증강 기법은 크게 전통적 영상처리 기반 데이터 증강 기법과 생성적 적대 신경망 (Generative Adversarial Network, GAN) [10]을 이용한 합성

데이터 증강 기법으로 분류할 수 있다. 딥러닝을 활용한 의료 영상 분석에서 흔히 사용되는 전통적 영상처리 기반 데이터 증강 기법으로는 무작위로 뒤집기, 회전하기, 자르기 등 영상의 물리적 형태를 변환하거나, 밝기 조절, 역상, Contrast Limited Adaptive Histogram Equalization (CLAHE) 등 영상의 밝기를 조절하는 기법들이 있다[11]. 이러한 전통적인 방법들로 학습된 딥러닝 모델은 원본 데이터가 보유하고 있는 속성들의 다양한 변이를 배우지 못해 그 성능이 부분적으로만 향상된다는 한계점을 가지고 있다.

GAN을 이용한 합성 데이터 증강 기법은 데이터가 부족한 상황에도 딥러닝에 성공적으로 적용된 또 다른 대안 방법이다. GAN은 비지도 학습 (unsupervised learning) 방식의 인공지능 알고리즘 중 하나로 원본 학습 데이터의 특징을 스스로 학습하여 새로운 데이터를 생성해낸다. 원본 데이터가 가지는 속성들의 다양한 변이를 학습 할 수 있는 동시에 원본 데이터와 유사한 분포를 가지는 데이터를 생성할 수 있다는 점에서 데이터 증강 기법에 많이 적용되어 뛰어난 성능을 보여주었다[12], [13], [14], [15], [16], [17].

### 제 3절 관련 연구의 한계점

데이터 증강 기법을 적용할 때 학습 모델이 가장 우수한 성능에 도달하기 위해서 얼마만큼의 증강 데이터가 필요한지 알 수 없다. 또한, GAN을 이용하여 합성 데이터 증강을 수행한 관련 연구 사례들을 보면 저자들이 합성 데이터의 양과 증강에 사용한 데이터의 양에 대한 명확한 설명과 기준이 제시되어 있지 않다. Frid-Adar et al[18]은 데이터 증강 개수를 임의적으로 증가시켜 가며 모델 성능의 변화를 관찰하는 실험을 진행하였다. 그러나 해당 연구에서는 데이터 개수를 일정한 간격 없이 임의적으로 설정하여 증가시켰다는 점에서 정략적이지 않고, 연구자 개인이 임의로 성능이 가장 높을 것이라고 추측되는 시점에서 데이터 증강을 멈추기 때문에 해당 시점이 실제로 가장 우수한 시점이라고 결정 짓기 어렵다. 또한, 수동 작업이 개입되므로 적절한 증강 개수를 찾는 데 까지 오랜 시간이 소요된다는 한계점들이 존재한다.

### 제 4절 연구의 목적

본 연구에서는 GAN을 기반으로 한 합성 데이터 증강을 수행하였을 때 모델이 최고의 성능에 달성하기 위해 필요한 데이터 증강의 적절 양을 자동으로 탐색해줄 수 있는 방법을 제안하였다. 이는 가장 우수한 성능에 최적화된 모델을 자동적으로 찾아주고 본 방법을 통해 의료 영상 데이터가 부족한 상황에서도 높은 성능을 지닌 딥러닝 모델을 개발할 수 있도록 하는 것을 목적으로 하였다.

본 연구에서 사용된 실험 데이터는 2차 병원에서 수집한 부비동

x-ray 영상이며, 본 데이터를 통해 모델 학습 및 모델 성능을 평가하고, 모델의 일반성을 확보하기 위해 3차 병원에서 추가적으로 수집한 부비동 x-ray 영상을 이용하여 외부 검증 (external validation)을 진행하였다.

## 제 2 장 연구 방법

### 제 1 절 데이터 수집 및 구성

본 논문에서 사용된 의료영상 데이터들은 서울특별시 보라매 병원과 분당서울대병원 이비인후과 주도로 임상 시험 심사위원회 (Institutional Review Board, IRB)에서 승인된 연구 10-2020-173와 B-2010-645-401에서 촬영된 것이며, 촬영된 의료영상의 이용을 승인 받았다.

2차 의료기관인 서울특별시 보라매병원에서 2010년 1월부터 2020년 1월까지 상악동염 (maxillary sinusitis)을 진단할 때 가장 많이 활용되는 x-ray의 일반촬영법인 water's view와 한달 이내에 촬영한 CT를 통해 만성부비동염 (chronic sinusitis, CRS) 진단을 받은 250명의 환자 데이터를 수집하였다. 외부 검증 (external validation)을 위해 분당서울대병원에서 2003년 8월부터 2019년 4월까지 동일한 조건으로 105명의 환자에 대한 데이터를 수집하였다. 그 중 부비동 x-ray를 찍고 2주 이내에 CT를 촬영한 환자들만 실험용 데이터로 사용하였으며, 급성 부비동염 (acute sinusitis) 및 x-ray와 CT 촬영 기간 동안에 병변이 급격히 악화되거나 호전 된 환자들은 제외하였다. 그림 2.1은 본 연구의 연구 대상 선정 기준 및 제외 기준을 제시한다. 모든 데이터 세트는 이비인후과 임상치의 소견을 걸쳐 한쪽 상악동에만

염증이 있을 경우 “left” 혹은 “right sinusitis”, 양쪽에 염증이 있을 경우 “bilateral sinusitis”, 염증이 없을 경우 “clear”로 레이블링 (labeling)이 이루어졌다 (그림 2.2). 딥러닝 모델 학습을 위한 입력 데이터의 클래스는 두 개의 범주 (1 = “sinusitis” (부비동염), 0 = “clear sinus” (정상))로 설정하였다.

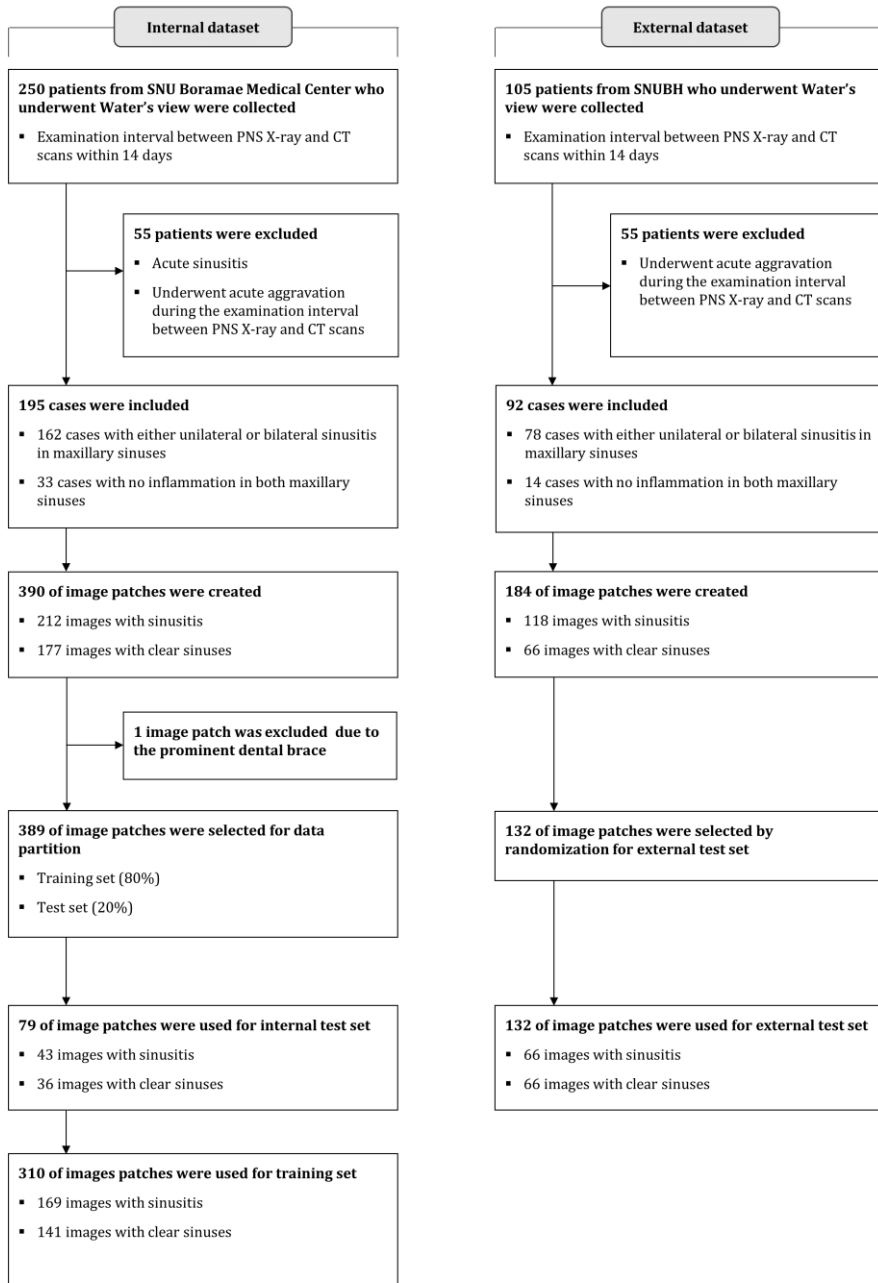


그림 2.1 연구 대상 선정 기준 및 제외 기준

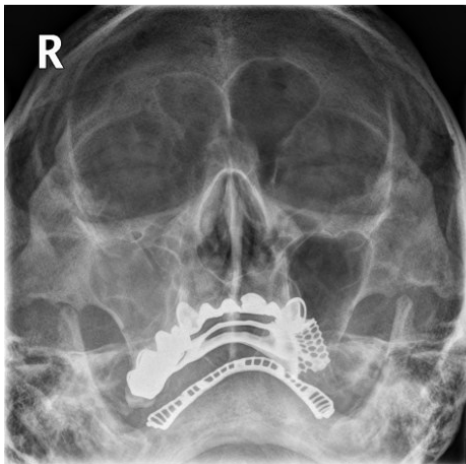




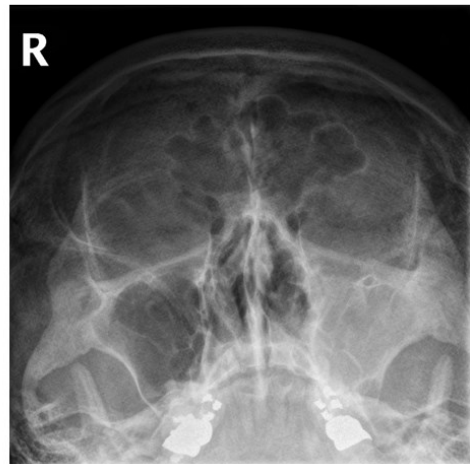
(a) Clear Sinus



(b) Bilateral Sinusitis



(c) Right: Sinusitis, Left: Clear



(d) Right: Clear, Left: Sinusitis

그림 2.2 부비동 x-ray 데이터 샘플 및 의사 소견에 따른 분류 기준

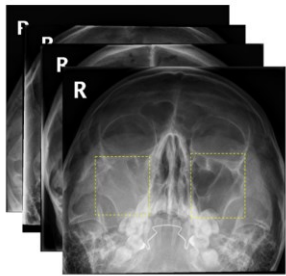
## 제 2 절 데이터 전처리 과정

데이터 전처리 과정은 이비인후과 임상의를 통해 상악동 (maxillary)의 해부학적 부위를 기준으로 관심영역 (region of interest, ROI)을 경계 박스 (Bounding box)로 표시하고, labelImg annotation 툴을 사용하여 각 ROI의 좌표에 대한 주석을 xml 파일로 저장하였다. 이후 MATLAB (ver. 2019a (The Math Works, Inc., 2019 Computer Software) 소프트웨어를 이용하여 ROI를 잘라내어 (cropping) 이미지 패치 (image patch) 를 생성하였다. 그 결과, 내부 데이터 세트는 총 389개의 이미지 패치가 생성되었고 그 중, 부비동염이 포함된 이미지 패치 개수는 212개, 정상은 177개 였다. 외부 검증용 데이터 세트에서는 총 187개의 이미지 패치 중에서 118개는 부비동염, 66개는 정상이었다. 내부 데이터 세트와 외부 검증용 데이터 세트의 클래스 비율은 각각 1:1.2, 1:1.8로 나타났다.

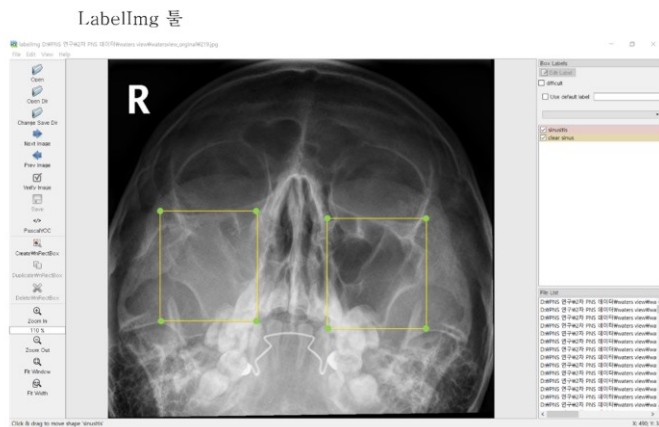
딥러닝 알고리즘 개발을 위해 전체 데이터 개수에서 80%는 학습데이터 (training set), 20%는 평가 데이터로 random stratified sampling을 이용하여 분할하였다. Random stratified sampling 방법은 학습 데이터를 나눌 때 무작위로 샘플링을 하되, 기존의 데이터 세트의 클래스 비율이 학습 데이터 세트와 검증 데이터 세트에서도 동일하게 유지시켜준다. 그림 2.3와 표 2.1은 이미지 패치 생성 과정과 딥러닝 모델 개발을 위한 데이터 세트의 구성을 나타낸다.

표 2.1 부비동 x-ray 데이터의 구성

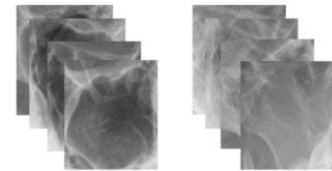
Numbers of Image Patches			
Data/Class	Clear Sinus	Sinusitis	Total
Training Set (80%)	141	169	310
Test Set (20%)	36	43	79
Total	177	212	389



데이터 파일 업로드



저장 시 xml  
파일이 생성됨



자르기 (cropping)



```

<object>
  <name>sinusitis</name>
  <pose>Unspecified</pose>
  <truncated>0</truncated>
  <difficult>0</difficult>
  <bndbox>
    <xmin>135</xmin>
    <ymin>384</ymin>
    <xmax>382</xmax>
    <ymax>662</ymax>
  </bndbox>
</object>
<object>
  <name>clear sinus</name>
  <pose>Unspecified</pose>
  <truncated>0</truncated>
  <difficult>0</difficult>
  <bndbox>
    <xmin>557</xmin>
    <ymin>401</ymin>
    <xmax>808</xmax>
    <ymax>677</ymax>
  </bndbox>
  
```

그림 2.3 이미지 패치 생성 과정

### 제 3 절 ACGAN 모델을 이용한 데이터 생성

GAN은 2014년 Goodfellow에 의해 개발된 생성모델 (generative model)을 기반으로 한 기법으로 최근 의료 분야에서 데이터 증강 기법으로 널리 사용되어 왔다. GAN은 데이터를 생성해주는 생성기 (generator)와 생성된 합성 데이터와 실제 원본 데이터를 판별하는 판별기 (discriminator)로 구성되어 있다. 이 두 네트워크는 서로 대립 (adversarial)하며 서로의 성능을 점차 개선해 나간다는 것이 이 모델의 주요 개념이다. 생성기 ( $G$ )는 학습하고자 하는 원본 데이터  $x$ 의 분포 (distribution)를 알아내 데이터 분포 (data distribution)를 최대한 정확히 모사하여 생성된 샘플을 실제 데이터와 구별하기 어렵게 만드는 것을 목적으로 한다. 한편 판별기 ( $D$ )에서는 생성기에서 출력된 합성 데이터가 실제 원본 데이터에서 입력 받은 데이터인지 혹은 생성기에서 만들어진 데이터인지를 구별하여 각각의 경우에 대한 확률을 예측한 뒤, 생성기를 업데이트한다. 생성기 ( $G$ )와 판별기 ( $D$ )의 대립 관계는 two-players minimax game의 개념을 따르며 수식 2.1와 같다.

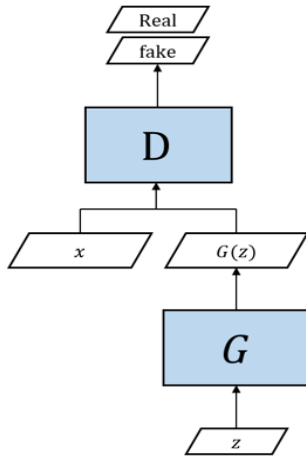
$$\min_G \max_D V(D, G) = E_{x \sim P_{data}(x)} [\log D(x)] + E_{z \sim P_z(z)} [\log(1 - D(G(z)))] \quad 2.1$$

$P_{data}$ 와  $P_z(z)$ 는 원본 데이터와 합성 데이터에 대한 각각의 확률 분포 (probability distribution)를 의미한다.  $G(z)$ 는 생성기가 랜덤 벡터 노이즈 (random vector noise)  $z$ 를 입력으로 받아 원본 데이터와 유사한 분포를 가진 샘플을 만들어낸 것을 의미한다. 생성기는 원본

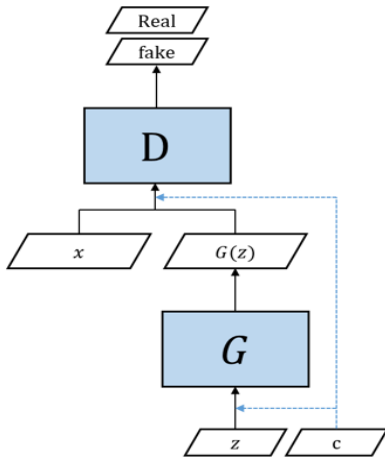
데이터와 구별할 수 없을 정도의 가짜 이미지를 생성하는 능력을 향상시키기 위해  $D(G(z))$  을 1이 되도록 학습하고 식별기는 원본 데이터와 실제 데이터를 구별해 낼 수 있는 능력을 향상시키기 위해  $D(G(z)) = 0$  이 되도록 학습한다[10].

본 연구에서 GAN 기반의 데이터 증강을 수행하기 위해 GAN의 변형된 모델 중 하나인 Auxiliary Classifier GAN (ACGAN) [19]을 사용하였다. Conditional GAN (CGAN)은 외부의 추가 정보 (class label 혹은 modality의 데이터 등)를 생성기와 식별기에 입력하여 생성된 이미지의 품질 (예: 해상도)을 개선시키기 위해 도입된 변형 모델이다[20]. ACGAN은 CGAN 원리와 유사하나 식별기가 데이터의 클래스를 예측할 수 있다는 점에서 차이가 있다 (그림 2.4).

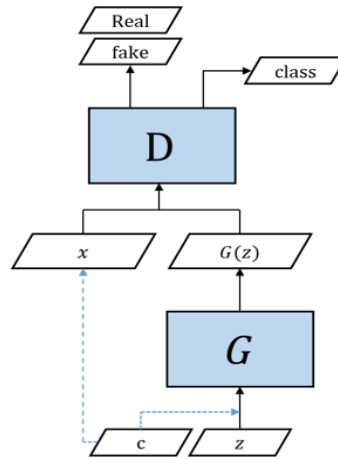
ACGAN 학습을 위해 사용된 생성기와 식별기의 아키텍처 (architecture)는 그림 2.5 에 구체적으로 나타나 있다. ACGAN 학습을 위해 사용된 하이퍼파라미터 (hyper-parameter)는 다음과 같다. 최적화 함수 (optimizer function)는 Adam, 학습율 (learning rate)는 0.0002에 Beta 0.5, 배치 크기 (batch size)는 64, 그리고 30000 에폭 (epoch)으로 설정하여 학습하였다. ACGAN의 학습은 NVIDIA GeForce GTX 1050 Ti GPU 환경에서 수행되었고 학습시간은 약 5시간 정도 소요되었다. ACGAN에서 생성된 상악동 이미지 패치와 원본 이미지는 그림 2.6 에서 볼 수 있다.



(a) GAN



(b) CGAN



(c) ACGAN

그림 2.4 GAN, CGAN, ACGAN 비교 ( $z$ : 노이즈 벡터,  $c$ : 클래스 라벨,  $x$ : 원본 이미지,  $G$ : 생성기 (generation),  $D$ : 식별기 (discriminator))

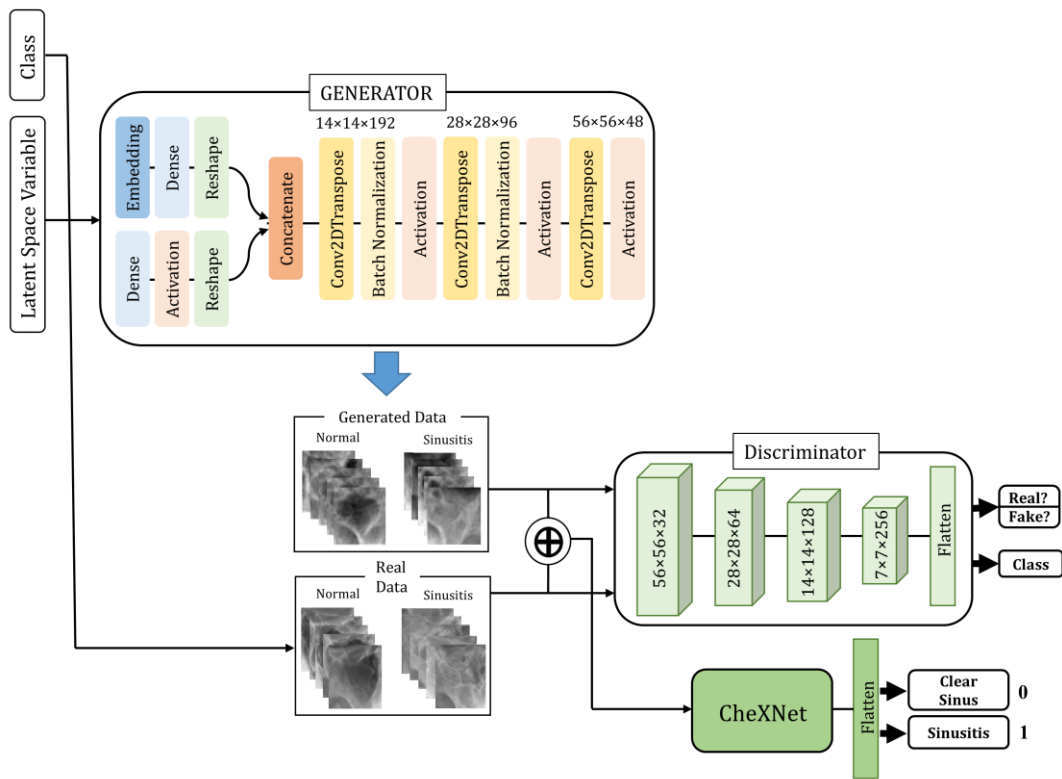


그림 2.5 AC-GAN 아키텍처와 부비동 x-ray 분류 방법



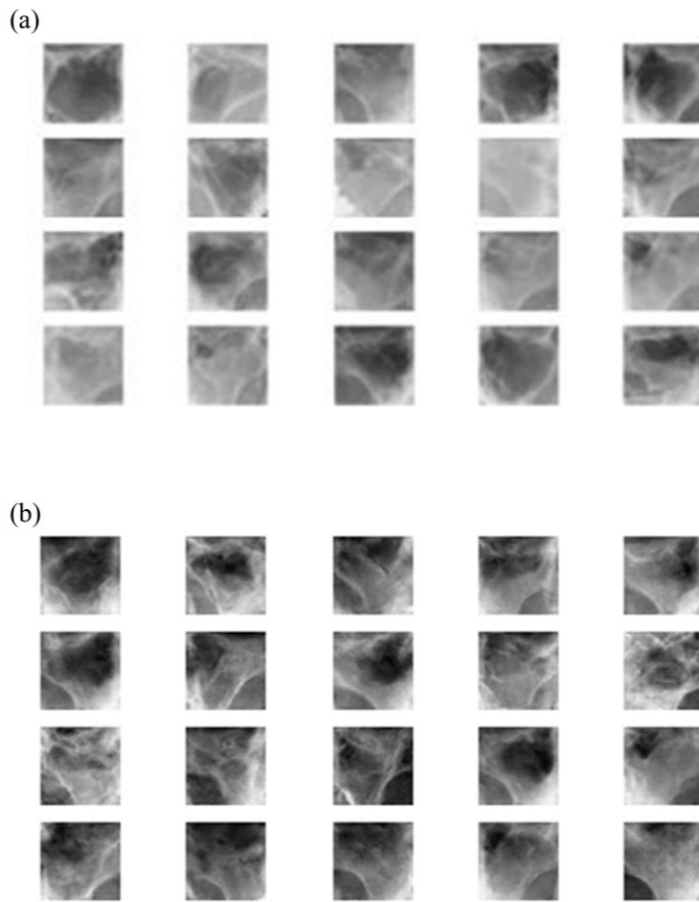


그림 2.6 ACGAN 모델을 통해 생성된 샘플

## 제 4절 데이터 증강 기법

본 연구에서는 GAN을 이용한 합성 데이터 증강 기법을 통해 학습 데이터의 양을 증대시키고 자동화 알고리즘을 통해 최고의 성능을 보이는 최적화된 모델을 자동적으로 찾아주는 과정을 수행하였다. 더불어 전통적 영상 처리 기반의 데이터 증강도 별도로 수행하여 두 기법 간의 성능을 비교하였다.

현재 학습 데이터의 클래스 비율은 1:1.2 (“sinusitis” 이미지 패치 개수: 169, “clear sinus” 이미지 패치 개수: 141)이다. 클래스간의 불균형이 큰 편은 아니지만 데이터 증강 기법의 효과를 명확하게 측정하기 위해 데이터 증강 기법을 적용하기 전에 ACGAN에서 생성된 합성 데이터를 소수 클래스 (minority class)에 결합시켜 클래스 균형을 맞추었다.

### 4.1 전통적 데이터 증강 기법

의료 영상 분석을 위해 흔히 사용되는 가장 일반적인 전통적 데이터 증강 기법은 기하학적 변환 (geometric transformation) 및 영상의 강도 변환 (intensity transformation) 등이 있다. 그림 2.7와 같이 본 연구에서는 부비동 x-ray 이미지 패치가 보유하고 있는 영상 특성을 최대한 보존하기 위해 기하학적 변환 {좌우 뒤집기 (horizontal flipping)과 회전}만 적용하였다. 회전의 경우  $-10$  에서  $10$ 도 범위 내에서 무작위로 회전 (random rotation) 하도록 설정하였으며 두 종류 모두 원본 데이터에만 적용하였다.

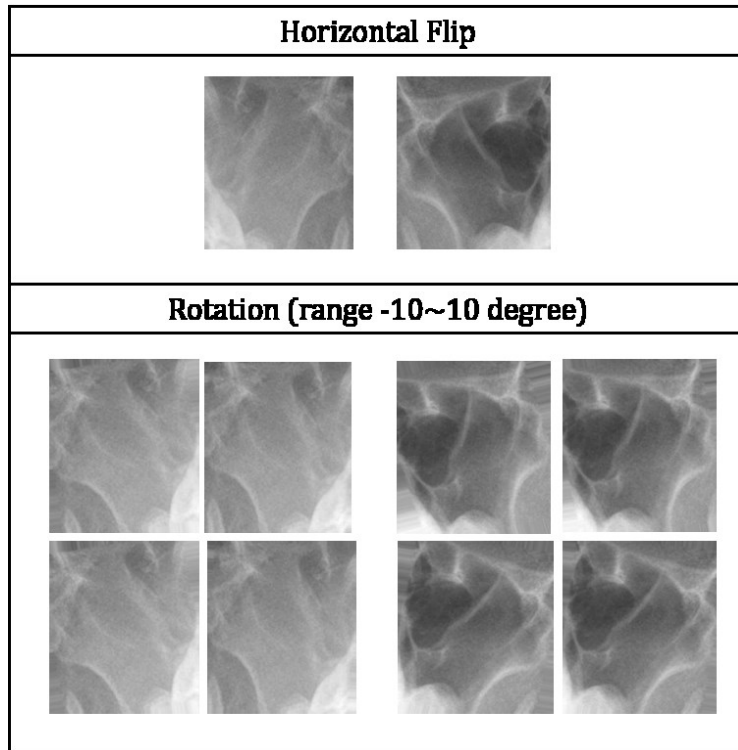


그림 2.7 전통적 데이터 증강 기법을 적용한 결과 샘플: (좌우 뒤집기 및 회전)

#### 4.2 GAN 기반 합성 데이터 증강 기법

GAN을 이용하여 합성 데이터 증강을 수행하기 위해 기존의 학습 데이터 세트를 이용하여 학습한 ACGAN 생성기의 가중치 (weight)를 업로드하여 데이터를 생성하였다. 그리고 합성 데이터를 기존의 학습 데이터와 결합시킨 후 딥러닝을 진행하였다.

본 연구에서는 슬라이딩 윈도우 (sliding window) 개념을 적용하여 모델 성능이 가장 높은 시점에 최적화된 모델과 그에 해당하는 데이터 증강의 배율 (본 연구에서는 이를 최적의 배율 ( $K_{optimal}$ )로 정의)을 자동적으로 찾아주는 방식으로 접근하였다.

슬라이딩 윈도우는 리스트 형태의 배열이 주어졌을 때 지정된 윈도우 범위 내에 속해있는 모든 요소 값들을 특정 문제에 맞게 연속적으로 처리해주는 기능을 한다[21]. 본 연구에서 수행되는 슬라이딩 윈도우 과정은 아래와 같다.

- ① 배열의 첫 번째 인덱스 (index)부터 시작한다.
- ② 슬라이딩 윈도우 범위 내에 속한 성능 지표 값들을 차례대로 비교하며 최댓값을 지역 변수에 업데이트 한다.
- ③ 윈도우는 우측으로 한 칸씩 이동하며 ②번 과정을 반복한다. 이때 출력되는 최댓값이 이전에 위치한 윈도우의 최댓값보다 같거나 증가하였는지 판단한다.
- ④ 만일 같거나 증가하였을 경우 배수는 1씩 증가하며 다음 데이터 증강 단계로 넘어가고 감소하였을 경우 모든 프로세스는 멈추게 되며, 감소하기 바로 전 시점의 성능 지표에 해당되는 배율 값과 학습 모델이 최종적으로 출력된다.

그림 2.8은 위 과정을 구체화한 순서도와 전체적인 시스템의 흐름을 나타내는 순서도를 보여준다. GAN을 이용한 합성 데이터 증강은 ACGAN 모델의 생성기에서 기존 학습 데이터 개수의 배수 1만큼 합성 데이터를 생성하여 기존의 학습 데이터와 결합하게 되면서 시작된다. 이어서 딥러닝 모델 학습 및 모델 평가 과정이 끝날 때 마다 특정 성능 지표를 리스트 형태의 배열로 저장한다 (*AUC\_array*). 본 연구에서는 AUC 값을 기준으로 성능이 가장 우수한 시점을 관찰하였다. *AUC\_array* 배열의 길이가 윈도우 크기

( $w$ )보다 크면 슬라이딩 윈도우 과정으로 넘어가게 되고 작거나 같을 경우 배수는 1씩 증가하며 다음 데이터 증강 단계로 넘어간다. 슬라이딩 윈도우는  $AUC\_array$ 의  $[n-w]$ 번째 인덱스부터 시작되며  $w$  범위 내에서 최댓값을 찾아 지역 변수인  $window\_AUC\_max$ 에 업데이트 된다. 매 반복 마다 새로 업데이트 되는  $window\_AUC\_max$ 가 이전  $window\_AUC\_max$  값 보다 증가하거나 같으면 그 값은  $AUC\_max$ 에 업데이트 되면서 배수는 1씩 증가하며 ( $K = K + 1$ ) 다음 데이터 증강을 수행한다. 각 윈도우에서 출력한 최댓값들에서 감소하는 시점이 보이면 모든 과정은 멈추게 되고 최적의 배율 ( $K_{optimal}$ )과 그 배율에 해당되는 학습 모델 ( $M_{optimal}$ )이 출력된다.

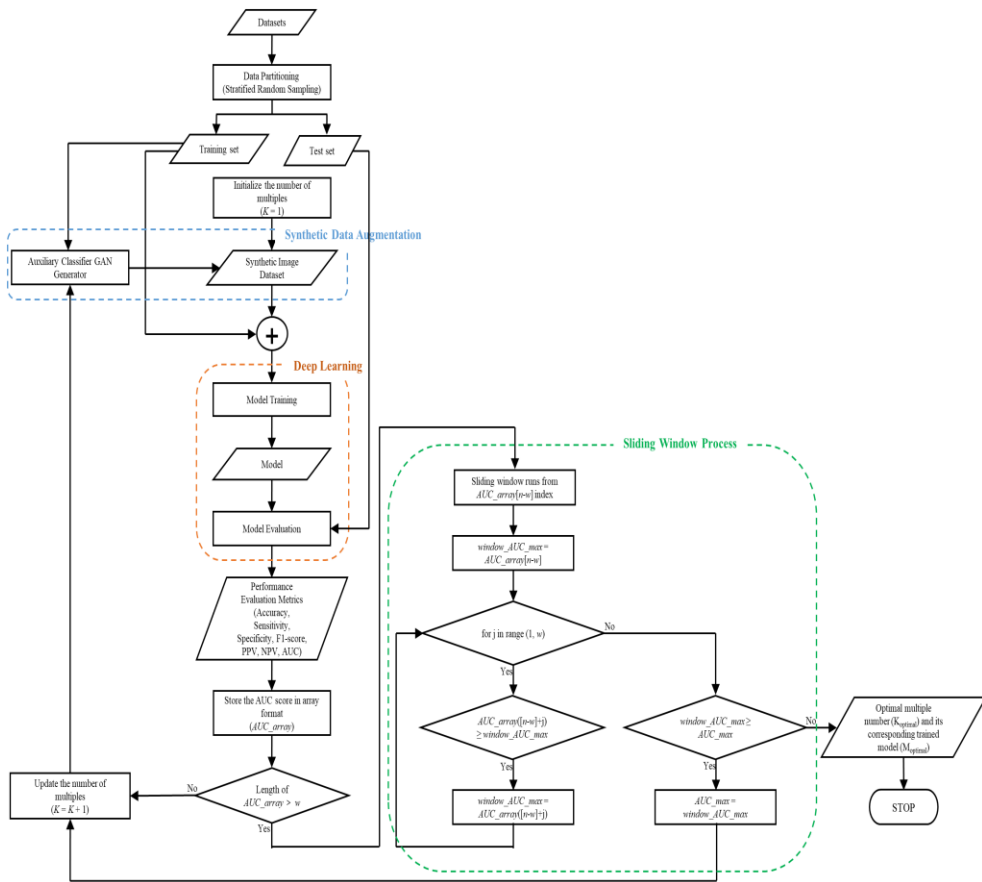


그림 2.8 슬라이딩 윈도우를 이용한 GAN 기반의 데이터 증강 기법 순서도

## 제 5절 분류 모델 학습 과정

본 연구에서는 부비동염 진단을 위해 사전학습 (pre-trained)된 CheXNet 모델 [22]을 이용하여 전이 학습 (transfer learning) 하였다. CheXNet은 DenseNet 모델 [23]을 기반으로 하였으며, 모델 학습을 위해 사용된 흉부 x-ray와 본 연구에서 사용되는 부비동 x-ray의 영상 특성이 유사하다는 점을 고려하여 본 모델을 선택하였다. 모델 학습 과정에서 입력 데이터로 받은 상악동의 이미지 패치는 이진 보간법 (bilinear interpolation)을 이용하여  $56 \times 56$  픽셀 크기로 조정되었다. 전이 학습 과정에서 전결합 계층 (fully-connected layer)의 마지막 레이어는 sinusitis와 clear sinus 두 개의 범주로 출력 가능하도록 수정하였다.

모델의 하이퍼파라미터는 학습율 (learning rate)은 0.0001, 배치 크기 (batch size)는 8, 에폭 (epochs)은 25, 손실 함수 (loss function)는 binary cross entropy, 그리고 최적화 함수 (optimizer function)는 Adam으로 설정하였으며, 본 하이퍼파라미터들은 fine-tuning 과정에서 가장 높은 성능을 보여 지정되었다. CheXNet 모델 학습은 NVIDIA Tesla K80 GPU가 장착된 구글 클라우드 플랫폼 (Google Cloud Platform) 서버에서 수행되었다.

## 제 6절 모델 성능 평가 및 유의성 검증 방법

모델 성능을 검증하기 위해 혼동 행렬 (confusion matrix)을 이용하여 정확도 (accuracy), 민감도 (sensitivity), 특이도 (specificity), F1-score, 양성 예측 값(positive predictive value, PPV) 및 음성 예측 값(negative predictive value, NPV)을 계산하였다 (그림 2.9). 수식 2.2 ~ 2.7 은 각 성능 지표에 대한 수식을 나타낸다. 또한, 곡선 아래의 면적(area under the curve, AUC)과 수신기 작동 특성(receiver operating characteristic, ROC) 곡선도 계산되었다. 모델 평가는 고정된 내부 검증 데이터 세트와 외부 검증 데이터 세트를 통해 이루어졌으며 모든 성능은 default cutoff 값인 0.5를 기준으로 계산되었다.

추가로 GAN을 기반으로 한 클래스 균형 및 전통적 데이터 증강을 수행한 CheXNet 모델과 GAN 기반의 클래스 균형 및 데이터 증강을 수행한 모델 간의 통계적 유의성을 검증하여 진단 성능을 비교하였다. 정확도의 유의성을 검증하기 위해 bonferroni correction을 통한 McNemar's test가 사용되었다. P-value가 0.05보다 작으면 통계적으로 유의함을 간주하였다. 통계 분석은 R for Windows 4.0.3 (R Foundation for Statistical Computing, Vienna, Austria)과 IBM SPSS Statistics ver. 26.0 (IBM Corporation, Armonk, NY, 미국)을 사용하여 수행되었다.



		Actual Value	
		Actual Value Positive	Actual Value Negative
Test Outcome	Test Outcome Positive	TP (True Positive)	FP (False Positive)
	Test Outcome Negative	FN (False Negative)	TN (True Negative)

True Positive (TP): Sinusitis를 Sinusitis라고 진단함.

True Negative (TN): Clear Sinus를 Clear Sinus라고 진단함.

False Positive (FP): Sinusitis를 Clear Sinus라고 진단함.

False Negative (FN): Clear Sinus를 Sinusitis라고 진단함.

그림 2.9 혼동 행렬

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad 2.2$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad 2.3$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad 2.4$$

$$\text{F1-score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad 2.5$$

$$\text{PPV} = \frac{TP}{TP + FP} \quad 2.6$$

$$\text{NPV} = \frac{TN}{FN + TN} \quad 2.7$$

### 제 3 장 연구 결과

본 연구에서는 내부 검증 데이터 세트와 외부 검증 데이터 세트를 통해 네 가지의 다른 실험 조건에서 분류 모델의 성능을 분석하였다. 네 가지 실험은 아래와 같다.

- ① 기존 학습 데이터로만 학습시켜 CheXNet 모델의 성능을 평가 (OD)
- ② GAN 기반의 클래스 균형을 한 데이터로 학습시켜 CheXNet 모델의 성능을 평가 (GCB)
- ③ GAN 기반의 클래스 균형과 전통적 데이터 증강을 적용 후 학습시켜 CheXNet 모델의 성능을 평가 (GCB+CDA)
- ④ GAN 기반의 클래스 균형 및 데이터 증강을 적용하여 학습시킨 CheXNet 모델의 성능을 평가 (GCB+OMGDA)

④의 경우 본 연구에서 제안한 GAN 기반의 데이터 증강 기법으로 최적의 배율일 때의 모델로 학습한 것이다. 모든 실험에 대한 성능 평가는 내부 검증 데이터 세트(부비동염 = 43, 정상 = 36)와 외부 검증용 데이터 세트(부비동염 = 66, 정상 = 66)로 평가되었다.

## 제 1 절 GAN 기반의 데이터 증강 기법의 최적 배율

슬라이딩 윈도우를 통해 GAN 기반의 데이터 증강 기법의 최적의 배율 ( $K_{\text{optimal}}$ )을 자동적으로 탐색해본 결과, 가장 높은 진단 성능은  $K_{\text{optimal}}$ 이 17일 때 이었다. 슬라이딩 윈도우의 최댓값이 감소하는 시점은 모델을 30번 학습 했을 때 나타났다. 즉, 증강 데이터 샘플 개수를 29 ( $n = 10,112$ )배만큼 증폭시켰을 때를 의미한다. 표 3.1, 그림 3.1과 3.2을 통해 GAN 기반의 데이터 증강 샘플 개수가 분류 모델의 성능에 미치는 영향과 각 배율에 대한 CheXNet 모델의 ROC 커브를 확인할 수 있으며, 그림 3.3 슬라이딩 윈도우가 적용되는 과정을 시각화한 것이다. 정확도는 점차 증가하면서 5,718개의 학습 데이터( $K_{\text{optimal}} = 17$ )에서 최고 성능이 관찰 되었음을 확인할 수 있다 (Accuracy = 0.949, Sensitivity = 0.977, Specificity = 0.917, F1-score = 0.953, PPV = 0.93, NPV = 0.71, AUC = 0.964). 이 후 정확도는 0.89 주변에서 계속 진동 하였고 더 이상 향상되지 않았다.

표 3.1 내부 테스트 세트를 이용하여 학습 데이터 개수의 증가에 따른 CheXNet 모델의 성능 평가

Number of Training sets (original images +synthetic images)	Accuracy	Sensitivity	Specificity	F1-score	PPV	NPV	AUC
(×2) N= 648	0.772	0.698	0.861	0.769	0.857	0.705	0.882
(×3) N= 989	0.810	0.767	0.861	0.815	0.868	0.756	0.888
(×4) N= 1,324	0.810	0.767	0.861	0.815	0.868	0.756	0.902
(×5) N= 1,662	0.848	0.814	0.889	0.854	0.897	0.800	0.891
(×6) N= 2,000	0.848	0.814	0.889	0.854	0.897	0.800	0.891
(×7) N= 2,338	0.848	0.767	0.944	0.846	0.943	0.777	0.926
(×8) N= 2,676	0.848	0.814	0.889	0.854	0.897	0.800	0.943
(×9) N = 3,014	0.848	0.837	0.861	0.857	0.878	0.816	0.943
(×10) N = 3,352	0.886	0.837	0.944	0.889	0.947	0.829	0.935

(×11) N= 3,690	0.861	0.861	0.861	0.871	0.881	0.838	0.926
(×12) N= 4,028	0.861	0.814	0.917	0.864	0.921	0.805	0.935
(×13) N= 4,366	0.873	0.884	0.861	0.884	0.884	0.861	0.949
(×14) N= 4,704	0.873	0.907	0.833	0.886	0.867	0.882	0.952
(×15) N= 5,042	0.873	0.861	0.889	0.881	0.902	0.842	0.940
(×16) N= 5,380	0.899	0.930	0.861	0.909	0.889	0.912	0.944
(×17) N= 5,718	0.949	0.977	0.917	0.955	0.933	0.971	0.964
(×18) N= 6,056	0.911	0.954	0.861	0.921	0.891	0.939	0.952
(×19) N= 6,394	0.873	0.861	0.889	0.881	0.902	0.842	0.948
(×20) N= 6,732	0.886	0.884	0.889	0.894	0.905	0.865	0.943
(×21) N= 7,070	0.873	0.814	0.944	0.875	0.946	0.810	0.943
(×22) N= 7,408	0.873	0.837	0.917	0.878	0.923	0.825	0.943
(×23) N= 7,746	0.899	0.884	0.917	0.905	0.927	0.868	0.956

(×24) N= 8,084	0.899	0.884	0.917	0.905	0.927	0.868	0.956
(×25) N= 8,422	0.861	0.861	0.861	0.871	0.881	0.838	0.933
(×26) N= 8,760	0.911	0.907	0.917	0.918	0.929	0.892	0.946
(×27) N= 9,098	0.911	0.954	0.861	0.921	0.891	0.939	0.952
(×28) N= 9,436	0.911	0.907	0.917	0.918	0.929	0.892	0.951
(×29) N= 9,774	0.873	0.930	0.806	0.889	0.851	0.906	0.944
(×30) N= 10,112	0.873	0.930	0.806	0.889	0.851	0.906	0.944

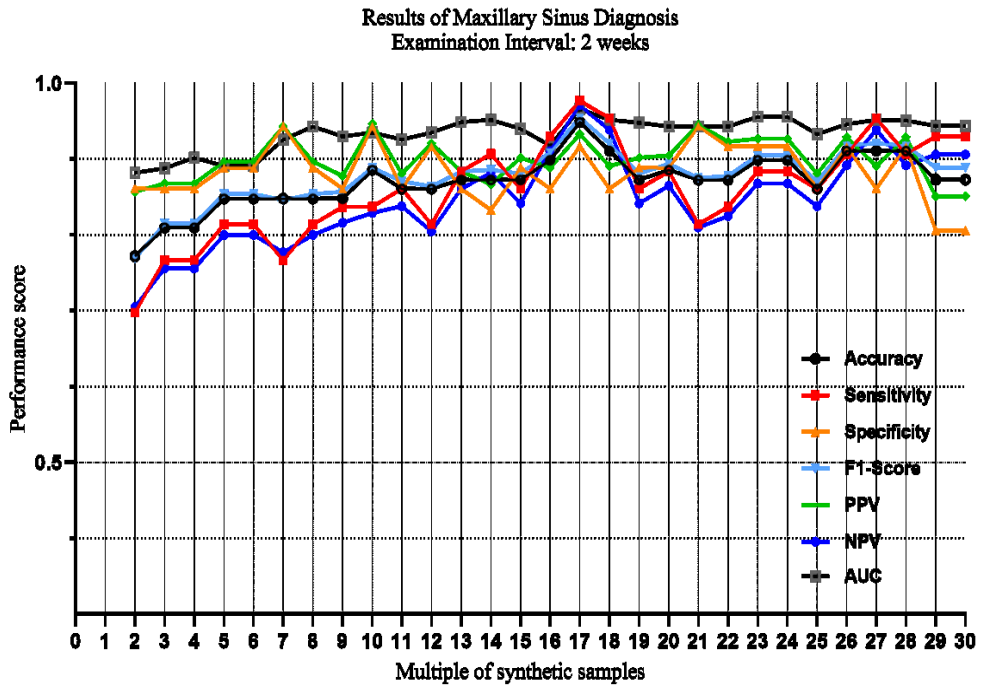
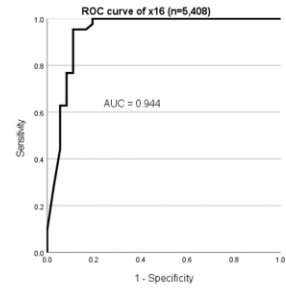
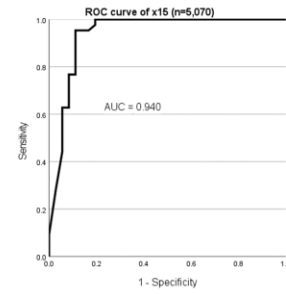
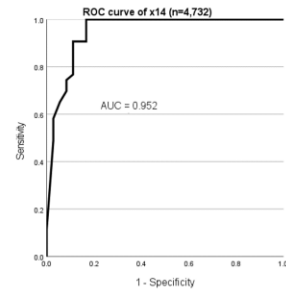
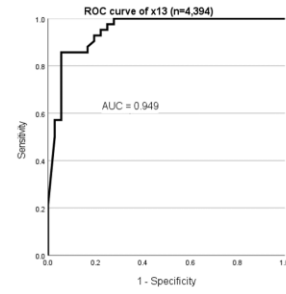
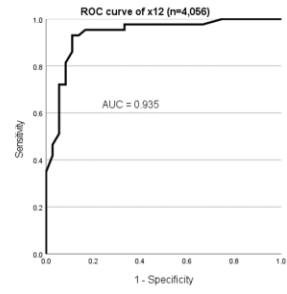
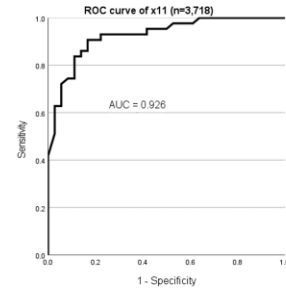
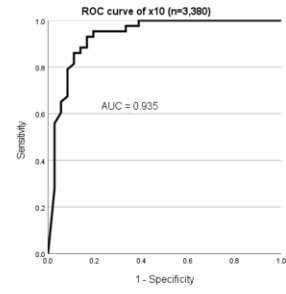
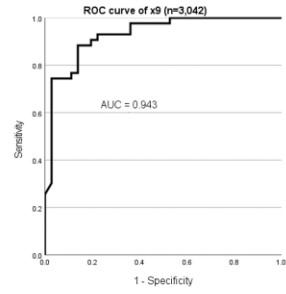
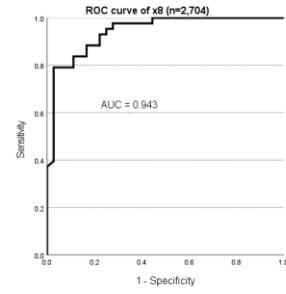
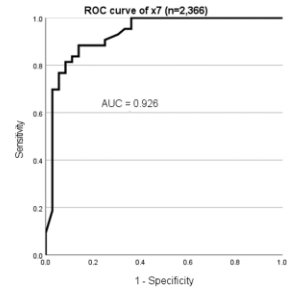
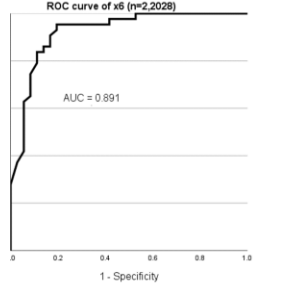
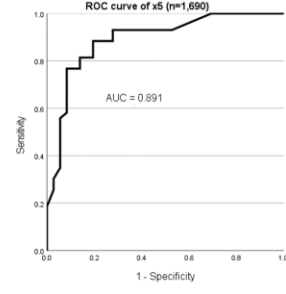
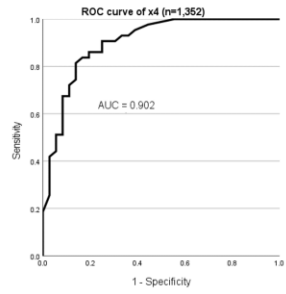
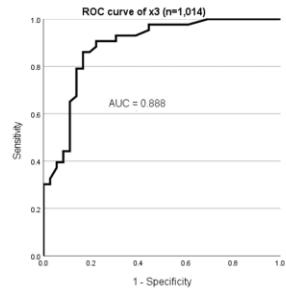
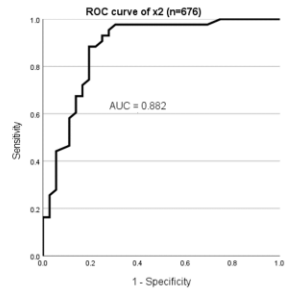


그림 3.1 표 3.1에 대한 산점도 (scatter plot)





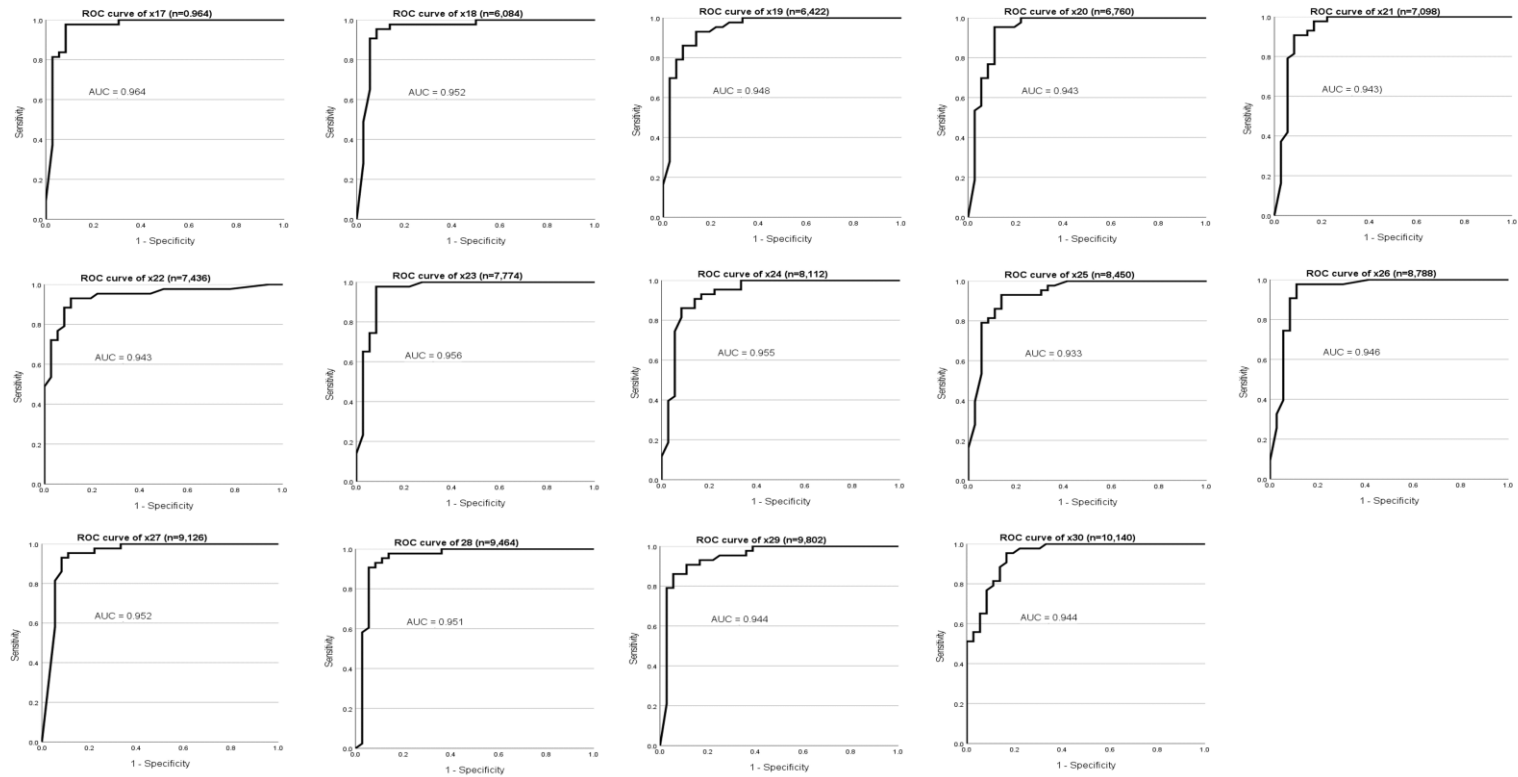


그림 3.2 각 배율에 대한 CheXNet 모델의 ROC 곡선

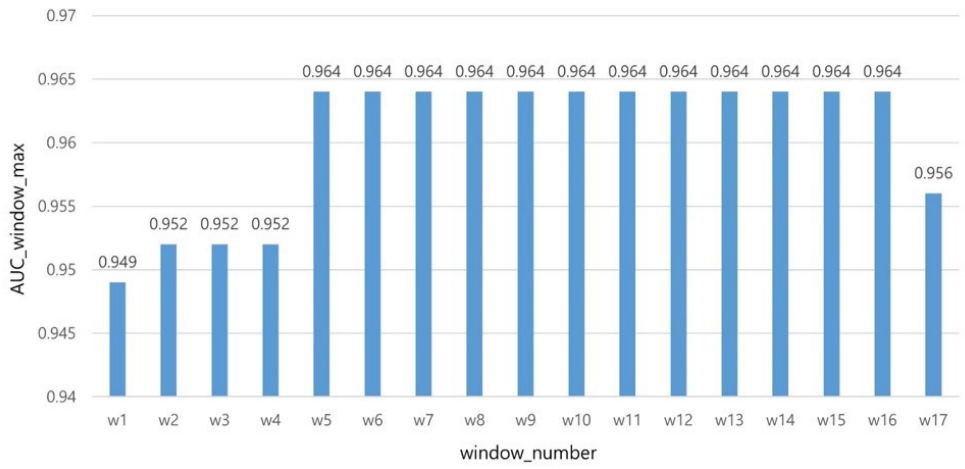
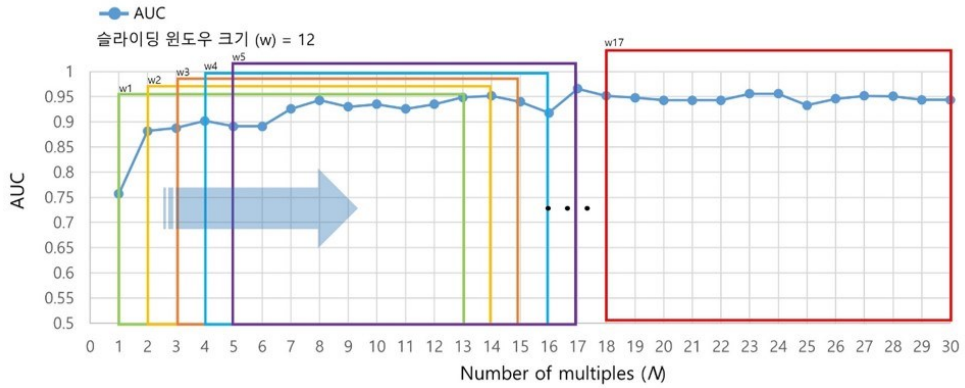


그림 3.3 슬라이딩 윈도우의 시각화

## 제 2 절 모델 성능 평가

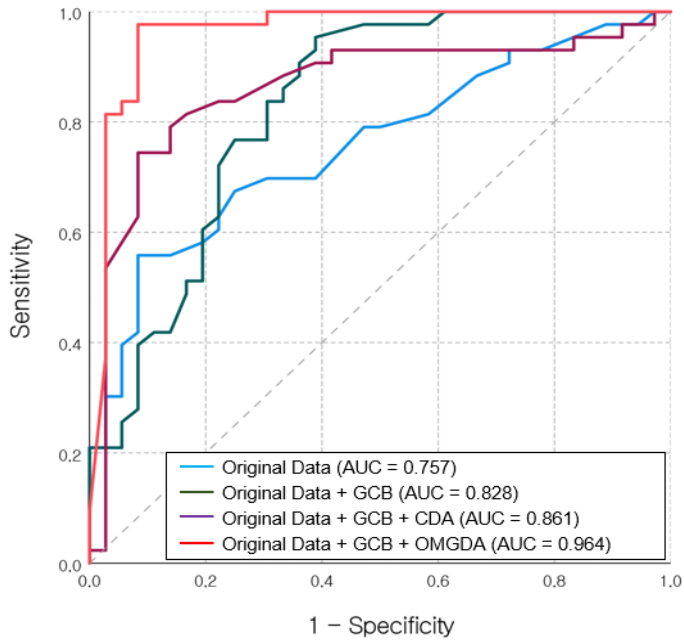
상약동염 진단을 위해 사용된 CheXNet 모델은 4 가지의 조건 (OD, GCB, GCB+CDA, GCB+OMGDA) 아래 모델 학습을 수행하였고 내부 및 외부 검증을 수행하였다 (표 3.2). 내부 검증 결과, 본 연구에서 제안한 GAN 기반의 합성 데이터 증강 기법의 최적의 배율로 학습한 모델 (GCB+OMGDA)에서 가장 높은 성능을 보였다 (accuracy: 0.949, sensitivity: 0.977, specificity: 0.917, F1-score: 0.955, PPV: 0.933, NPV: 0.971, AUC: 0.964). GCB+CDA 모델은 OD 모델과 GCB 모델 보다 높은 성능을 보였지만 (accuracy: 0.823, sensitivity: 0.744, specificity: 0.917, F1-score: 0.821, PPV: 0.914, NPV: 0.750, AUC: 0.861), GCB+OMGDA 모델보다는 낮았다. 외부 검증에서도 GCB+OMGDA 모델은 다른 모델들에 비해 가장 높은 성능을 보였다 (accuracy: 0.917, sensitivity: 0.924, specificity: 0.909, F1-score: 0.917, PPV: 0.910, NPV: 0.923, AUC: 0.955). 그림 3.4은 4개의 모델의 내부 및 외부 검증에 대한 ROC 커브를 나타낸다.

통계적 유의성 검증 결과 GCB+OMGDA 모델은 내부 검증 및 외부 검증에서 OD, GCB, GCB+CDA 모델들과 정확도를 비교하였을 때 통계적으로 유의함을 볼 수 있었다. 내부 검증의 경우 OD 모델과 비교하였을 때는  $P=.002$ , GCB 모델과 비교하였을 때는  $P=.001$ , 그리고 GCB+CDA 모델과의 비교에서는  $P=.045$ 이 나왔고, 외부 검증에서는 각 모델에 대하여  $P=.001$ ,  $P=.006$ ,  $P=.002$ 로 나왔다. 반면 GCB+CDA 모델은 OD 모델과 GCB 모델에 비해 통계적으로 차이가 나타나지 않았다. 내부 검증의 경우 GCB+CDA 모델과 OD

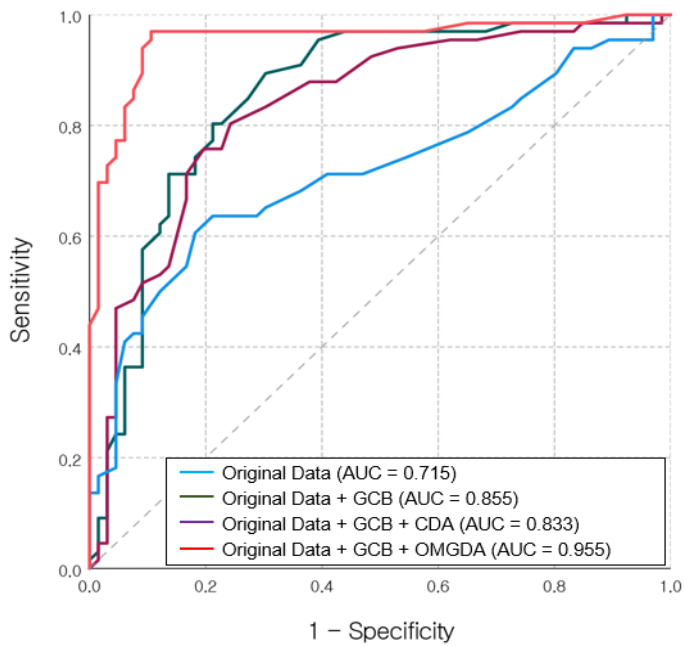
모델을 비교하였을 때는  $P=.537$ 이 나왔으며 GCB+CDA 모델과 GCB 모델은  $P=1.000$ 이 나왔다. 외부 검증에서는 각 모델에 대하여  $P=.112$ 와  $P=1.000$ 으로 나타났다 (그림 3.5). 통계적 유의성 검증 결과를 통해 본 연구에서 사용한 ACGAN을 통해 생성된 이미지는 데이터 증강 기법에 성공적으로 사용될 수 있음을 검증할 수 있었다.

표 3.2 CheXNet 모델의 내부 및 외부 검증 결과

Method	Accuracy	Sensitivity	Specificity	F1-score	PPV	NPV	AUC
<b>Internal test dataset (n = 79)</b>							
OD (original data)	0.734	0.628	0.861	0.720	0.844	0.660	0.757
GCB	0.747	0.721	0.778	0.756	0.795	0.700	0.828
GCB + CDA	0.823	0.744	0.917	0.821	0.914	0.750	0.861
GCB + OMGDA	0.949	0.977	0.917	0.955	0.933	0.971	0.964
<b>External test dataset (n = 132)</b>							
OD (original data)	0.689	0.591	0.788	0.656	0.736	0.658	0.715
GCB	0.788	0.727	0.849	0.774	0.828	0.757	0.855
GCB + CDA	0.780	0.712	0.849	0.764	0.825	0.747	0.833
GCB + OMGDA	0.917	0.924	0.909	0.917	0.910	0.923	0.955



(a) Internal Validation



(b) External Validation

그림 3.4 CheXNet 모델의 내부 및 외부 검증 결과에 대한 ROC 곡선

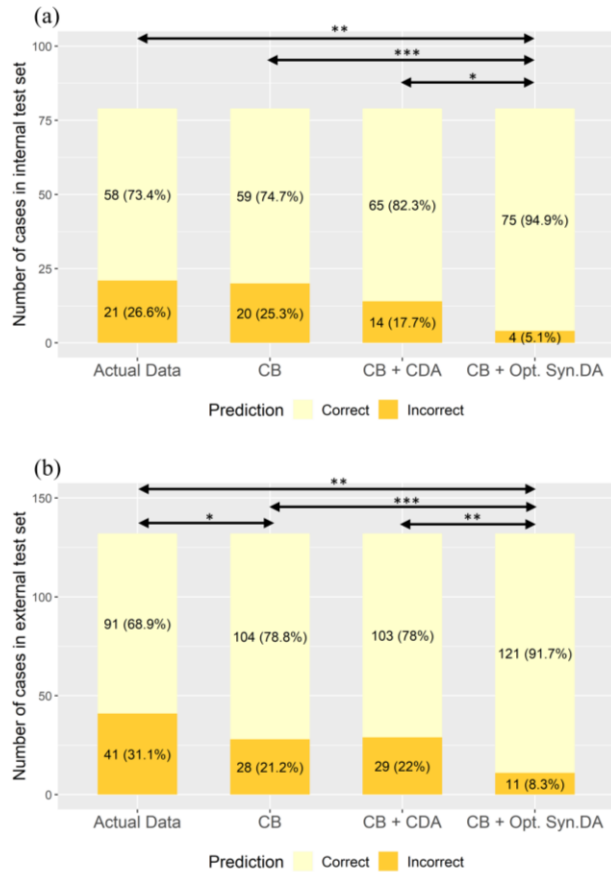


그림 3.5 통계적 유의성 검증 결과

## 제 4 장 고찰

본 연구에서는 부비동 x-ray 촬영과 CT 촬영 사이의 간격이 최대 14일인 195명의 환자를 내부 데이터 세트로 사용하였다. 3차 병원에서는 대부분의 환자들이 이미 1차 병원에서 만성 부비동염 진단을 받았기 때문에 부비동 x-ray 영상 대신 부비동 CT를 진단 도구로 사용하는 경우가 많았다. 반대로 진단이 불확실하거나 증상이 호전된 경우에는 x-ray 촬영을 실시하였다. 따라서 2주 간격 내에 x-ray와 CT를 모두 촬영한 환자는 거의 없었다. 195명의 환자 데이터로 딥 러닝을 수행하기에는 데이터가 많이 부족한 상황이었다.

이를 극복하기 위해, GAN을 이용하여 데이터를 생성하고 합성 데이터를 이용하여 데이터 증강을 수행하였다. 또한, 슬라이딩 윈도우 개념을 접목시켜 GAN을 기반으로 한 합성 데이터 증강 시, 가장 우수한 성능에 최적화된 모델과 최적의 배율을 자동적으로 찾아줄 수 있는 알고리즘을 구축하였다.

GAN 기반의 데이터 증강 기법의 결과에 따르면 증강 샘플 수의 크기가 어느 정도 증가하였을 때 모델 성능이 점차 떨어지기 시작했으며, 이후 별다른 개선이 없는 것으로 관측되었다. 이러한 현상에 대해 두 가지 가능성을 설명할 수 있다. 첫 번째는 ACGAN 모델의 모드 붕괴 (mode collapse) 현상이다. 모드 붕괴는 GAN 모델을 학습하는 과정에서 발생할 수 있는 일반적인 문제이며, 이는 생성기에서 생성되는 합성 데이터의 다양성이 제한되는 현상을 의미한다. 실제 모드 붕괴가 발생하였는지 확인하기 위해 다중 스케일 구조 유사성 (Multi-scale structural similarity, MS-SSIM) [19], [27]을 사용하여 이미지 유사성을 정량적으로 측정하였으며, 최적의



배율인 17배 이후의 이미지들에 대한 MS-SSIM 점수의 변화를 관찰하였다.

MS-SSIM은 사람의 육안으로는 보이지 않는 이미지의 속성을 민감하게 검출해준다. 이러한 MS-SSIM은 [0, 1] 범위 사이의 값을 출력하는데, 출력 값이 높을수록 지각적으로 더 유사한 이미지에 해당하며, 값이 낮을수록 이미지 다양성이 향상됨을 의미한다. **부비동염 및 정상 클래스의 원본 학습 데이터와 ACGAN으로 생성된 데이터에서 100쌍의 데이터를 무작위로 선택하였다.** 그 결과, ACGAN 기반의 생성 이미지들의 MS-SSIM 점수는 약 0.5 주위로 진동 현상을 보였고, 이는 원본 데이터의 MS-SSIM 점수와 큰 차이가 없었다 (그림 4.1). 이는 본 연구에서 학습된 **ACGAN 모델에서는 모델 붕괴가 발생하지 않았음을 확인할 수 있었다.**

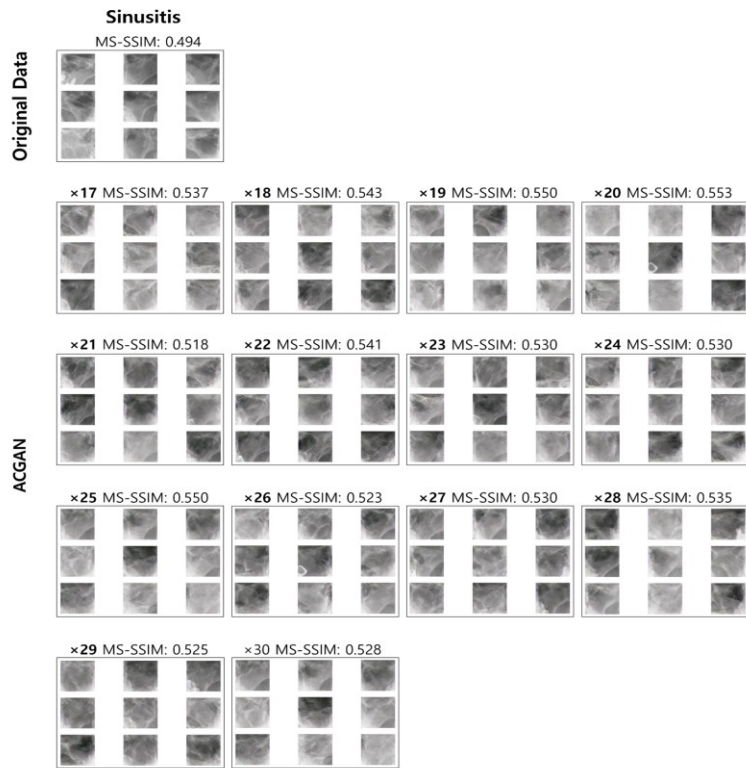
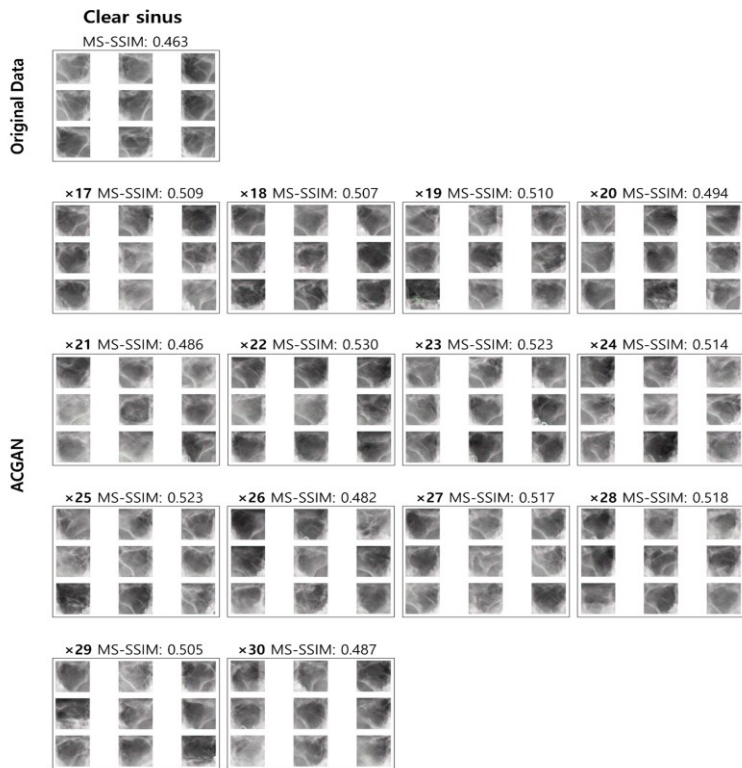


그림 4.1 MS-SSIM 결과

두 번째로는 본 연구에서 학습된 모든 모델은 고정된 학습 에폭 (epochs = 25)으로 학습되었음을 고려할 때, 데이터 증강으로 인해 학습 데이터가 샘플 수 가 증가하면서 모델의 과도한 학습으로 인해 학습 세트에 대해서 과적합 현상이 초래될 수 있었다는 것이다.

외부 검증에서 GAN 기반의 클래스 균형 및 전통적 데이터 증강 기법을 적용한 모델 (GCB+CDA)이 GAN 기반 클래스 균형 모델 (GCB) 보다 모델의 정확도가 낮아지는 현상을 보였다. 이는 모델이 내부 테스트 데이터에 상대적으로 과적합 하였기 때문일 수 있다고 생각한다. 또한 원본 데이터로만 (OD) 학습된 모델의 성능은 외부 검증에서 내부 검증 결과보다 현저히 낮았다. 이는 학습 데이터 부족으로 일어났을 가능성이 있다.

CNN은 현재 영상 분석에 가장 많이 사용되고 있는 분류 모델이다. 최근 연구 중 딥러닝을 통한 부비동염 x-ray 분석에서 VGG-16, VGG-19[24] 및 ResNet-101[25]을 통해 다수결 결정 알고리즘 (majority decision algorithm) 구현하였던 사례가 있다[26]. 내부 검증에서는 본 연구에서 제안된 모델의 정확도, 민감도 및 AUC 값이 이전 부비동 진단 관련 연구에서 제안한 모델들의 성능보다 높았음을 확인할 수 있다. 내부 검증은 딥러닝 모델의 진단 성능을 실질적으로 과대 평가할 수 있기 때문에 외부 검증은 모든 진단 기술을 검증하는데 있어 매우 중요하다. 외부 검증에서는 본 연구에서 제안한 모델의 민감도는 더 높았지만, 정확도는 다른 모델들과 큰 차이가 없었다. 부비동 x-ray를 통한 판독은 높은 민감도가 요구되는 초기 진단용 도구이며 본 연구에서 제안된 모델은 다른 모델들에 비해 가장 높은 민감도를 보였다.

이전 연구[18]에서는, 증강 데이터 샘플 개수가 특정한 규칙

없이 증대되었기 때문에, 데이터 증강을 어느 시점에서 멈추는 것이 좋을지에 대한 명확한 기준이 없었다. 본 연구에서 제안한 방법은 슬라이딩 윈도우를 이용하여 최고의 진단 성능을 보이는 데이터 증강의 최적 배율을 자동으로 탐색할 수 있었다. 이를 통해 본 연구에서 제안한 방법은 더욱 정량적이고 자동화로 인한 편리성을 제공함과 동시에 GAN을 이용한 합성 데이터 증강에 대한 효과를 입증할 수 있었다.

본 연구의 한계점은 다음과 같다. 첫째, 이상적으로는 부비동 x-ray 와 CT 촬영을 모두 당일 날 수행한 환자들을 대상으로 선정해야 했지만, 당일에 두 촬영을 모두 수행한 환자는 40명 밖에 되지 않았다. 따라서, 급성 부비동염으로 진단 받은 환자와 x-ray와 CT 촬영 기간 동안에 상태가 악화된 환자들은 제외하고 촬영 간격 기간이 14일 이내인 환자들을 실험용 데이터로 선정하였다. 둘째, GAN 기반의 합성 데이터 증강을 수행하기 위해 ACGAN 모델을 사용하였지만 다른 GAN 모델들을 적용하여 비교해보지는 않았다. 본 연구에서 ACGAN 모델은 이미지의 품질을 개선할 수 있다는 점에서 임의로 선택한 것이며 최근 개발된 새로운 GAN 기법들을 적용함으로써 더 높은 성능을 얻을 수 있다는 가능성도 배제할 수 없다. 마지막으로, 본 연구에서는 다른 의료 기관(3차 병원)에서 추가적으로 수집한 데이터를 통해 외부 검증을 수행하였다. 하지만 의료기관 마다 내재된 특성 (예: 의사의 진단 기술, 장비 유형 및 품질 등)이 다르므로 이로 인해 발생할 수 있는 데이터의 품질 차이가 연구 결과에 영향을 미쳤을 수도 있다[28]. 따라서 모델의 신뢰성을 높이기 위해 다양한 의료기관에서 획득한 데이터 추가적인 모델 평가 절차가 필요하다.

## 제 5 장 결론

본 연구에서는 실제 의료 환경에서 데이터가 부족한 상황에서도 딥러닝 기술을 적용하여 고성능의 모델을 개발할 수 있는 방법을 제안하였다. GAN을 기반으로 한 합성 데이터 증강 기법을 이용하여 데이터 부족으로 인해 발생할 수 있는 과적합 문제를 극복하고 외부 검증을 통해 모델의 일반성을 검증할 수 있었다.

또한, GAN을 활용한 의료 영상 분석 관련 연구에서 제시되지 않았던 증강 데이터의 적절 양에 대한 기준을 본 연구에서는 최고의 성능에 최적화된 모델을 자동적으로 탐색해주는 방법을 제안함으로써 개선할 수 있었다. 추후 데이터 증강 기법 관련 연구에서 본 연구에서 제안된 방법이 x-ray 영상 뿐만 아니라 다양한 의료 영상에 적용되고 특히 데이터가 충분하지 않은 1차 및 2차 의료기관에서 딥러닝을 적용하여 관련 연구 및 상업화에 활발하게 활용 하는데 기여할 수 있을 것으로 기대한다.

본 연구에서 제안한 GAN 기반의 합성 데이터 증강 기법은 데이터가 적은 상황에서도 부비동염 진단을 잘 수행할 수 있었으며, 본 연구에서 제안한 방법을 통해 임상에서의 진단 정확도 향상에도 도움이 될 수 있기를 기대한다.

## 참고 문헌

- [1] M.Y. Chen, H.S. Chiang, E. Lughofer and E. Egrioglu, "Deep Learning: emerging trends, applications, and research challenges," *Soft Computing*, vol. 24, pp. 7835–7838, 2020.
- [2] E. Montagnon, M. Cerny, A. Cadrin-Chênevert, et al., "Deep learning workflow in radiology: a primer," *Insights Imaging*, vol. 11, No. 22, 2020.
- [3] Z. Lai, and H. Deng, "Medical image classification based on deep features extracted by deep model and statistic feature fusion with multilayer perceptron," *Comput. Intell. Neurosci.* vol. 2018, pp. 2061516, 2018.
- [4] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *J. Digit. Imaging* vol. 32, no. 4, pp. 582–596, 2019.
- [5] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *J. Digit. Imaging* vol. 32, no. 4, pp. 582–596, 2019.
- [6] X. Liu, J. Wu, Z. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*. vol. 39, no. 2, pp. 539–550, 2019.
- [7] N.V. Chawla, K.W. L.O. Hall, W.P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*. vol.16, pp. 321–357, 2002.
- [8] H. Han, W.Y. Wang, B.H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *International Conference on Intelligence Computing*. vol. 3644. pp. 878–887, 2005.
- [9] H. Han, Y. Bai, E. Garcia, S. Li, "ADASYN: adaptive synthetic sampling approach for imbalanced learning," in *proceedings of the International Conference on Neural Networks*, 2008.
- [10] I. Goodfellow, et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems*. 2014.
- [11] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, "Differential data augmentation techniques for medical imaging classification tasks," *AMIA Annu. Symp. Proc.* vol. 2017, pp. 979–984, 2017.
- [12] I. Alnujaim, and Y. Kim, "Augmentation of Doppler radar data using generative adversarial network for human motion analysis," *Healthc. Inform. Res.* vol. 25, pp. 344–349, 2019.
- [13] T. Malygina, E. Elicheva, and I. Drokin, "Data augmentation with GAN: improving chest X-ray pathologies prediction on class-imbancing cases," in *Analysis of Images, Social Networks and Texts*. AIST 2019. Lecture Notes in Computer Science, vol. 11832, W. van der Aalst. et al., Eds. Springer, Cham, 2019, pp. 321–334.

- [14] M. Loey, F. Smarandache, and N. E. M. Khalifa, "Within the lack of chest COVID-19 dataset: a novel detection model based on GAN and deep transfer learning," *Symmetry* vol. 12, pp. 651, 2020.
- [15] S. Albahli, "Efficient GAN-based Chest Radiographs (CXR) augmentation to diagnose coronavirus disease pneumonia," *Int. J. Med. Sci.* vol. 17, pp. 1439-1448, 2020.
- [16] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "CovidGAN: data augmentation using auxiliary classifier GAN for improved COVID-19 detection," *IEEE Access* vol. 8, pp. 91916-91923, 2020.
- [17] T. Fujioka, et al., "Efficient anomaly detection with generative adversarial network for breast ultrasound imaging," *Diagnostics (Basel)* vol. 10, pp. 456, 2020.
- [18] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing* vol. 321, pp. 321-331, 2018.
- [19] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proceedings of the 34th International Conference on Machine Learning*. vol. 70. PMLR, 2017, pp. 2642-2651.
- [20] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proceedings of the 34th International Conference on Machine Learning*. vol. 70. PMLR, 2017, pp. 2642-2651.
- [21] B.A. Kusuma, N. Nuryani, Damanto, "Heart Sounds Determination Based on Sliding Window Maximum Method," *Journal of Physics: Conference Series*. vol. 1204, 2019.
- [22] P. Rajpurkar, et al. (2017, Nov.). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. [Online]. Available: [arxiv.org/abs/1711.05225](https://arxiv.org/abs/1711.05225). 2017.
- [23] G. Huang, Z. Liu, L. Maaten, K.Q. Weinberger, "Densely Connected Convolutional Networks," *Computer Vision and Pattern Recognition*. 2017.
- [24] K. Simonyan, and A. Zisserman. (2014, Sept.). Very deep convolutional networks for large-scale image recognition. [Online]. Available: [arxiv.org/abs/1409.1556](https://arxiv.org/abs/1409.1556).
- [25] K. He, X. Zhang, S. Ren, and J. Sun. (2017, Nov.). "Deep residual learning for image recognition," Presented at IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV. Available at <https://doi.org/10.1109/CVPR.2016.90>
- [26] H. G. Kim, K. M. Lee, E. J. Kim, and J. S. Lee, "Improvement diagnostic accuracy of sinusitis recognition in paranasal sinus X-ray using multiple deep learning models," *Quant. Imaging Med. Surg.* vol. 9, pp. 942-951, 2019.
- [27] B. Erol, S. Z. Gurbuz, and M. G. Amin, "Motion classification using kinematically sifted ACGAN-synthesized radar micro-Doppler

signatures,” *IEEE Transactions on Aerospace Electronic Systems*  
vol. 56, no. 4, pp. 3197–3213, 2020.

[28] M. J. Willemink, et al., “Preparing medical imaging data for machine learning,” *Radiology* vol. 295, pp. 4–15, 2020.



## Abstract

# A Study on the Generation of Synthetic Data using Generative Adversarial Network for Optimization of diagnostic performance on paranasal sinusitis radiography

Hye–Min Moon

Interdisciplinary Program in Bioengineering

The Graduate School

Seoul National University

Recently, rapid development of deep learning has led to a state-of-the-art performance in a wide range of computer vision tasks, mainly through large-scale datasets. However, acquiring a large amount of dataset is restricted and challenging in medical domain.

This study aimed to investigate the feasibility of synthetic data augmentation using generative adversarial networks (GAN) by suggesting an automatic method to determine the ideal amount of data augmentation in order to develop deep learning-based model optimized for best performance in a limited datasets.

This retrospective study evaluated the Water ' s view

radiographs of 250 patients diagnosed with chronic sinusitis, who underwent radiographic examination, between January 2010 and January 2020. Image patches of 177 healthy sinuses and 212 maxillary sinusitis were created and randomly split into a training set (80%) and a test set (20%) to develop a deep learning model. During synthetic data augmentation process, Auxiliary Classifier GAN (ACGAN) was applied to generate synthetic images from the original training set. A method to determine the optimal amount of GAN-generated synthetic data to train the model with the highest performance was proposed herein. Transfer learning techniques were applied using a pre-trained CheXNet model for classification. The model was trained using conventional data augmentation and the proposed synthetic data augmentation.

The performance of the models was evaluated based on the statistical analyses of the accuracy, sensitivity, specificity, positive and negative predictive values, area under the curve (AUC), and receiver operating characteristic, in both internal and external datasets. The experimental results verified that the proposed approach achieved a higher performance than conventional data augmentation, with an accuracy of 0.949, sensitivity of 0.917, specificity of 0.955, and an AUC of 0.964 using internal test set, and 0.917, 0.924, 0.909, and 0.955, respectively, using external test set. These values showed significant differences compared with the model trained using conventional data augmentation.

The findings of this study suggest that the proposed deep

learning approach could be used to assist radiologists and improve diagnosis with the use of deep learning technique even in a presence of lack of training data.

**Keywords : Data Augmentation, Computer-aided diagnosis system, Generative Adversarial Network, Artificial Intelligence**

**Student Number : 2018-20423**

## 감사의 글

가장 먼저 언제나 저의 곁에서 사랑과 믿음을 주신 존경하는 부모님께 깊은 감사의 인사 드립니다. 어떠한 상황에서도 묵묵히 곁에서 지켜봐 주시며 깊은 격려와 응원, 그리고 조언을 해주신 저희 가족분들께 진심으로 감사드리며, 넘치도록 주신 사랑과 믿음에 보답하여 자랑스러운 딸이 될 수 있도록 열심히 노력하겠습니다.

인턴 기간을 포함하여 4년 넘게 부족한 저를 지도해주시고 보살피 주신 김성완 교수님께 감사 드립니다. 연구자로서의 소양과 지혜 뿐 아니라 사회에 나아가 잘 적응할 수 있도록 전달해 주신 많은 조언들과 가르침들을 마음 깊이 새기며 살아가겠습니다. 좋은 연구를 수행할 수 있도록 공학적/이론적으로 많은 도움을 주시고 또한 지혜로운 연구자가 되기 위해 많은 조언을 해주신 공현중 교수님께 깊은 감사를 드립니다. 연구를 잘 진행할 수 있도록 도움주신 이비인후과 김진엽 선생님께 감사드립니다. 귀한 시간 내주시어 학위 심사와 소중한 조언 해주신 이정찬 교수님과 공현중 교수님께 감사의 말씀 드립니다.

대학원 생활과 생체 모델링 및 제어 (BMC) 연구실 생활 동안 뜻 깊고 즐거운 추억을 만들어 주신 모든 연구원 분들께도 감사 드립니다. 짧은 기간 함께했지만 처음 인턴으로 들어왔을 당시 많은 도움과 웃음을 주신 이치원 박사님, 김명준 박사님, 남형석 선생님, 홍나영 연구원, 그리고 대학원 생활 동안 많은 지식을 공유해 주시고 조언해 주신 조민우 교수님, 임민혁 선생님, 김윤재 박사님, 전병준 연구원, 조예은 연구원, 그리고 옆에서 많은 응원과 웃음을 안겨 주었던 윤단 연구원, 문경빈 연구원, 김영균 연구원, 김병수 연구원에게 감사의 말씀 드리며, 소중한 추억들 깊이 간직하겠습니다.