SOTA 이미지 분류 모델의 AUC 기반 성능 비교 : 의료 오픈소스 데이터의 문제와 실제 임상 환경을 반영한 복원 기반 데이터 증강 접근

이 건
숭실대학교 정보통계보험수리학과
학번 : 20220499
20220499@ssu.ac.kr

연구 질문 요약(Summary of research of questions)

1. Accuracy 기준으로 SOTA로 평가된 이미지 분류 모델들이 AUC 기준에서도 동일한 우수성을 보이는가?
   → SOTA 모델들이 Accuracy와 AUC를 계산 한 결과 Accuracy가 높은 모델이 AUC가 가장 높지는 않았다. VIT가 Accuracy Sota를 달성했으나 AUC 관점에서는 EfficientNet이 Sota 성능을 보였다.

2. Kaggle 등 오픈소스 의료 영상 데이터는 현실의 클래스 불균형을 제대로 반영하지 못하는데, 이를 어떻게 보완할 수 있는가?
   → 복원 기반 데이터 증강 기법을 통해 음성 데이터를 증강하였으며 새롭게 증강한 데이터만 기존 균형 데이터만 사용하여 훈련시킨 모델에 test한 결과 Accuracy가 0.9766으로 음성데이터에 특성에 알맞게 보완되어 증강되었다.

3. 다양한 연도별 이미지 분류 모델들 중, 의료 데이터 환경에서 AUC 기준으로 가장 우수한 모델은 무엇인가?
   → Resnet18. Densenet121, EfficientNet, VIT의 Accurcy와 AUC를 계산한 결과 VIT의 Accuracy가 가장 높았으나 AUC측면에서는 EfficientNet가 가장 우수한 결과를 보였다.

4. 모델이 병변 분류를 할 때 질병 이미지의 구도 각도 배경 등이 아닌 병변 정의에 맞게 병변을 인식하고 분류를 하였는가?
   → XAI기법의 Grad-Cam을 사용한 결과 병변의 위치를 잘 파악하고 분류를 하였다.

프로젝트 배경 및 목적(Motivation)

본인 과는 정보통계보험수리학과로 의료데이터를 다루는 교수님들이 여러분 계시고 영향을 받아 본인은 평소 의료 영상 데이터 처리에 큰 관심을 가지고 있으며, 특히 의료 분야에서의 데이터 불균형 문제에 주목해 왔다. 실제 임상 환경에서는 대부분의 영상이 정상(음성)이고, 질환(양성)은 매우 소수에 불과한다. 이러한 불균형 구조에서는 단순한 Accuracy 지표로는 모델의 진정한 성능을 평가하기 어렵고, 민감도와 특이도 간의 균형을 평가할 수 있는 AUC가 더욱 중요한 metric이 된다.

이러한 현실을 반영해, 본인은 이미지 분류 분야에서 일반적으로 사용되는 사전학습 모델들 특히 경진대회 등에서 Accuracy 기준으로 SOTA로 평가된 모델들이 과연 AUC 기준에서도 우수한 성능을 보일지에 대해 의문을 가지게 되었다. 많은 연구와 실무에서는 Accuracy 중심의 모델 선정이 이루어지고 있지만, 의료 영상처럼 클래스 불균형이 뚜렷한 상황에서는 AUC를 기준으로 한 모델 선택이 더 적절할 수 있다고 생각한다.

이 프로젝트는 단순히 모델 간 비교를 넘어서, 의료 환경에 적합한 평가 지표와 모델 선택 기준을 제시하려는 시도다. 만약 Accuracy 기반 SOTA 모델이 AUC 기준에서는 그렇지 않다는 결과가 도출된다면, 이는 현재 의료 영상 분석에 쓰이는 모델 평가 방식에 대한 근본적인 재고를 유도할 수 있을 것이다. 나아가 보다 정밀하고 신뢰할 수 있는 진단 보조 시스템 개발에 기여할 수 있는 통찰을 제공할 수 있을 것이다.

또한 의료 이미지 데이터는 민감정보를 포함하고 있어, 일반 연구자나 학생이 실제 병원 데이터를 직접 구하거나 활용하기는 매우 어렵다. 따라서 의료 분야에서 프로젝트를 진행하려면 오픈소스 데이터를 사용할 수밖에 없는 현실적인 한계가 존재한다. 그러나 대부분의 오픈소스 의료 데이터는 클래스 간 균형이 잘 맞춰져 있어, 실제 임상 환경의 양성과 음성 간 불균형 구조를 반영하지 못하는 한계가 있다. 본 프로젝트는 이러한 제약을 극복하기 위해, U-Net 기반 복원 모델을 활용한 데이터 증강 기법을 적용하여 정상 데이터를 증가시키고, 오픈소스 데이터로도 실제 병원 데이터에 가까운 불균형 구조를 시뮬레이션 할 것이다. 이처럼 단순한 회전·반전이 아닌 복원 기반 증강 방식을 통해, 오픈소스 데이터만으로도 현실 세계에 기반한 의료 AI 프로젝트를 수행할 수 있다는 가능성을 제시하고자 한다.

또한 증강한 데이터가 실제 음성 데이터 분포에 맞게 생성되었는가 확인하기 위해 EfficientNet의 새로운 모델을 만들어 기존 증강되지 않는 데이터로 훈련을 진행하여 증강시킨 데이터를 Test하여 Accuracy를 계산하여 실제 음성 데이터 분포에 맞게 생성되었는지 확인하고자 한다.

또한 모델이 병변 분류를 할 때 병변 이미지의 구도 각도 배경이 아닌 병변 정의에 맞게 Classification을 하였는지 확인하기위해 XAI 기법의 Grad-Cam을 사용하여 모델이 병변 정의에 맞게 병변의 위치를 잘 파악하고 분류 하였는지 확인하고자 한다.

‘

1 데이터셋(Data Set)

Brain Tumor (MRI Scans) 데이터를 사용
https://www.kaggle.com/datasets/rm1000/brain-tumor-mri-scans?select=healthy
healty data : 2000장 gloma data : 1621장 meningioma data : 1645장 pituitary 1757장
데이터는 1 : 1 : 1: 1 구조로 균형적인 형태이다.

1.1 의료 오픈소스 데이터의 문제점 그럼에도 의료 오픈소스 데이터를 사용 하여야만 하는 이유

Kaggle 데이터는 클래스 간 균형(1:1:1:1)이 잘 맞춰져 있어 실험에 적합해 보이지만, 실제 의료 환경을 반영하지는 못한다. 실제 병원 데이터는 음성(정상)이 양성(질병)보다 훨씬 많아 클래스 불균형이 심하며, 이는 의료 AI 모델에서 반드시 고려해야 할 현실이다. 또한 의료 데이터 같은 경우 민감 데이터이기 때문에 일반적인 학부생이 현실적인 의료 데이터를 구하기는 힘들고 오픈소스의 의료 데이터를 사용하여야 하나 오픈소스의 의료 데이터는 현실 세계를 반영하지 못하기에 데이터 증강을 통해 오픈소스로도 현실 세계에 맞는 프로젝트를 진행할 수 있음을 알리고자 한다. Kaggle은 실험의 공정성과 비교 가능성, 학습자의 접근성을 위해 균형 잡힌 데이터를 제공하지만, 이로 인해 현실성이 떨어지는 한계가 있다. 본 프로젝트는 과제 명세에 따라 오픈소스 데이터만 사용해야 하므로 Kaggle 데이터를 활용한다. 그러나 현실을 더 잘 반영하기 위해, 정상 데이터를 U-Net 기반 복원 모델을 이용해 증강하여 실제 병원과 유사한 불균형 구조를 시뮬레이션 한다. 이 증강 방식은 단순한 회전·반전이 아닌 복원 기반 증강이며, 복원 이미지가 원본과 유사해 다양성이 낮을 수 있는 한계는 있지만,

1. 실제 의료 환경에 가까운 데이터 비율을 구현하고,
2. U-Net을 실질적인 의료 데이터 분석에 확장 적용한다는 점에서 의미 있는 실험이다.

2. 방법론 (Mehotd)

2.1 METRIC : AUC (Area Under the ROC Curve)

AUC는 ROC 곡선(Receiver Operating Characteristic Curve) 아래의 면적으로 정의되며, 모델이 무작위 분류보다 얼마나 잘 구별하는지를 나타내는 지표이다. 특히 클래스 불균형이 심한 환경에서 정확도(Accuracy)가 왜곡될 수 있는 문제를 보완해 주기 때문에, 의료 영상처럼 정확한 판별이 중요한 문제에 적합하다.

2.2 데이터 증강(Method)

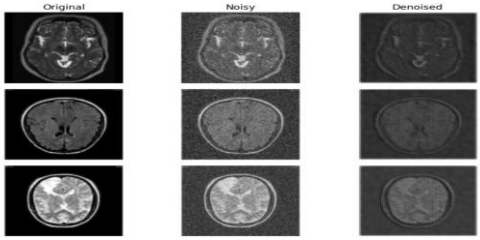
앞서 언급했듯 Kaggle 원본 데이터는 1:1:1:1 로 균형적인 데이터 형태이다. 하지만 실제 임상 데이터에서 음성 클래스가 더 많은 구조를 반영하기 위해서는 데이터를 균형이 아닌 불균형 데이터로 조절을 함으로써 실세계에 현상을 반영할 것이다. 양성 데이터를 제거함으로써 불균형을 만드는 데에는 기존 이미지가 1만장도 되지 않아 현재 사이즈 자체로도 모델이 제대로 된 학습을 하지 못할 수 있기에 데이터를 버리기 보다는 정상 데이터를 증강시켜 실세계를 반영한 불균형 구조로 만들었다.

2.2.1 데이터 증강 : 복원 기반 증강 방식 적용

일반적인 rotation, flip과 같은 증강 기법이 아닌, 본 프로젝트에서는 U-Net 기반의 복원 모델을 사용한 증강 방식을 적용하였다. 원본 정상 이미지에 0.1, 0.2, 0.3, 0.4, 0.5 수준의 가우시안 노이즈를 추가하여 U-Net으로 노이즈를 제거 학습 수행하여 학습된 U-net 모델을 통해 복원된 이미지를 생성하는 으로 증강을 하였다. 이 복원된 이미지가 데이터 증강에 역할을 할 것이고 정상 데이터에서만 데이터 증강을 하여 데이터를 균형적인 데이터에서 실세계의 현상을 반영한 불균형 데이터로 만들었다. 이를 통해 오픈소스 데이터 기반의 연구에서도 실제 의료 현장을 반영한 실험 설계가 가능함을 보여주는 사례를 제시할 것이다.

2.2.2 생성 결과 및 생성 데이터셋

기존 1:1:1:1 비율에서 healthy(음성) 0.1 0.2 0.3 0.4 0.5의 가우시안 노이즈를 추가하여 복구하여 데이터를 증강한 결과 8:1:1:1의 불균형 비율을 구축하여 현실세계의 음성과 양성간 불균형을 구축하였다.

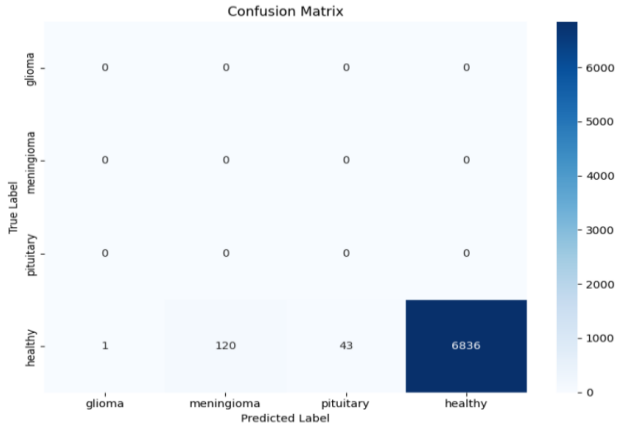


[Figure1] 원본 이미지와 가우시안 노이즈 0.1을 부과하여 증강한 이미지

[Figure1]의 Original 이미지는 원본 데이터이고 Denoised는 가우시안 노이즈 0.1을 더해서 복원한 이미지로 복원한 이미지가 새로운 훈련 데이터로 저장이 된다. 가우시안 노이즈를 0.1 0.2 0.3 0.4 0.5 를 부여하여 복원하여 생성한 이미지를 데이터 증강 기법으로 훈련 데이터의 음성 데이터로 구축하여 불균형 형태를 형성하였다.

2.2.2. 증강된 음성 이미지의 실제 음성 분포 적합성 검증

증강된 음성(healthy) 이미지가 실제 음성 이미지의 분포를 충실히 반영하고 있는지를 검증하기 위하여, 증강 기법의 효과를 간접적으로 평가하는 실험을 수행하였다. 먼저, 증강되지 않은 원본 데이터(healthy, meningioma, pituitary, glioma)를 이용하여 EfficientNet 기반 분류 모델을 학습시켰다. 이후, 학습된 모델에 대해 증강된 음성 이미지를 테스트 데이터로 입력하였고, 모델이 이를 실제 음성(healthy)으로 분류하는지를 확인하였다. 이를 통해 생성된 음성 이미지가 기존 음성 이미지와 유사한 분포를 가지는지를 간접적으로 평가하였다.



[Figure2] 증강된 데이터로 Test를 한 혼동행렬

[Figure2] 를 본다면 위에 EfficientNet 모델은 증강된 데이터로 훈련되지 않고 기존 원본 데이터(healthy, meningioma, pituitary, glioma)를 이용하여 EfficientNet 기반 분류 모델을 학습된 모델로 증강 된 데이터 healthy(음성)만 test한 결과 Accuracy: 0.9766으로 실제 증강된 데이터가 음성 분포에 적합하여 생성이 되었다.

2.3 실험조건

실험에서는 SOTA 모델의 모든 파라미터를 학습 가능하게 설정하였다. 이는 분류층(classification head)만 학습했을 때보다 전체 모델(feature extractor 포함)을 함께 학습한 경우 정확도가 더 높았기 때문이다. 최적화는 Adam 옵티마이저(eps=1e-8, lr=0.0001, weight_decay=1e-4)를 사용하였다.

Loss 는 cross_entrpoy_loss 로 사용하였으며 학습률은 ReduceLROnPlateau 스케줄러로 조절했으며, test 손실이 3 회 연속 줄지 않으면 학습률을 0.1 배로 감소시켰다. 컴퓨팅 자원을 고려하여 최대 에폭은 10 으로 설정하고, 테스트 손실 기준으로 2 회 연속 개선되지 않으면 조기 종료하는 방식으로 학습을 진행하였다. 이미지 전처리(Transform) 과정에서는 SOTA 모델이 224x224 크기의 이미지를 입력으로 받도록 설계되어 있기 때문에, 모든 이미지를 해당 크기로 Resize 해주는 것이 필요하다. VIT 같은 경우 이는 이미지가 일관된 크기를 가져야 패치 단위로 분할하고 transformer 구조에 맞게 처리할 수 있기 때문이다. 또한 RandomHorizontalFlip 을 사용하여 이미지를 무작위로 좌우 반전시켜 데이터 다양성을 높혔다. 뇌 MRI 는 좌우 대칭 구조이므로 이를 활용해 모델의 일반화 성능을 향상시키기 위해 사용하였다.

2.4 데이터 증강을 적용한 모델

데이터 증강 기법을 적용한 결과, 총 11,914 장의 훈련(Train) 데이터와 2,109 장의 테스트(Test) 데이터로 구성된 최종 데이터셋을 구축하였다. 본 연구에서는 실제 임상 환경에서 질병 보유자보다 건강한 사람이 훨씬 많은 분포 특성을 반영하기 위해, 훈련 데이터 내 healthy(음성) 클래스의 수를 8,400 장으로 증강함으로써 고의적인 데이터 불균형을 유도하였다. 반면, glioma, meningioma, pituitary 의 각 질병 클래스는 각각 1,134 장, 1,151 장, 1,229 장으로 구성되어 있으며, 이는 양성(질병) 대비 음성(정상) 데이터가 불균형하게 분포하는 실제 의료 영상 환경을 모사한 것이다.
테스트 데이터는 glioma 487 장, meningioma 494 장, pituitary 528 장, healthy 600 장으로 구성되어 있으며, 평가의 공정성을 확보하기 위해 모든 테스트 샘플은 증강되지 않은 원본 이미지로만 구성하였다.
분류 모델로는 ResNet-18, DenseNet-121, EfficientNet, Vision Transformer(ViT) 모델을 사용하였으며, 각 모델의 성능은 Accuracy, AUC, Sensitivity(Recall), Specificity, F1-score 의 지표를 기반으로 비교 평가하였다. 이를 통해 불균형한 데이터 분포 상황에서도 각 모델이 얼마나 안정적으로 음성 및 양성 클래스를 구분할 수 있는지를 정량적으로 분석하였다.

2.5 데이터 증강을 적용하지 않은 모델

본 연구에서 사용된 초기 데이터셋은 총 4,914장의 훈련(Train) 데이터와 2,109장의 테스트(Test) 데이터로 구성되어 있으며, 모든 샘플은 증강되지 않은 원본 이미지로만 이루어져 있다. 훈련 데이터는 glioma 1,134장, meningioma 1,151장, pituitary 1,229장, healthy 1,400장으로 구성되어 있어 클래스 간 데이터 분포가 대체로 균형을 이루며, 약 1:1:1:1의 구조를 갖는다. 테스트 데이터 또한 glioma 487장, meningioma 494장, pituitary 528장, healthy 600장으로

구성되어 있으며, 모델 성능 평가를 위한 클래스 균형이 유지되도록 설계되었다. 이후 [2.4절]에서 다룰 다양한 모델 중, EfficientNet은 AUC 관점에서 가장 우수한 성능을 보인 모델로 나타났다. 이에 따라, 본 연구에서는 증강이 적용되지 않은 1:1:1:1 비율의 초기 데이터셋을 기반으로 EfficientNet 모델을 학습하고, 해당 모델의 성능 지표(Metric)를 계산하였다. 이를 바탕으로, 복원 기법을 통해 현실 세계의 데이터 분포를 모사하여 음성(healthy) 클래스의 수를 증가시킨 불균형 데이터셋으로 학습된 모델과의 성능을 비교하였다. 본 비교를 통해, 증강 기반의 시뮬레이션 모델이 균형 데이터 기반 모델 대비 성능 차이가 크지 않음을 보임으로써, 현실적인 데이터 분포를 반영한 증강 전략의 타당성을 제시하고자 한다.

3. 결과 (Results)

3.1 데이터 증강을 적용한 모델 Metric

모델명	Accuracy	AUC	Recall	Specificity	F1-score
Resnet18	0.9806	0.999040	0.9944	0.9615	0.9914
Densenet121	0.9810	0.998193	0.9913	0.9494	0.9880
EfficientNet	0.9848	0.999051	0.9944	0.9615	0.9914
VIT	0.9858	0.998637	0.9732	0.9717	0.9923

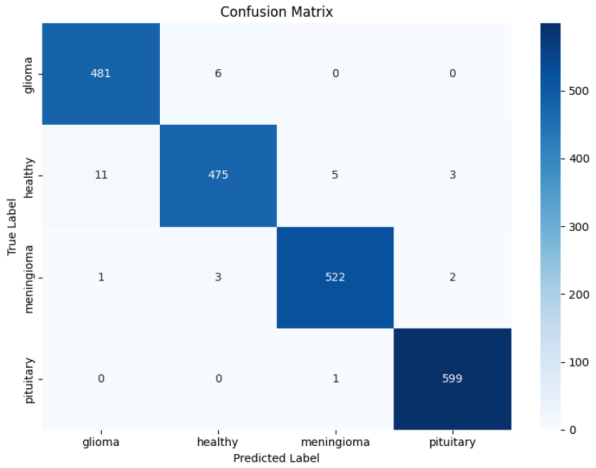
[Table1] 데이터 증강을 적용한 SOTA 모델의 의료 Metric

[Table1] 같은 경우 각 모델의 훈련 결과가 하나의 ipynd가 아닌 모델별로 ipynd파일을 만들었기에 모델에 대한 metric이 서로 다른 ipynd에 출력되었기에 수작업으로 표로 만들었다.

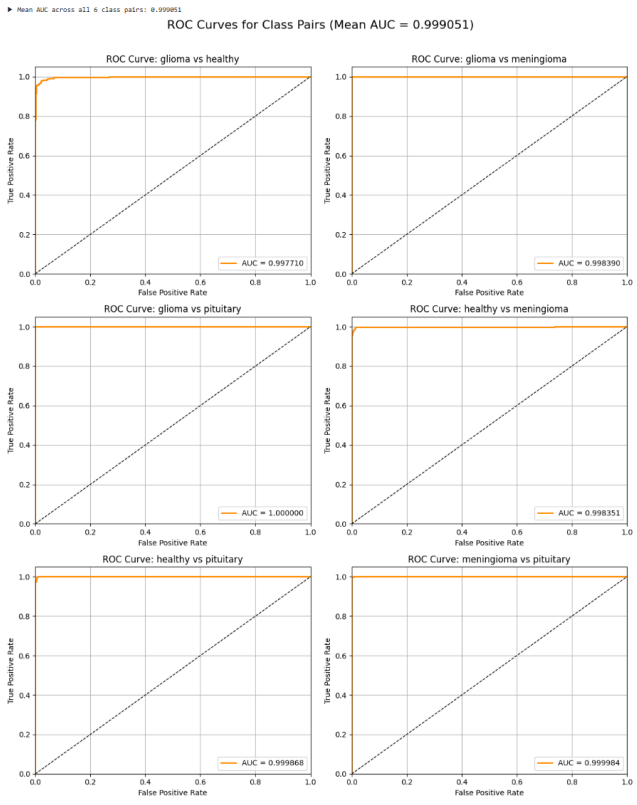
모델별 성능 비교 결과, ResNet-18, DenseNet-121, EfficientNet, Vision Transformer(ViT) 순으로 최근 연도에 발표된 모델일수록 Accuracy가 점진적으로 향상되는 경향을 보였다. 이는 모델 아키텍처의 발전이 실제 분류 정확도의 향상으로 이어진다는 일반적인 기대와 일치하는 결과이다.

그러나 AUC(Area Under the ROC Curve) 측면에서는 이러한 성능 향상이 동일한 양상으로 나타나지 않았다. 예를 들어, 가장 높은 Accuracy를 기록한 ViT(0.9858)는 AUC에서 EfficientNet(0.999051)보다 낮은 0.998637을 기록하였으며, DenseNet-121 또한 Accuracy는 ResNet-18보다 높았지만 AUC는 오히려 낮았다. 이는 Accuracy 지표가 단일 임계값 기준의 분류 정확도를 반영하는 반면, AUC는 모델의 전반적인 분류 민감도와 특이도를 종합적으로 평가하는 지표로서, 모델의 확률 예측 분포의 품질까지 포함한 보다 정교한 측면을 평가함을 시사한다. 따라서, 최신 모델일수록 Accuracy는 향상되는 경향을 보이나, AUC는 반드시 그에 비례하여 향상되지 않을 수 있으며, 이는 모델 선택 시 평가 지표 간 균형 있는 고려가 필요함을 의미한다.

결론적으로 Accuracy가 우수할수록 AUC도 비례하여 우수하지는 않으며 Accuarcy 관점에서는 최근 연도에 발표된 VIT가 SOTA 성능을 보였으나 AUC 관점에서는 EfficientNet이 SOTA 성능을 보였다.



[Figure3] 3.1에서 SOTA 모델을 보인 EfficientNet의 혼동행렬



[Figure4] 3.1에서 SOTA 모델을 보인 EfficientNet의 AUC

[Figure3] 과 [Figure4] 은 데이터 증강을 적용한 EfficientNet의 Metric을 보여주는 지표로서 높은 성능지표를 보이고 있다. 훈련 데이터로 음성 데이터가 증강된 데이터가 적용이 되었지만 테스트 데이터는 증강이 적용되지 않은 원본 이미지로 구축되어 있다.

3.2 데이터 증강을 적용한 EfficienNet과 데이터 증강을 적용하지 않은 EfficientNet 비교

모델명	Accuracy	AUC	Recall	Specificity	F1-score
EfficientNet (데이터 증강 적용)	0.9848	0.999051	0.9944	0.9615	0.9914
EfficientNet (데이터 증강 적용X)	0.9862	0.998939	0.9987	0.9971	0.9977

[Table2] 데이터 증강이 적용한 EfficientNet과 데이터 증강을 적용하지 않은 EfficientNet의 Metric

위 2개의 서로 다른 ipynd에 만들어져있기에 metric이 서로 다른 ipynd에 출력되어있기에 출력결과를 갖고 수작업으로 테이블을 만들었다.

[Table2] 결과를 본다면 결과는 매우 인상적이며, 데이터 증강 방식의 효과에 대해 중요한 시사점을 제공한다. 동일한 EfficientNet 모델을 기반으로, 증강된 불균형 데이터(healthy 클래스 과다)를 활용한 모델은 Accuracy 0.9848, AUC 0.999051을 기록하였고, 반면 데이터 증강 없이 균형 잡힌 원본 데이터로 학습된 모델은 Accuracy 0.9862, AUC 0.998939로, 오히려 AUC 지표에서 증강 모델이 더 높은 성능을 보였다.
이는 본 연구에서 사용한 복원 기반 데이터 증강 기법의 유효성을 보여주는 결과로 해석할 수 있다. 본 기법은 단순한 이미지 회전, 자르기, 밝기 조절과 같은 전통적인 증강 방식이 아니라, 음성(healthy) 클래스의 실제 분포 특성을 모사하여 다양성을 확보하도록 설계된 복원 중심의 생성적 증강 방법이다. 이러한 복원 방식은 기존 healthy 데이터의 구조적 패턴과 시각적 특징을 유지하면서도, 보다 다양한 표현을 포함하도록 하여 모델이 음성 클래스에 대한 표현력을 향상시킬 수 있도록 도와준다.
따라서, 이와 같은 증강 방식은 단순히 데이터 양을 늘리는 것 이상의 효과를 가지며, 모델이 음성과 양성 클래스 간의 미세한 경계를 더 정밀하게 학습할 수 있도록 기여한 것으로 보인다. 그 결과, Accuracy와 같이 단일 임계값 기반의 평가 지표에서는 소폭의 차이가 있었지만, 모델의 전체적인 분류 민감도와 특이도를 반영하는 AUC에서는 오히려 더 우수한 결과를 보인 것이다.
결론적으로, 복원 기반 증강 기법은 단순한 양적 확장이 아니라 분포적 정합성과 표현 다양성의 확보를 통해 모델의 일반화 성능에 긍정적인 영향을 미칠 수 있음을 본 실험을

통해 확인할 수 있었으며, 이는 향후 현실 세계의 불균형 의료 데이터 처리에 있어 효과적인 접근법으로 활용될 수 있음을 시사한다.

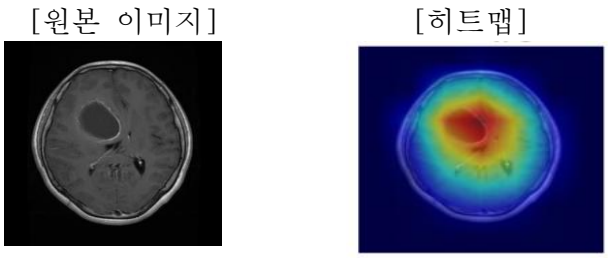
3.3 도출

모델별 성능 비교 결과, 연도별로 최신에 발표된 모델일수록 Accuracy는 향상되는 경향을 보였으나, AUC 측면에서는 이러한 향상이 반드시 비례하지는 않았다. 특히, ViT가 Accuracy 기준 SOTA 성능을 보인 반면, AUC 기준에서는 EfficientNet이 가장 우수한 성능(0.999051)을 나타냈다. 이는 분류 임계값에 기반한 Accuracy와 달리, AUC가 모델의 예측 확률 분포 품질까지 반영하는 보다 정교한 지표임을 시사한다. 또한, 본 연구에서 제안한 복원 기반 데이터 증강 기법은 불균형 데이터 구성에도 불구하고 높은 AUC 성능을 기록하며, 해당 증강 방식의 효과성과 현실적 유효성을 입증하였다.

4 추가분석

모델의 테스트 정확도가 0.9848로 매우 높아, 실제로 병변 부위를 보고 판단했는지 혹은 병변(종양)이 아닌 구도, 각도, 배경을 보고 classification을 하였는지 확인하기 위해 XAI 기법의 Grad-CAM을 사용하였다. XAI는 모델의 의사결정 과정을 사람이 이해할 수 있도록 돕는 기술이며, Grad-Cam은 모델이 어떤 이미지 영역을 보고 예측했는지 시각적으로 보여주는 방법이다. 이를 통해 모델이 단순히 병변 이미지의 외형이나 배경이 아닌 실제 병변 부위를 근거로 분류했는지 시각적으로 검증하였다.

[glioma] Glioma는 뇌나 척수에 생기는 종양으로, 신경교세포에서 발생하는 대표적인 뇌종양이다.



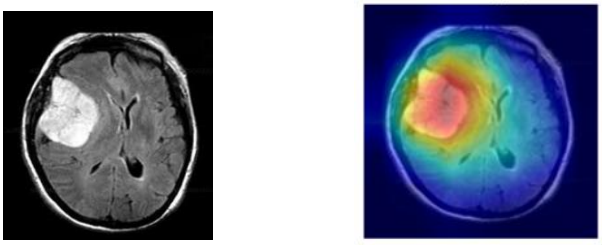
[Figure5] Glioma의 원본 이미지와 Grad-Cam을 적용한 히트맵

[Figure5]은 Grad-CAM에서 생성된 히트맵은 모델이 분류를 할 때 어느 영역을 중요하게 생각했는지를 시각적으로 보여주는 지도이며, 색상은 중요도를 나타낸다. 이때 빨간색은 모델이 예측을 내릴 때 가장 주목한 부분을 의미하며, 해당 영역이 병변의 위치와 잘 일치할 경우 모델이 실제

병변 부위를 효과적으로 인식하고 있음을 의미합니다.(파랑색으로 갈 수록 덜 중요함을 의미한다.) 따라서 히트맵에서 병변 부위에 선명한 빨간색이 나타난다면, 이는 모델이 병변을 정확하게 탐지하고 판단에 활용하고 있다는 것을 시각적으로 보여주는 결과다.

[menigioma] meningioma 는 뇌와 척수를 둘러싼 막, 즉 뇌막(meninges)에서 발생하는 종양을 말한다

[원본 이미지] [히트맵]

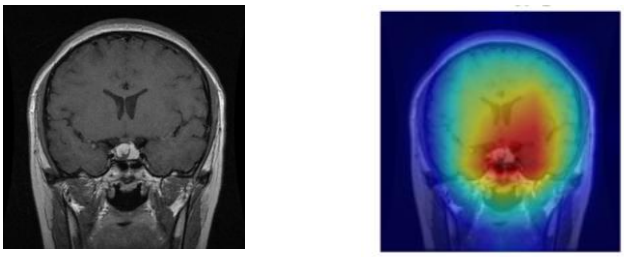


[Figure6] menigioma의 원본 이미지와 Grad-Cam을 적용한 히트맵

마찬가지로 뇌막에 발생한 종양을 잘 탐지하고 이 부분으로 인해 menigioma 라고 분류를 했음을 알 수 있다.

[Pituitary] Pituitary 종양은 뇌하수체에서 발생하는 종양으로, 호르몬 분비에 영향을 미치거나 주변 뇌 조직을 압박할 수 있다.

[원본 이미지] [히트맵]



[Figure7] Pituitary의 원본 이미지와 Grad-Cam을 적용한 히트맵

마찬가지로 뇌하수체에 발생한 종양을 잘 감지하였음을 알 수 있었다.

5 영향력 (Impact)

본 연구의 결과는 불균형한 의료 영상 데이터 환경에서의 분류 모델 개발에 있어 실질적이고 의미 있는 영향을 미칠 수 있다. 특히, 복원 기반 데이터 증강 기법이 클래스 불균형 상황에서도 높은 AUC 성능을 기록함으로써, 기존의 단순한 양적 증강 접근법을 넘어, 실제 데이터 분포를 고려한 정교한 증강 방식의 가능성을 제시하였다. 이로 인해 의료 인공지능

연구자, 영상 진단 시스템 개발자, 그리고 병원 등에서 실제 모델을 적용하고자 하는 임상의들에게 현실적이고 효과적인 데이터 구성 전략을 제공할 수 있다. 본 연구에서 제안한 방식은 특히 건강한 사람의 수가 압도적으로 많은 현실 의료 환경을 시뮬레이션하고자 할 때 유용하게 활용될 수 있으며, 데이터 수집이 제한적인 환경에서도 효과적인 분류 성능을 확보할 수 있는 대안을 제시한다. 반면, 이미 균형 잡힌 데이터를 보유한 기관이나, 복원 품질 통제가 어려운 환경에서는 본 기법의 적용이 오히려 성능 저하나 예측 신뢰도 감소로 이어질 수 있어 주의가 필요하다.
결론적으로 본 연구는 의료 데이터의 불균형 문제에 대해 보다 실용적이고 현실적인 접근법을 제시하였으며, 향후 의료 인공지능 분야에서의 데이터 설계 및 증강 전략 수립에 있어 중요한 참고가 될 수 있는 잠재력을 지닌다.

6 Challenge Goal (기대효과)

1. 의미 있는 모델 성능 비교 (가장 열정을 느끼는 challenge goal)
본 프로젝트에서 가장 열정을 느낀 도전 목표는 단순히 Accuracy 지표만으로 모델 성능을 평가하는 기존의 방식에서 벗어나, 불균형 데이터 상황에 더 적합한 AUC 지표를 활용해 SOTA 모델들을 재평가하는 것이었다. 의료 영상 데이터는 현실적으로 질병(양성)보다 건강한 사례(음성)가 훨씬 많은 불균형 구조를 가지므로, Accuracy만으로 모델을 평가할 경우 실제 현장에서의 성능을 제대로 반영하지 못할 수 있다.
이에 따라 연도별 주요 SOTA 모델들(ResNet-18, DenseNet-121, EfficientNet, ViT)을 동일한 조건에서 학습시켜 Accuracy와 AUC를 모두 비교 분석하였다. 그 결과, 가장 높은 Accuracy를 기록한 ViT는 AUC 기준에서는 EfficientNet보다 낮은 성능을 보였으며, 오히려 EfficientNet이 AUC 기준에서 가장 우수한 성능을 나타냈다. 이로써 Accuracy가 높은 모델이 항상 AUC에서도 우수하지는 않다는 점을 실증적으로 확인할 수 있었다.
이러한 분석은 단순한 수치 비교를 넘어, 실제 의료 환경에 적합한 모델을 평가하고 선택하는 데 있어 보다 현실적인 기준을 제시할 수 있다는 점에서 중요한 의미를 가진다. Accuracy 기반 평가에 익숙한 기존 관행에 문제를 제기하고, AUC 기반의 재해석을 통해 더 신뢰도 높은 모델 선택이 가능함을 보여주었다는 점에서 본 과제는 의미 있는 도전 과제였다고 생각한다

2. 의료 오픈소스로도 유의미한 의료 프로젝트를 진행 가능함을 제시
의료 오픈소스 데이터를 활용한 본 프로젝트는, 기존 공개 데이터들이 대부분 클래스 간 균형을 이루고 있어 실제 임상 환경의 불균형 구조를 반영하지 못한다는 한계를 지적하고, 이를 극복하고자 했다. 이를 위해 U-Net 기반 복원 모델을 활용한 데이터 증강 기법을 적용하여 정상(음성) 데이터를 의도적으로 증강함으로써 현실적인 클래스 불균형을 시뮬레이션하였다. 이 과정을 통해, 오픈소스 기반 연구 환경에서도 실제 의료 현장의 데이터 특성을 반영한 실험 설계가 충분히 가능함을 보여주는 사례를 제시했다.

3. 과제 기반 실제 확장 적용(개인적 차원에서 Challenge goal)
과제 2에서 다룬 U-Net 기반 복원 모델을 단순히 이론 실습에 그치지 않고, 본 프로젝트에 실질적으로 적용함으로써, 학습 내용을 실제 문제 해결에 연결하는 경험을 쌓고자 합니다. MNIST 같은 단순 데이터셋이 아닌 의료 영상이라는 실제적인 데이터에 적용해본다는 점에서 교육적 의의가 있으며, 이 과정을 통해 실습 기반 결과 도출과 학습의 확장성을 체험할 수 있었다.

6 Work Plan Evaluation
초기 제안서에서는 총 5개의 모델을 각각 독립적으로 학습시키는 계획이었으며, 모델 하나를 학습하는 데 하루가 걸릴 것으로 예상하여 모델 훈련 작업 기간을 약 5일로 산정하였다. 하지만 실제 구현 과정에서는 하나의 기본 모델 구조를 구축한 뒤, 코드의 일부만 수정하여 다른 모델 실험을 수행할 수 있었기 때문에, 모든 모델 실험을 하루 만에 끝낼 수 있었다. 이는 코드의 모듈화와 재사용 가능한 구조 설계 덕분으로, 당초 예상보다 훨씬 효율적으로 작업이 진행되었다.
또한 데이터 구성에 있어서도 차이가 발생하였다. 원래는 뇌종양이 인구 1만 명당 1명꼴로 발생한다는 실제 유병률을 반영하여 극단적인 데이터 불균형 상황을 구축하고자 하였으나, 해당 비율을 현실적으로 구현하기 위해서는 방대한 양의 컴퓨팅 자원과 시간이 필요했다. 이에 따라 실제로는 8:1:1:1 (glioma:meningioma:pituitary:healthy) 비율로 데이터를 증강하여 모델을 학습시켰다. 이는 현실의 분포를 완전히 반영하진 못하지만, 제한된 자원 내에서 불균형 문제를 다루는 현실적인 절충안이었다.
한편, 사전 계획에는 없었지만, 모델의 분류 정확도가 매우 높게 나타난 것을 계기로 XAI 기법 중 하나인 Grad-CAM을 추가적으로 적용하였다. 이는 모델이 병변을 올바르게 이해하고 있는지, 단순히 배경이나 이미지 구도를 기반으로

판단하고 있는 것은 아닌지에 대한 의문에서 출발한 작업이었다. Grad-CAM을 통해 시각화한 결과, 모델은 실제 병변 영역에 집중하여 판단하고 있었으며, 이는 단순한 정확도 수치 이상의 신뢰도를 확보하는 데 중요한 근거가 되었다.
이와 같이, 전반적인 작업 계획은 초안 대비 더 빠르게 진행되었고, 데이터 구성과 분석 해석 면에서는 예상과 다른 현실적 한계에 따라 전략적으로 조정이 이루어졌다.
결과적으로는 계획보다 효율적인 수행과 더 깊이 있는 해석을 병행할 수 있었던 작업이었다.
또한, 모델 평가에 있어서도 초기 흥미로운 결과를 얻을 수 있었다. 일반적으로 분류 정확도(Accuracy)가 높으면 의료 분야에서 주요하게 사용되는 평가 지표인 AUC(Area Under the Curve) 또한 높을 것으로 예상하지 않았고 이 예상이 맞아 떨어졌다. 일부 모델에서는 Accuracy가 높았음에도 불구하고 AUC가 상대적으로 낮게 측정되었으며, 이는 Accuracy가 클래스 간의 불균형을 충분히 반영하지 못하는 한계를 드러낸 것으로 해석된다.
예상과는 다르게 특히 흥미로웠던 점은, 균형 잡힌 데이터로 학습한 모델보다 복원 기반의 데이터 증강을 통해 학습한 모델의 AUC가 더 높게 나타났다는 점이다. 이는 예상과는 상반된 결과로, 복원 기반 증강 데이터가 실제 병변의 시각적 특성과 구조를 보다 자연스럽고 다양하게 반영하여, 모델이 병변의 경계나 특징을 보다 일반화 가능하게 학습하도록 도와준 결과로 보인다. 즉, 복원 기반 증강은 단순한 양적 확장 이상의 효과를 제공하며, 모델의 판별 능력을 정밀하게 향상시키는 데 기여했음을 알려준다.

7 Testing

본 연구에서는 제안한 데이터 증강 기법 및 분류 모델의 신뢰성과 타당성을 확보하기 위해 다양한 관점에서 정량적·정성적 검증 절차를 수행하였다. 첫 번째 검증은 증강된 healthy(음성) 데이터가 실제 정상 영상의 특징을 제대로 반영하고 있는지를 평가하기 위한 실험으로, 증강되지 않은 원본 데이터로만 학습한 EfficientNet 모델에 증강된 healthy 데이터를 입력하여 테스트하였다. 그 결과, Accuracy는 0.9766을 기록하였으며 이는 증강 데이터가 정상 데이터의 특징을 충분히 학습 가능한 품질로 생성되었음을 시사한다.
두 번째 검증은 데이터 증강을 통해 구성한 불균형 구조 기반 모델이 실제로 유의미한 결과 이루는지를 평가하기 위한 비교 실험이다. 이를 위해 동일한 EfficientNet 아키텍처를 기반으로, 하나는 클래스 간 균형이 유지된 원본 데이터로, 다른 하나는 정상 데이터를 증강하여 불균형을 유도한 데이터로 각각 학습시켰다. 그 결과, 불균형 데이터 기반

모델은 Accuracy는 다소 감소하였으나 AUC는 오히려 증가하는 결과를 보였다. 이는 Accuracy와 달리 AUC가 클래스 불균형 상황에서 보다 적합한 평가 지표로 기능함을 의미하며, 제안된 증강 방식이 실제 의료 환경에서 요구되는 민감도와 특이도 측면에서 더 나은 분류 성능을 제공함을 입증하였다.
마지막으로, 높은 Accuracy에도 불구하고 모델이 병변 자체가 아닌 배경, 촬영 구도 등의 비의학적 특징에 의해 분류를 수행한 것이 아닐까 하는 우려를 해소하기 위해, XAI(Explainable AI) 기법인 Grad-CAM(Gradient-weighted Class Activation Mapping)을 활용한 시각적 분석을 수행하였다. 분석 결과, 모델은 실제 병변 영역에 집중하여 활성화 지도를 형성하고 있었으며, 이는 모델이 의학적으로 타당한 기준에 따라 병변을 분류하고 있음을 뒷받침한다.
이러한 다층적 검증 절차를 통해 본 연구에서 제시한 복원 기반 증강 기법 및 분류 모델의 성능과 신뢰성은 충분히 확보되었으며, 실제 임상 환경에 적용 가능한 수준의 의료 AI 모델로서의 가능성을 제시하였다.

7 Collaboration (협업)

2025년 6월 4일, 정다흰 교수님과의 미팅을 통해 본 연구의 모델 성능에 대한 중요한 논의가 이루어졌다. 당시 논의의 핵심은 본 연구에서 사용된 모델이 매우 높은 Accuracy를 기록한 반면, 이와 같은 성능이 데이터 자체의 구조적 문제에서 기인한 것은 아닌지에 대한 우려였다. 특히, 모델이 실제 병변의 정의에 기반하여 분류를 수행하기보다는, 이미지의 구도, 촬영 각도, 배경과 같은 병변과 무관한 요소에 의존하여 분류 결정을 내리고 있을 가능성이 제기되었다.
이에 정다흰 교수님은 설명 가능한 인공지능(Explainable AI, XAI) 기법 중 하나인 Grad-CAM(Gradient-weighted Class Activation Mapping)을 활용한 시각적 분석을 제안하였다. 해당 기법은 모델이 분류 결정을 내릴 때 집중하는 이미지 영역을 시각적으로 표현할 수 있는 방법으로, 모델의 의사결정 근거를 확인하는 데 유용하다. 본 연구에서는 교수님의 제안을 반영하여 Grad-CAM을 적용한 분석을 수행하였고, 모델이 예측에 기여한 활성화 영역을 시각화한 히트맵을 원본 영상과 함께 비교하였다. 그 결과, 우려한 상황이었던 모델이 배경이나 구도와 같은 부차적 요소가 아닌, 실제 병변이 위치한 영역에 집중하여 분류를 수행하고 있음을 확인할 수 있었다. 이는 본 연구의 모델이 병변의 정의에 부합하는 시각적 특징을 기반으로 학습되었음을 입증하며, 모델의 예측 근거가 의학적으로 타당하다는 점을 강화하는 결과였다. 이와 같은 협업 과정을 통해, 단순한 정량적 지표를 넘어 정성적 분석을 통해 모델의 신뢰성과 해석 가능성을 높일 수

있었으며, 이는 의료 영상 분야에서 인공지능 모델을 임상에 적용하는 데 있어 중요한 검증 절차로 작용하였다.