JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 11, pp. 181-189, November 2026
https://doi.org/10.9708/jksci.2026.31.11.181

Analysis of Class Imbalance Mitigation Techniques on Performance and Grad-CAM Explainability in Medical Image Classification

Kwangho Ko*, Sooyeon Bong**, Hyunjun Kim***, Dongheon Lee**

*Professor, Dept. of Industrial Artificial Intelligence, Sungkyunkwan University, Suwon, Korea
**Student, Dept. of Industrial Artificial Intelligence, Sungkyunkwan University, Suwon, Korea
***Student, Dept. of Global Biomedical Engineering, Sungkyunkwan University, Suwon, Korea

[Abstract]

This study examines how class imbalance mitigation affects classification performance and Grad-CAM explainability in chest X-ray classification. Using the RSNA Pneumonia Detection dataset and a ResNet50V2 transfer learning model, we compared downsampling, oversampling, weighted cross entropy, and focal loss with a baseline and an augmentation-only control across repeated training seeds. Classification performance was evaluated alongside Grad-CAM localization, faithfulness, activation area, and seed stability. At a fixed threshold, imbalance mitigation increased recall but reduced precision, without consistent AUROC improvements. CAM activation areas decreased, and sampling-based methods showed lower seed agreement. Furthermore, insertion AUC increased alongside deletion AUC and was strongly correlated with positive prediction probability. These findings indicate that classification performance alone is insufficient to assess explanation quality, which should be evaluated independently while accounting for changes in model prediction probabilities.

▸ **Key words**: Class Imbalance, Pneumonia Detection, Chest X-ray, Grad-CAM, Explainable AI

[요 약]

본 연구는 클래스 불균형 완화 기법이 흉부 X-ray 분류 성능과 Grad-CAM 설명 품질에 미치는 영향을 분석하였다. RSNA 폐렴 탐지 데이터셋과 ResNet50V2 전이학습 모델을 활용하여 다운샘플링, 오버샘플링, 가중 교차 엔트로피, Focal Loss를 Baseline 및 데이터 증강 단독 대조군과 비교하고 각 조건을 반복 학습하였다. 분류 성능과 함께 Grad-CAM의 국소화, 충실도, 활성 면적 및 시드 간 안정성을 평가하였다. 분석 결과, 불균형 완화 기법은 고정 임계값에서 Recall을 높였으나 Precision은 감소하였고 AUROC의 일관된 개선은 확인되지 않았다. CAM 활성 면적은 감소하였으며, 샘플링 기반 기법의 시드 간 일치도도 낮았다. 또한 iAUC는 dAUC와 함께 증가하면서 양성 예측 확률과 강한 상관관계를 보였다. 이는 불균형 의료영상 모델에서 분류 성능과 설명 품질을 독립적으로 평가하고 예측 확률의 변화도 함께 고려해야 함을 시사한다.

▸ **주제어**: 클래스 불균형, 폐렴 탐지, 흉부 X-ray, Grad-CAM, 설명 가능 인공지능

• First Author: Kwangho Ko, Corresponding Author: Dongheon Lee
*Kwangho Ko (kwangho4162@skku.edu), Dept. of Industrial Artificial Intelligence, Sungkyunkwan University
**Sooyeon Bong (asfrrfyxhj@g.skku.edu), Dept. of Industrial Artificial Intelligence, Sungkyunkwan University
***Hyunjun Kim (Hyeonjung512@gmail.com), Dept. of Global Biomedical Engineering, Sungkyunkwan University
**Dongheon Lee (dongheonlee@skku.edu), Dept. of Industrial Artificial Intelligence, Sungkyunkwan University
• Received: 2026. 09. 07, Revised: 2026. 10. 06, Accepted: 2026. 11. 04.

Copyright © 2026 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945
www.kci.go.kr

I. Introduction

최근 딥러닝을 활용한 의료영상 진단 시스템은 높은 분류 성능을 보이고 있다. 그러나 폐렴 탐지와 같은 실제 임상 환경에서는 정상 대비 질환 샘플이 부족한 클래스 불균형(class imbalance)이 빈번히 발생하며, 이로 인해 모델이 소수 클래스인 질환을 놓치는 문제가 나타난다[1][2]. 이를 해결하기 위해 학습 데이터의 분포를 조정하는 데이터 수준의 기법과 손실함수를 조정하는 알고리즘 수준의 기법 등 다양한 방법론이 제안되어 왔다[3][4]. 이러한 기법의 효과는 주로 분류 성능 지표로 평가되어 왔다. 그러나 최근 연구들은 불균형 완화가 모델의 예측 확률과 예측 행동 자체를 바꿀 수 있음을 보였다[5][6]. 즉, 불균형 완화의 효과는 분류 성능만으로는 충분히 평가하기 어렵다.

한편 의료영상 분야에서는 모델의 판단이 실제 병변에 근거한 것인지 확인할 수 있어야 한다. 판단 근거를 확인할 수 없다면 판독의가 오판을 걸러내기 어렵기 때문이다. 이를 위해 설명 가능 인공지능(Explainable AI, 이하 XAI)이 활발히 연구되고 있으며, 그중 그래디언트 기반 클래스 활성화 맵(Gradient-weighted Class Activation Mapping, 이하 Grad-CAM)이 널리 활용되고 있다[7]-[9]. 그러나 Grad-CAM 설명은 재학습만으로 달라질 수 있고, 이를 평가하는 지표에도 한계가 있어 그 자체로 신뢰하기 어렵다[10]-[12].

불균형 완화 기법이 모델의 예측 확률을 바꾼다면 설명의 형태뿐 아니라 그 평가 지표 값도 함께 달라질 가능성이 있다. 그럼에도 이를 체계적으로 비교한 연구는 제한적이다. 불균형 완화 기법은 의료 AI 개발에서 널리 적용되며[3][4], 임상 AI 모델은 배포 이후에도 재학습이 필요하다는 점이 강조되고 있다[13]. 따라서 기법 선택이나 재학습에 따라 설명이 달라진다면 판독의는 일관되지 않은 근거를 제공받게 되고, 지표상의 개선을 실제 설명의 개선으로 오인하여 설명의 신뢰도를 과대평가할 수 있다.

이에 본 연구는 서로 다른 불균형 완화 전략이 Grad-CAM 설명을 어떻게 변화시키는지 분석하는 데 초점을 둔다. RSNA 폐렴 탐지 데이터셋과 ResNet50V2를 사용하여 데이터 수준 기법과 알고리즘 수준 기법을 여러 랜덤 시드(random seed)로 반복 학습하고, Grad-CAM 활성 영역의 공간적 분포, 시드 간 설명의 안정성, 국소화(localization)·충실도(faithfulness) 지표와 모델 예측 확률의 관계를 분석한다.

II. Related works

클래스 불균형이란 데이터셋 내의 클래스 간 샘플 수의 분포가 치우친 상태를 뜻한다. 불균형 환경에서 학습된 모델은 다수 클래스 위주로 예측하는 경향을 보인다[1][2]. 의료 데이터의 클래스 불균형 완화 기법은 크게 데이터 수준과 알고리즘 수준의 접근법으로 구분된다[3][4].

데이터 수준의 기법은 학습 데이터의 클래스 분포를 조정하는 방식이다. 다운샘플링(downsampling)은 다수 클래스의 샘플 일부를 무작위로 제거해 클래스 비율을 맞추는 방법이다. 구현이 단순하고 학습 속도가 빠르다는 장점이 있지만 제거된 샘플 중 유용한 정보가 손실될 수 있다는 한계점이 있다. 오버샘플링(oversampling)은 소수 클래스 샘플을 복제해 수를 늘리므로 데이터 손실이 없지만 동일 샘플이 반복 학습되므로 과적합 위험이 있다[14]. 데이터 증강(data augmentation)은 소수 클래스 이미지에 다양한 변환을 적용하여 새로운 샘플을 생성함으로써 과적합 위험을 낮추고 데이터 다양성을 확보한다[15].

알고리즘 수준의 기법은 데이터 분포를 직접 변경하기보다 손실함수에 클래스 불균형을 반영한다. 가중 교차 엔트로피(Weighted Cross Entropy, 이하 WCE)는 클래스별 가중치를 손실에 부여하여 소수 클래스를 잘못 예측할 경우 더 큰 페널티를 부과하는 방식으로, 일반적으로 클래스 빈도의 역수로 가중치를 설정한다[16]. Focal Loss는 이미 높은 확신으로 맞히는 쉬운 샘플의 손실을 자동으로 줄이고, 어려운 샘플에 집중하도록 설계되었다[17]. 이러한 기법들의 효과는 주로 Accuracy, AUROC(Area Under the ROC Curve), F1-score, Recall 등의 분류 성능 지표로 평가되어 왔다.

XAI는 딥러닝 모델의 의사결정 과정을 사람이 이해할 수 있도록 해석하는 기술을 의미하며, 크게 내재적 해석 가능성과 사후 설명 방식으로 구분된다[18]. 의료영상과 같은 복잡한 이미지 데이터를 다루는 딥러닝 기반 모델에서는 예측 결과에 대해 별도의 설명을 생성하는 접근인 사후 설명 방식이 널리 사용되어 왔으며, LIME, SHAP, Grad-CAM 등이 대표적인 방법이다[8][9][19][20]. Grad-CAM은 마지막 합성곱 층의 특징맵과 기울기를 이용하여, 모델이 특정 클래스를 예측할 때 주목한 영역을 설명맵으로 나타낸다[8].

설명의 품질을 정량적으로 평가하기 위한 지표로는 국소화 지표와 충실도 지표가 주로 사용된다. IoU(Intersection over Union)와 PG(Pointing Game)는 병변 영역과 설명맵 간의 일치도를 측정하는 국소화 지표이며, iAUC(Insertion

www.kci.go.kr

AUC)와 dAUC(Deletion AUC)는 설명맵의 중요 영역을 점진적으로 삽입·삭제하면서 모델 예측 변화를 측정하는 충실도 지표이다[21]-[23].

불균형 완화 기법은 분류 성능 외에 모델 출력의 성질도 바꿀 수 있다. 임상 예측 모델에서 다운샘플링, 오버샘플링 등의 기법은 판별력을 개선하지 못한 채 소수 클래스의 발생 확률을 과대추정하여 확률 보정을 악화시켰으며, Recall 향상 효과는 임계값 조정만으로도 유사하게 얻을 수 있었다[5]. 또한 불균형 완화가 모델의 예측 행동을 크게 바꿀 수 있어, 성능 비교만으로는 그 영향을 충분히 파악할 수 없음이 지적되기도 했다[6]. 다만 위 연구들은 임상 예측 모델이나 변수 단위 설명을 대상으로 하여, 영상 모델의 공간적 설명이 불균형 완화에 따라 어떻게 달라지는지는 다루지 않았다.

의료영상에서 설명맵 신뢰성의 한계도 보고되었다. 흉부 X-ray에서 Grad-CAM 등의 설명맵은 전용 탐지·분할 모델보다 국소화 성능이 낮았고, 동일 구조의 모델을 다시 학습하는 것만으로 맵이 달라졌다[10]. 또한 Grad-CAM의 국소화 성능은 영상의학 전문의보다 크게 낮았으며, 그 격차는 작고 형태가 복잡한 병변에서 가장 크고, 모델의 예측 확신도가 낮을수록 국소화 성능도 낮았다[11]. 평가 지표 측면에서는 iAUC와 dAUC가 설명맵 값의 순위만 반영하고, 학습 분포를 벗어난 교란 이미지에 의존한다는 한계가 지적되었다[12]. 그러나 불균형 완화처럼 학습 조건이 바뀌어 예측 확률 자체가 이동할 때, 설명의 형태와 안정성, 지표의 해석이 어떻게 달라지는지는 충분히 다루어지지 않았다.

이에 본 연구는 동일한 데이터셋과 모델 구조에서 여러 불균형 완화 기법을 반복 학습한다. 이를 통해 예측 확률의 변화가 Grad-CAM 설명의 공간적 분포와 시드 간 안정성, 국소화·충실도 지표의 해석에 미치는 영향을 함께 분석한다는 점에서 기존 연구와 구별된다.

III. Results and Analysis

1. Methodology

설명 품질을 병변 영역과의 공간적 일치도로 평가하기 위해, 본 연구에서는 병변 위치 어노테이션을 제공하는 Kaggle의 RSNA Pneumonia Detection Challenge(2018) 데이터셋을 사용했다[24]. 해당 데이터셋은 총 26,683명의 환자에 대한 흉부 정면 X-ray 이미지와 bounding box 형태의 병변 위치 어노테이션으로 구성되어 있다. 레이블은 patientID 단위로 폐렴 여부에 따라 이진 분류(비폐렴: 0, 폐렴: 1)로 구성되며, 클래스 분포는 비폐렴 20,671장, 폐렴 6,012장으로 약 3.4:1의 불균형 비율을 보인다. 변환된 PNG 이미지를 학습 80%, 검증 10%, 시험 10%로 분할했으며, 데이터 분할은 모든 반복 실험에서 동일하게 유지하였다. 데이터셋의 최종 분할 결과는 Table 1과 같다.

Table 1. Train-Validation-Test Split

Split	Negative	Positive	Total
Train	16,536	4,810	21,346
Validation	2,067	601	2,668
Test	2,068	601	2,669



Fig. 1. Model architecture

본 연구의 분류 모델은 ImageNet으로 사전 학습된 ResNet50V2를 기반으로 한 전이학습 구조를 사용했다. ResNet50V2는 의료영상 분류에서 베이스라인으로 널리 활용되며, Grad-CAM과 결합 시 효과적으로 병변 영역을 강조할 수 있다[25]. ResNet50V2 백본(backbone) 위에 평균풀링(Global Average Pooling)과 밀집층(fully connected) 분류기를 추가했으며, 백본의 마지막 합성곱 블록(conv5_x)과 이후 정규화층을 학습 가능하도록 설정하여 미세 조정(fine-tuning)을 수행하였다. 학습은 Adam(학습률 1.4×10^{-3}), 배치 크기 64, 최대 20 에폭으로 수행하였으며, 검증 AUROC를 기준으로 학습률 감소(×0.1, patience 2), 조기 종료(patience 5) 및 최종 모델 선택을 수행하였다. 각 실험 조건은 10개의 랜덤 시드로 반복 학습하였으며, 데이터 증강은 계산 비용을 고려하여 5회 반복하였다. 전체 구조는 Fig. 1과 같다.

입력 이미지는 224×224 픽셀 크기로 조정하고, 단일 채널 X-ray를 3채널로 복제한 뒤 ResNet50V2의 입력 형식에 맞게 정규화(-1 ~ +1) 하였다. 해당 전처리는 학습, 검증, 시험 세트에 모두 동일하게 적용했다.

불균형 완화 기법에 따른 분류 성능과 설명 품질의 변화를 체계적으로 추적하기 위해 데이터 수준과 알고리즘 수준의 기법을 비교하였다. 데이터 수준 기법에서는 다운샘플링과 오버샘플링을 적용하였고, 알고리즘 수준에서는 WCE와 Focal Loss를 사용하였다. 또한 불균형 완화 기법을 적용하

지 않고 데이터 증강만 수행한 조건을 대조군으로 추가하였다. 모든 불균형 완화 기법은 학습 세트에만 적용하고, 검증 및 시험 세트는 원본 분포를 유지했다. Baseline은 불균형 완화 기법과 데이터 증강을 모두 적용하지 않은 조건이며, 데이터 증강은 대조군에만 적용하였다.

다운샘플링과 오버샘플링은 Table 2와 같이 학습 데이터의 클래스 비율을 1:1로 조정하였다. 데이터 증강은 Table 3과 같이 임상적 의미를 보존하는 범위의 변환을 학습 이미지에 적용하였다.

Table 2. Downsampling, Oversampling

Class	Before	Down	Over
Negative	16,536	4,810	16,536
Positive	4,810	4,810	16,536

Table 3. Augmentation

Aug.	Parameter	Probability
Horizontal Flip	-	p = 0.5
Rotation	±15°	p = 0.5
Scaling	±10%	p = 0.5
Brightness	±20%	p = 0.5

WCE의 손실함수(WL)와 Focal Loss의 손실함수(FL)는 다음과 같이 정의하였다.

$$WL = -\sum_{c} w_c y_c \log(p_c) \quad (1)$$

$$FL(p_t) = -\alpha_t (1-p_t)^{\gamma} \log(p_t) \quad (2)$$

식(1)에서 w_c는 클래스 c에 대한 가중치, y_c는 정답 레이블, p_c는 모델 예측 확률이다. 본 연구에서는 다수 클래스의 가중치를 1로 정규화하여 $w_0 = 1.0, w_1 = 3.4378$로 설정했다.

식(2)에서 α_t는 클래스 가중치, $(1-p_t)^{\gamma}$는 쉬운 샘플의 손실을 줄여 어려운 샘플에 집중하도록 하는 조절 파라미터이다. 본 연구에서는 $\alpha = 0.75$, $\gamma = 1.0$으로 설정했다.

기법 간 설명 품질의 변화를 비교하기 위해 모든 실험에서 해석이 직관적이고 구현이 표준화된 Grad-CAM을 사용했다. ResNet50V2의 고수준 공간 정보를 보존하는 최상위 합성곱층인 conv5_block3_out 출력을 기준으로 폐렴 클래스에 대한 Grad-CAM을 계산하였으며, 산출된 활성화 맵은 [0, 1] 범위로 정규화한 뒤 Bilinear 보간법을 통해 224×224 크기로 변환하여 설명 품질 평가에 활용하였다. 사용된 Grad-CAM의 파이프라인은 Fig. 2와 같다.

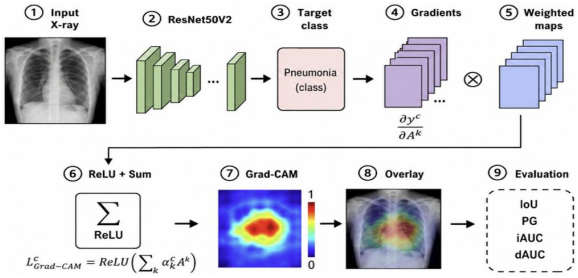


Fig. 2. Grad-CAM Pipeline

Table 4. XAI metrics

Metrics	Expressions
IoU	$IoU = \frac{\lvert M_{CAM} \cap M_{GT} \rvert}{\lvert M_{CAM} \cup M_{GT} \rvert}$
PG	$PG = \begin{cases} 1, \text{if } (x_{\max}, y_{\max}) \in B_{GT} \\ 0, otherwise \end{cases}$
iAUC	$IAUC \approx \sum_{t=1}^{T} \frac{s_{t-1} + s_t}{2} (r_t - r_{t-1})$
dAUC	$DAUC \approx \sum_{t=1}^{T} \frac{s_{t-1} + s_t}{2} (r_t - r_{t-1})$

모델의 분류 성능은 AUROC, AUPRC, Precision, Recall, F1-score로 평가하였으며, 분류 임계값은 0.5를 기준으로 산출하였다.

설명 품질은 동일한 폐렴 양성 시험 영상 601장을 대상으로 국소화(IoU, PG)와 충실도(iAUC, dAUC)를 평가하였다. IoU는 CAM을 임계값 0.5로 이진화한 영역과 병변 영역의 겹침 정도로, PG는 최대 활성화 위치의 병변 포함 여부로 산출하였으며, IoU는 임계값 0.2~0.7 범위에서도 추가로 확인하였다.

iAUC와 dAUC는 CAM 값이 높은 순서대로 픽셀을 순차적으로 복원하거나 제거할 때의 예측 확률 변화를 측정하였다. 삽입·삭제는 픽셀의 5%씩 20단계로 진행하였으며, Gaussian Blur(51×51)를 공통 기준 영상으로 사용하였다. 각 지표의 수식은 Table 4와 같으며, 여기서 T는 전체 단계 수, r_t는 복원 또는 제거된 픽셀의 비율, s_t는 해당 단계에서의 모델 예측 확률을 나타낸다.

클래스 불균형 완화 기법의 영향을 분석하기 위해 Baseline과 4가지 불균형 완화 기법의 분류 및 XAI 지표를 비교하였다. 반복 학습 결과를 평균과 표준편차로 제시하고, 시드 단위 Welch t-검정과 Holm 다중비교 보정을 적용하였다. 추가로 CAM 활성 면적(정규화 CAM ≥ 0.5인 픽셀 비율), 시드 간 CAM 일치도(시드 쌍별 CAM 상관

www.kci.go.kr

계수의 평균) 및 양성 예측 확률(p_0)을 분석하여 설명의 공간적 특성과 재현성, 평가 지표와의 관계를 검토하였다.

2. Experimental Results

2.1 Classification and XAI Evaluation Results

본 절에서는 클래스 불균형 완화 기법에 따른 분류 성능과 Grad-CAM 설명 품질의 변화를 정량적으로 비교하였다. 모든 결과는 반복 학습의 평균과 표준편차로 제시하였다.

Table 5와 같이 Baseline의 AUROC는 0.878이었으며, 네 가지 불균형 완화 기법은 0.870~0.875로 유사한 수준을 보였다. 임계값 0.5에서 불균형 완화 기법은 Recall을 0.574에서 0.802~0.829로 높였으나, Precision은 0.671에서 0.490~0.512로 낮아졌고, F1-score는 0.609~0.624로 Baseline과 유의한 차이가 없었다.

이는 불균형 완화 기법이 모델의 판별력을 개선하기보다 양성 판정 경향을 강화하였음을 보여준다. 반면 불균형 완화 기법을 적용하지 않은 데이터 증강은 Recall과 Precision이 Baseline과 유사하였으며, AUROC는 0.882로 소폭 높았다.

Table 6은 동일한 폐렴 양성 시험 영상 601장에 대한 Grad-CAM 설명 품질 평가 결과를 나타낸다. 국소화 지표에서 PG는 모든 불균형 완화 기법에서 증가하는 경향을 보였으나(0.338 → 0.375~0.444), IoU는 Focal Loss를 제외하고 감소하였으며 오버샘플링에서 0.134로 가장 낮았다. 다만 Holm 다중비교 보정 후 국소화 지표에서 유의한 개선은 확인되지 않았다. IoU 임계값을 0.2~0.7로 변경한 민감도 분석에서도 기법 간 경향은 유지되었으며, 오버샘플링의 IoU 저하는 임계값 0.6과 0.7에서 유의하였다(Holm 보정 $p = 0.024$, 0.014). PG의 증가가 IoU의 개선으로 일관되게 이어지지는 않았다.

충실도 지표에서는 iAUC가 Baseline의 0.532에서 0.662~0.796으로 증가하였으나, 낮을수록 유리한 dAUC 역시 0.367에서 0.553~0.675로 상승하였다(Holm 보정 $p < 0.001$). 이처럼 두 충실도 지표가 상반된 평가 방향을 보였으며, 불균형 완화 기법을 적용하지 않은 데이터 증강에서는 두 지표 모두 Baseline과 유사하였다. 따라서 iAUC의 증가만으로 설명 품질의 개선을 판단하기는 어렵다.

2.2 CAM Characteristics, Stability, and Faithfulness Analysis

앞선 결과에서 설명 품질 지표의 변화 방향이 일관되지 않음에 따라, Grad-CAM 활성 영역의 공간적 특성과 반복 학습에 따른 안정성을 추가로 분석하였다.

Table 7과 같이 Baseline의 CAM 활성 면적은 전체 영상의 0.145였으나, 불균형 완화 기법 적용 후 0.079~0.107로 감소하였다(Fig. 3). 특히 오버샘플링은 0.079로 가장 작은 활성 면적을 보였으며, IoU 역시 가장 낮았다. 이는 활성 영역이 특정 위치에 집중되면서 PG가 증가하더라도, 실제 병변 영역을 충분히 포함하지 못해 IoU가 감소할 수 있음을 시사한다. 따라서 CAM 활성 면적의 축소를 설명 품질의 향상으로 직접 해석하기는 어렵다.

Table 5. Classification Performance by Imbalance Mitigation Methods

Method	AUROC	AUPRC	Precision	Recall	F1-score
Baseline	0.878±0.003	0.693±0.006	0.671±0.021	0.574±0.051	0.617±0.022
WCE	0.875±0.005	0.688±0.011	0.510±0.031	0.809±0.051	0.624±0.014
Focal Loss	0.874±0.005	0.689±0.010	0.512±0.032	0.802±0.052	0.623±0.010
Downsampling	0.870±0.004	0.669±0.011	0.490±0.065	0.829±0.083	0.609±0.034
Oversampling	0.871±0.005	0.678±0.010	0.511±0.047	0.802±0.054	0.621±0.026
Augmentation	0.882±0.002	0.699±0.005	0.700±0.043	0.512±0.092	0.585±0.052

Table 6. Grad-CAM Evaluation Metrics by Imbalance Mitigation Methods

Method	IoU	PG	iAUC	dAUC
Baseline	0.184 ± 0.039	0.338 ± 0.125	0.532 ± 0.038	0.367 ± 0.034
WCE	0.162 ± 0.039	0.404 ± 0.097	0.735 ± 0.037	0.608 ± 0.045
Focal Loss	0.195 ± 0.048	0.444 ± 0.090	0.662 ± 0.036	0.553 ± 0.034
Downsampling	0.163 ± 0.035	0.375 ± 0.113	0.796 ± 0.064	0.675 ± 0.086
Oversampling	0.134 ± 0.041	0.390 ± 0.113	0.715 ± 0.036	0.572 ± 0.067
Augmentation	0.154 ± 0.031	0.312 ± 0.061	0.517 ± 0.030	0.351 ± 0.029

www.kci.go.kr

Table 7. CAM Activation Area and Seed Agreement

Method	CAM Area	Seed Agreement
Baseline	0.145±0.025	0.657
WCE	0.088±0.017	0.681
Focal Loss	0.103±0.027	0.660
Downsampling	0.107±0.027	0.564
Oversampling	0.079±0.013	0.509
Augmentation	0.147±0.035	0.661

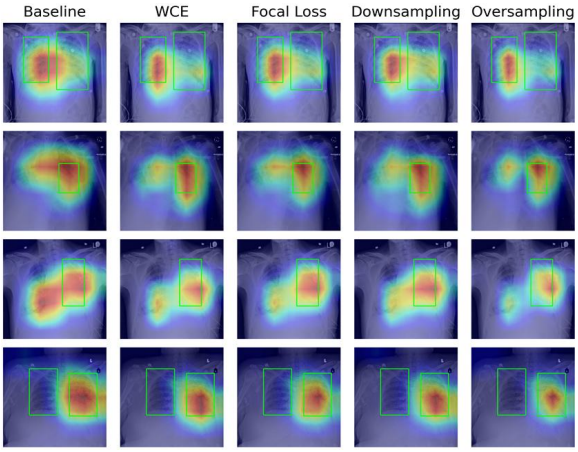


Fig. 3. Grad-CAM by imbalance mitigation method (Seed-Averaged)

반복 학습에 따른 설명의 재현성을 확인하기 위해 동일한 시험 영상에 대한 시드 간 CAM 일치도를 비교하였다. Baseline의 일치도는 0.657이었으며, WCE와 Focal Loss는 각각 0.681과 0.660으로 유사한 수준을 보였다. 반면 다운샘플링과 오버샘플링은 각각 0.564와 0.509로 낮았다(Fig. 4). 이는 두 기법에서 학습 시드에 따른 설명의 공간적 변동성이 상대적으로 크게 나타났음을 보여준다. 학습 시드마다 무작위 변환이 적용되는 데이터 증강의 일치도는 0.661로 Baseline과 유사하였다.

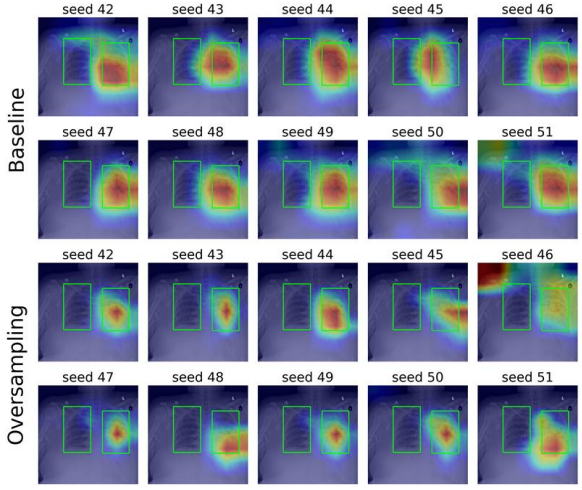


Fig. 4. Grad-CAM across training seeds (Baseline vs. Oversampling)

충실도 지표의 상반된 변화를 해석하기 위해 양성 시험 영상의 평균 예측 확률(p_0)과 iAUC의 관계를 분석하였다. 불균형 완화 기법은 p_0를 0.543에서 0.673~0.777로 높였으며(Holm 보정 $p < 0.001$), Fig. 5와 같이 iAUC는 기법과 관계없이 p_0와 강한 양의 상관관계를 보였다(Spearman $\rho = 0.93$). 동일 기법 내에서도 p_0가 높은 시드일수록 iAUC가 높았으며, 데이터 증강 단독 조건에서는 p_0와 iAUC가 모두 Baseline과 유사하였다. 따라서 본 실험 조건에서 iAUC의 증가는 모델의 예측 확률 수준과 밀접하게 연관되어 있으며, 이를 설명 품질의 개선으로 직접 해석하기는 어렵다.

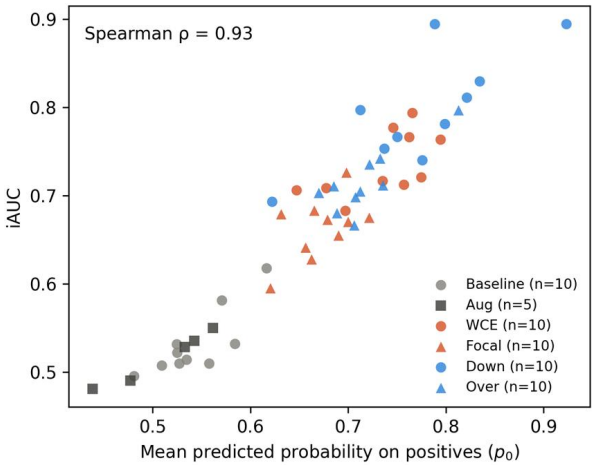


Fig. 5. Relationship between p_0 and iAUC

2.3 Subgroup Analysis

본 절에서는 병변의 크기와 모델의 분류 결과에 따라 Grad-CAM의 국소화 성능이 어떻게 달라지는지 추가로 분석하였다. 이를 위해 폐렴 양성 시험 영상을 영상 대비 병변 면적의 삼분위에 따라 Small, Medium, Large의 세 그룹(각 약 200장)으로 구분하고, 각 그룹의 IoU와 PG를 비교하였다.

Table 8. Grad-CAM IoU by Lesion Size

Method	Small	Medium	Large
Baseline	0.098	0.202	0.251
WCE	0.110	0.184	0.191
Focal Loss	0.126	0.222	0.237
Downsampling	0.097	0.183	0.209
Oversampling	0.092	0.152	0.159
Augmentation	0.080	0.166	0.214

www.kci.go.kr

Table 9. Grad-CAM Pointing Game by Lesion Size

Method	Small	Medium	Large
Baseline	0.139	0.336	0.537
WCE	0.175	0.414	0.623
Focal Loss	0.204	0.471	0.655
Downsampling	0.148	0.378	0.598
Oversampling	0.160	0.385	0.623
Augmentation	0.106	0.306	0.522

Table 8과 Table 9에서 모든 기법은 병변 크기가 증가할수록 IoU와 PG가 높아지는 경향을 보였다. 작은 병변의 IoU는 0.080~0.126으로 전반적으로 낮았으며, PG 역시 0.106~0.204로 큰 병변에 비해 낮게 나타났다. 이는 불균형 완화 기법의 종류와 관계없이 작은 병변에서 Grad-CAM의 국소화 성능이 제한적임을 보여준다.

추가로 모델이 폐렴 양성 영상을 정분류한 경우(TP)와 오분류한 경우(FN)를 구분하여 설명 품질을 비교하였다.

Table 10. Grad-CAM Localization Performance by Prediction Outcome

Method	TP		FN	
	IoU	PG	IoU	PG
Baseline	0.244	0.464	0.103	0.174
WCE	0.184	0.468	0.064	0.135
Focal Loss	0.224	0.512	0.078	0.165
Downsampling	0.185	0.430	0.049	0.090
Oversampling	0.157	0.462	0.041	0.100
Augmentation	0.216	0.466	0.083	0.140

Table 10과 같이 모든 기법에서 FN의 IoU와 PG가 TP보다 낮았다. Baseline의 경우 IoU는 TP 0.244에서 FN 0.103으로, PG는 0.464에서 0.174로 감소하였다. 이러한 경향은 다른 기법에서도 동일하게 관찰되었으며, 모델이 폐렴을 탐지하지 못한 사례에서 Grad-CAM 역시 병변 영역을 충분히 반영하지 못할 가능성을 시사한다.

이러한 결과는 의료영상 분류 모델의 설명을 평가할 때 전체 평균 성능뿐 아니라 병변의 크기와 예측 결과에 따른 국소화 성능의 차이도 고려할 필요가 있음을 보여준다. 다만 기법별 Recall 차이로 인해 FN에 해당하는 영상 수와 구성이 달라지므로, 기법 간 FN 지표의 직접적인 비교에는 주의가 필요하다.

IV. Conclusions

본 연구에서는 클래스 불균형 완화 기법이 의료영상 분류 성능과 Grad-CAM 기반 설명 품질에 미치는 영향을 분석하였다. 이를 위해 RSNA 폐렴 탐지 데이터셋과 ResNet50V2 모델을 활용하여 데이터 및 알고리즘 수준의 불균형 완화 기법을 반복 학습하고, 분류 성능뿐 아니라 Grad-CAM의 국소화, 충실도, 공간적 특성 및 안정성을 종합적으로 비교하였다.

실험 결과, 불균형 완화 기법은 Recall을 높였으나 Precision은 감소하였으며, AUROC와 F1-score에서는 뚜렷한 개선이 확인되지 않았다. 설명 품질 측면에서도 PG는 증가하는 경향을 보였으나 IoU는 기법에 따라 상이한 변화를 나타냈다. 이는 불균형 완화에 따른 분류 성능의 변화가 반드시 Grad-CAM 설명 품질의 개선으로 이어지는 것은 아님을 보여준다.

Grad-CAM의 공간적 특성을 분석한 결과, 불균형 완화 기법 적용 후 CAM 활성 면적은 Baseline 대비 약 26~46% 감소하였다. 활성 영역이 특정 위치에 집중되더라도 실제 병변과의 공간적 일치도가 반드시 높아지는 것은 아니었으며, 이러한 결과는 CAM의 시각적 집중만으로 설명 품질을 판단하기 어렵다는 점을 시사한다. 또한 반복 학습에 따른 시드 간 CAM 일치도는 다운샘플링과 오버샘플링에서 상대적으로 낮게 나타나, 불균형 완화 전략에 따라 설명의 공간적 안정성에도 차이가 발생할 수 있음을 확인하였다.

충실도 평가에서는 iAUC와 dAUC가 함께 증가하는 상반된 양상이 관찰되었다. 특히 iAUC는 모델의 양성 예측 확률과 강한 양의 상관관계(ρ = 0.93)를 보여, 설명 품질 지표의 변화가 예측 확률 수준과 밀접하게 연관되어 있음을 확인하였다. 따라서 충실도 지표의 상승을 실제 설명 품질의 향상으로 직접 해석하기보다는, 모델 출력의 변화와 다른 설명 품질 지표를 함께 고려할 필요가 있다.

추가 분석에서는 작은 병변과 모델이 폐렴을 탐지하지 못한 사례에서 Grad-CAM의 국소화 성능이 상대적으로 낮게 나타났다. 이는 의료 AI의 설명을 판독 보조에 활용할 때 개별 영상에서 제시되는 설명의 정확성과 안정성에 대한 검토가 중요함을 시사한다. 따라서 불균형 완화 기법을 평가할 때는 분류 성능과 함께 국소화·충실도 지표 및 반복 학습에 따른 설명의 안정성을 검토할 필요가 있다.

다만 본 연구는 단일 데이터셋과 모델 구조 및 Grad-CAM에 한정되며, bounding box 기반 병변 평가와 Grad-CAM의 낮은 공간 해상도에도 한계가 있다. 향후 다양한 의료영상 데이터셋과 모델 구조 및 설명 기법을 대상으로 분석을 확장하여, 불균형 완화 전략에 따른 설명 품질의 변화와 안정성을 추가로 검증할 필요가 있다.

www.kci.go.kr

ACKNOWLEDGEMENT

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (RS-2025-02309552).

REFERENCES

[1] J.M. Johnson, and T.M. Khoshgoftaar, “Survey on deep learning with class imbalance,” Journal of Big Data, Vol. 6, No. 1, pp. 1-54, March 2019. DOI: 10.1186/s40537-019-0192-5

[2] H. He, and E.A. Garcia, “Learning from imbalanced data,” IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 9, pp. 1263-1284, September 2009. DOI: 10.1109/TKDE.2008.239

[3] M. Buda, A. Maki, and M.A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” Neural Networks, Vol. 106, pp. 249-259, October 2018. DOI: 10.1016/j.neunet.2018.07.011

[4] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, “Handling imbalanced medical datasets: review of a decade of research,” Artificial Intelligence Review, Vol. 57, Art. no. 273, September 2024. DOI: 10.1007/s10462-024-10884-2

[5] R. van den Goorbergh, M. van Smeden, D. Timmerman, and B. Van Calster, “The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression,” Journal of the American Medical Informatics Association, Vol. 29, No. 9, pp. 1525-1534, August 2022. DOI: 10.1093/jamia/ocac093

[6] A. Stando, M. Cavus, and P. Biecek, “The effect of balancing methods on model behavior in imbalanced classification problems,” Proc. Fifth Int. Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA), PMLR Vol. 241, pp. 16-30, Turin, Italy, September 2023.

[7] A. Adadi, and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” IEEE Access, Vol. 6, pp. 52138-52160, September 2018. DOI: 10.1109/ACCESS.2018.2870052

[8] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” Proc. IEEE Int. Conf. Computer Vision (ICCV), pp. 618-626, Venice, Italy, October 2017. DOI: 10.1109/ICCV.2017.74

[9] R.-K. Sheu, and M.S. Pardeshi, “A survey on medical explainable AI (XAI): recent progress, explainability approach, human interaction and scoring system,” Sensors, Vol. 22, No. 20, Art. no. 8068, October 2022. DOI: 10.3390/s22208068

[10] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, et al., “Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging,” Radiology: Artificial Intelligence, Vol. 3, No. 6, Art. no. e200267, November 2021. DOI: 10.1148/ryai.2021200267

[11] A. Saporta, X. Gui, A. Agrawal, et al., “Benchmarking saliency methods for chest X-ray interpretation,” Nature Machine Intelligence, Vol. 4, No. 10, pp. 867-878, October 2022. DOI: 10.1038/s42256-022-00536-x

[12] T. Gomez, T. Fréour, and H. Mouchère, “Metrics for saliency map evaluation of deep learning explanation methods,” Proc. Int. Conf. Pattern Recognition and Artificial Intelligence (ICPRAI), Lecture Notes in Computer Science, Vol. 13363, pp. 84-95, June 2022. DOI: 10.1007/978-3-031-09037-0_8

[13] J. Feng, R.V. Phillips, I. Malenica, A. Bishara, A.E. Hubbard, L.A. Celi, and R. Pirracchio, “Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare,” npj Digital Medicine, Vol. 5, Art. no. 66, May 2022. DOI: 10.1038/s41746-022-00611-y

[14] G.E.A.P.A. Batista, R.C. Prati, and M.C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” ACM SIGKDD Explorations Newsletter, Vol. 6, No. 1, pp. 20-29, June 2004. DOI: 10.1145/1007730.1007735

[15] C. Shorten, and T.M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” Journal of Big Data, Vol. 6, Art. no. 60, July 2019. DOI: 10.1186/s40537-019-0197-0

[16] Y. Cui, M. Jia, T.Y. Lin, Y. Song, and S. Belongie, “Class-balanced loss based on effective number of samples,” Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 9268-9277, Long Beach, CA, USA, June 2019. DOI: 10.1109/CVPR.2019.00949

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” Proc. IEEE Int. Conf. Computer Vision (ICCV), pp. 2980-2988, Venice, Italy, October 2017. DOI: 10.1109/ICCV.2017.324

[18] F. Doshi-Velez, and B. Kim, “Towards a rigorous science of interpretable machine learning,” arXiv preprint arXiv:1702.08608, 2017.

[19] M.T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), pp. 1135-1144, San Francisco, CA, USA, August 2016. DOI: 10.1145/2939672.2939778

[20] S.M. Lundberg, and S.I. Lee, “A unified approach to interpreting model predictions,” Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 4765-4774, Long Beach, CA, USA, December 2017.

www.kci.go.kr

[21] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.R. Müller, "Evaluating the visualization of what a deep neural network has learned," IEEE Transactions on Neural Networks and Learning Systems, Vol. 28, No. 11, pp. 2660-2673, November 2017. DOI: 10.1109/TNNLS.2016.2599820

[22] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," Proc. British Machine Vision Conf. (BMVC), Paper 151, Newcastle, UK, September 2018.

[23] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," Proc. European Conf. Computer Vision (ECCV), pp. 543-559, Amsterdam, Netherlands, October 2016. DOI: 10.1007/978-3-319-46493-0_33

[24] G. Shih et al., "Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia," Radiology: Artificial Intelligence, Vol. 1, No. 1, e180041, January 2019. DOI: 10.1148/ryai.2019180041

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," Proc. European Conf. Computer Vision (ECCV), pp. 630-645, Amsterdam, Netherlands, October 2016. DOI: 10.1007/978-3-319-46493-0_38

Authors



Kwangho Ko received the B.S., M.S., and Ph.D. degrees in Automotive Engineering from Seoul National University, South Korea. Dr. Ko joined the faculty of the School of Applied AI Convergence at Sungkyunkwan University in 2024. He is currently a Professor. His research interests include deep learning, machine learning, and autonomous driving technology.



Sooyeon Bong is currently pursuing the B.S. degree in Industrial Artificial Intelligence at Sungkyunkwan University, South Korea. Her research interests include computer vision, explainable artificial intelligence, and medical image analysis.



Hyunjun Kim is currently pursuing the B.S. degree in Global Biomedical Engineering at Sungkyunkwan University, South Korea. His research interests include brain-inspired artificial intelligence, neurological disease diagnosis and prediction, and neuroimaging data mining.



Dongheon Lee is currently pursuing the B.S. degree in Industrial Artificial Intelligence at Sungkyunkwan University, South Korea. His research interests include computer vision, industrial defect inspection, and synthetic data generation. His research focuses on semiconductor wafer map defect analysis and vision-based quality inspection in manufacturing.

www.kci.go.kr