

# 다양한 의료영상 도메인에서 ConvNeXtV2의 분류 성능 평가 및 기존 딥러닝 모델과의 비교

김민서\*, 고승완\*\*, 김형국\*\*\*

## Evaluation of ConvNeXtV2 Classification Performance Across Diverse Medical Imaging Domains Compared with Existing Deep Learning Models

Min-Seo Kim\*, Seung-Wan Ko\*\*, and HyounG-Gook Kim\*\*\*

### 요 약

본 연구에서는 Kvasir-v2, HAM10000 및 ChestXray14 데이터셋을 대상으로 ConvNeXtV2-B의 의료영상 분류 성능을 CNN 기반 EfficientNetV2-M 및 트랜스포머 기반 ViT-B와 비교하였다. 실험 결과, ConvNeXtV2-B는 Kvasir-v2, HAM10000 및 ChestXray14의 주요 평가지표에서 모두 가장 높은 평균 성능을 나타냈다. 이를 통해 ConvNeXtV2-B가 서로 다른 영상 특성과 라벨 구조를 갖는 의료영상 환경에서 전반적으로 안정적인 분류 성능을 제공함을 확인하였다. 또한 혼동행렬, Grad-CAM, 클래스별 F1 점수 및 연산 효율성을 분석하여 모델별 오분류 양상과 판단 근거, 계산 비용의 차이를 종합적으로 비교하였다. 이러한 결과는 ConvNeXtV2-B가 다양한 의료영상 분류 환경에 적용할 수 있는 경쟁력 있는 현대적 CNN 모델임을 보여준다.

### Abstract

This study compared ConvNeXtV2-B with CNN-based EfficientNetV2-M and Transformer-based ViT-B on the Kvasir-v2, HAM10000, and ChestXray14 datasets. ConvNeXtV2-B achieved the highest average performance across the major metrics on all three datasets, demonstrating stable performance across diverse image characteristics and label structures. Confusion matrices, Grad-CAM, class-wise F1 scores, and computational efficiency were also analyzed to compare misclassification patterns, decision evidence, and computational costs. These results demonstrate that ConvNeXtV2-B is a competitive modern CNN for various medical image classification tasks.

### Keywords

ConvNeXtV2, medical image classification, deep learning, convolutional neural network, performance comparison

\* 광운대학교 전자융합공학과 학사과정  
- ORCID: <https://orcid.org/0009-0008-5473-9392>  
\*\* 광운대학교 동북아문화산업학부 학사과정  
- ORCID: <https://orcid.org/0009-0009-1658-0664>  
\*\*\* 광운대학교 전자융합공학과 교수(교신저자)  
- ORCID: <https://orcid.org/0000-0002-1518-4100>

• Received: Jun. 26, 2026, Revised: Aug. 10, 2026, Accepted: Aug. 13, 2026  
• Corresponding Author: HyounG-Gook Kim  
Dept. of Electronic Convergence Engineering, Kwangwoon University, 20  
Gwangun-ro, Nowon-gu, Seoul 01897, Korea  
Tel.: +82-2-940-5574, Email: [hkim@kw.ac.kr](mailto:hkim@kw.ac.kr)

## I. 서 론

의료영상 분류는 질환의 조기 발견과 진단 보조를 위한 중요한 인공지능 응용 분야이다. 내시경 영상, 피부 병변 영상, 흉부 X선 영상 등은 의료진의 판독 과정에서 주요 판단 근거로 활용되며, 딥러닝 기반 분류 모델은 이러한 영상으로부터 질환 또는 병변의 특징을 자동으로 학습하여 진단 보조의 효율성을 높일 수 있다. 특히 의료영상 데이터가 지속적으로 축적되는 환경에서 자동 분류 기술은 의료진의 판독 부담을 줄이고 일관된 보조 판단을 제공할 수 있다는 점에서 활용 가능성이 크다.

그러나 의료영상은 일반 자연영상과 달리 분류에 필요한 특징이 복잡적이다. 병변의 경계, 색상 변화, 질감, 형태와 같은 국소적 특징이 중요한 단서가 되는 동시에, 병변의 위치, 주변 조직과의 관계, 영상 전체의 구조적 변화와 같은 문맥 정보도 함께 고려되어야 한다. 또한 의료영상 데이터는 클래스 간 시각적 유사성, 클래스 불균형, 다중 라벨 구조와 같은 문제를 포함할 수 있으므로, 단일 데이터셋의 평균 성능만으로 모델의 적용 가능성을 판단하기 어렵다.

기존 의료영상 분류 연구에서는 합성곱 신경망(CNN, Convolutional Neural Network) 기반 모델과 트랜스포머 기반 모델이 주로 활용되어 왔다. CNN 기반 모델은 병변의 색상, 질감, 경계와 같은 국소적 시각 특징을 계층적으로 추출하는 데 강점을 가지며, 트랜스포머 기반 모델은 영상 내 넓은 영역의 관계를 학습하여 전역적 문맥 정보를 반영할 수 있다는 장점을 가진다. 그러나 의료영상 분류에서는 국소 특징과 전역 문맥 정보가 동시에 요구되는 경우가 많기 때문에, 특정 구조의 장점만으로 다양한 의료영상 도메인에 안정적으로 대응하기에는 한계가 있을 수 있다. 따라서 서로 다른 구조적 특성을 가진 모델들을 동일한 조건에서 비교하고, 데이터셋 특성과 라벨 구조에 따라 각 모델이 보이는 분류 특성을 분석할 필요가 있다.

최근 제안된 ConvNeXtV2[1]는 CNN의 계층적 특징 추출 구조를 유지하면서도 트랜스포머의 설계 요소를 반영한 모델이다. 이러한 구조는 CNN 기반 모델의 국소 특징 학습 능력을 유지하면서도 보다

넓은 영역의 특징을 반영할 수 있다는 점에서 의료영상 분류 문제에 적용 가능성을 가진다. 그러나 ConvNeXtV2를 서로 다른 의료영상 도메인에 적용하고, 기존 CNN 기반 모델 및 트랜스포머 기반 모델과 비교하여 성능적·구조적 특성을 분석한 연구는 아직 충분하지 않다.

이에 본 연구에서는 ConvNeXtV2-B가 다양한 의료영상 분류 환경에서 어떠한 성능 특성을 보이는지 분석하는 것을 목적으로 한다. 이를 위해 내시경 영상 데이터셋인 Kvasir-v2[2], 피부 병변 영상 데이터셋인 HAM10000[3], 흉부 X선 영상 데이터셋인 ChestXray14[4]를 사용하였다. 비교 모델로는 CNN 기반 모델인 EfficientNetV2-M[5]과 트랜스포머 기반 모델인 ViT-B[6]를 선정하였다. 세 모델은 동일한 데이터 분할, 전처리 및 학습 조건에서 평가하였으며, Accuracy, F1 점수, AUC(Area Under the Curve)와 같은 정량적 성능뿐만 아니라 혼동행렬, Grad-CAM(Gradient-weighted Class Activation Mapping), 클래스별 F1 점수 및 연산 효율성을 함께 분석하였다.

본 논문의 기여는 다음과 같다. 첫째, 서로 다른 영상 특성과 라벨 구조를 갖는 세 가지 의료영상 데이터셋에서 ConvNeXtV2-B, EfficientNetV2-M, ViT-B의 분류 성능을 동일 조건에서 비교하였다. 둘째, CNN 기반 모델, 트랜스포머 기반 모델, 현대적 CNN 기반 모델이 의료영상 도메인별로 보이는 성능 차이를 분석하였다. 셋째, 혼동행렬과 Grad-CAM을 활용하여 모델별 오분류 양상과 판단 근거의 차이를 정성적으로 확인하였다. 넷째, 클래스별 성능과 연산 효율성을 함께 분석하여 ConvNeXtV2-B의 의료영상 분류 적용 가능성을 성능 안정성과 계산 비용 측면에서 검토하였다.

본 논문의 구성은 다음과 같다. 제2장에서는 의료영상 분류에 적용된 관련 연구를 정리하고, 제3장에서는 실험 설계와 실험 결과를 제시한다. 제4장에서는 결론과 향후 연구 방향을 제시한다.

## II. 관련 연구

의료영상 분류에서는 제한된 데이터 환경에서도

높은 분류 성능을 확보하기 위해 사전학습 모델을 활용한 전이학습 기반 연구가 활발히 수행되어 왔다. 특히 CNN 계열 모델은 영상의 국소적 패턴을 계층적으로 추출하는 데 강점을 가지므로 흉부 X선, 피부 병변 영상, 내시경 영상 등 다양한 의료영상 분류 문제에 적용되어 왔다. 대표적으로 CheXNet은 DenseNet-121 구조를 기반으로 ChestXray14 데이터셋에서 폐렴 검출 및 흉부 질환 분류를 수행하였으며, CNN 기반 모델이 흉부 X선 영상의 진단 보조에 활용될 수 있음을 보였다[7]. 또한 HAM10000 피부 병변 분류 연구에서는 ResNet 및 Inception 계열 모델을 활용한 전이학습이 병변의 색상, 질감, 경계와 같은 시각적 특징을 학습하는데 효과적임을 확인하였다[8]. 이러한 연구들은 CNN 기반 모델이 의료영상의 국소 특징 추출에 유용함을 보여주지만, 클래스 불균형, 질환 간 동시 발생, 유사 병변 간 시각적 혼동이 존재하는 경우 평균 성능만으로 모델의 분류 특성을 충분히 설명하기 어렵다는 한계도 함께 나타낸다.

EfficientNet 계열 모델은 CNN 기반 구조 중에서도 모델 크기와 연산량을 고려하면서 분류 성능을 확보하기 위해 제안된 구조로, 의료영상 분류 연구에서도 활용되어 왔다. EfficientNetV2는 MBConv 및 Fused-MBConv 구조를 통해 특징 추출 과정의 효율성을 높인 모델이며, 다중 흉부 질환 및 코로나19 분류, 자기공명영상 기반 뇌종양 분류 등 다양한 의료영상 분류 문제에 적용되었다[9][10]. 이러한 연구들은 EfficientNetV2 계열이 비교적 효율적인 CNN 구조로 의료영상의 국소적 특징을 학습할 수 있음을 보여준다. 다만 의료영상에서는 병변 주변부, 해부학적 위치, 영상 전체의 구조적 변화가 함께 요구되는 경우가 많기 때문에, 효율적인 국소 특징 추출만으로는 복합적인 의료영상 특징을 충분히 표현하는 데 제한이 있을 수 있다.

트랜스포머 기반 모델은 영상 내 넓은 영역의 관계를 학습할 수 있다는 점에서 의료영상 분류에 적용되고 있다. ViT는 입력 영상을 패치 단위로 분할하고 셀프 어텐션을 통해 패치 간 관계를 학습하는 구조로, 흉부 X선, Kvasir, Kvasir-Capsule 및 다양한 2차원 생체의학 영상 분류 연구에 활용되었다

[11][12]. 이러한 연구들은 ViT가 의료영상에서 전역적 문맥 정보와 넓은 영역의 구조적 관계를 반영할 수 있음을 보여준다. 그러나 ViT는 충분한 학습 데이터와 계산 자원을 요구하며, 패치 단위 처리 과정에서 작은 병변의 경계, 미세한 질감 차이, 국소적 위치 정보가 약화될 가능성이 있다.

최근에는 CNN의 계층적 특징 추출 구조를 유지하면서 트랜스포머 계열 모델의 설계 방향을 일부 반영한 현대적 CNN 구조가 제안되고 있다. ConvNeXt는 이러한 흐름에서 제안된 모델로, CNN 기반 구조의 장점을 유지하면서도 더 넓은 영역의 특징을 반영할 수 있도록 설계되었다[13]. 이후 ConvNeXtV2는 ConvNeXt 구조를 확장하여 특징 표현의 다양성과 학습 안정성을 개선한 모델이다. 이러한 구조적 특성은 의료영상에서 요구되는 국소 특징과 문맥 정보를 함께 반영할 가능성을 가진다. 그러나 기존 연구들은 특정 모델이나 단일 의료영상 도메인에서의 성능 검증에 집중한 경우가 많아, 서로 다른 구조를 가진 모델들이 다양한 의료영상 도메인에서 어떠한 분류 특성을 보이는지 비교하는 데에는 한계가 있다.

### III. 실험 설계 및 성능 비교 분석

#### 3.1 실험 설계

본 연구에서는 ConvNeXtV2-B의 의료영상 분류 성능을 기존 딥러닝 모델과 비교 분석하기 위해 EfficientNetV2-M과 ViT-B를 비교 모델로 선정하였다. 두 모델은 각각 CNN 기반 모델과 트랜스포머 기반 모델의 대표적인 비교 기준으로 활용될 수 있으며, ConvNeXtV2-B가 기존 구조들과 비교하여 어떠한 성능적·구조적 특성을 보이는지 분석하는 데 적합하다.

EfficientNetV2-M은 성능과 연산 효율성의 균형을 고려한 CNN 기반 모델로, 의료영상 분류에서 효율 중심 합성곱 신경망 구조와 ConvNeXtV2-B를 비교하기 위한 기준으로 선정하였다. 이를 통해 ConvNeXtV2-B가 기존 효율 중심 CNN 모델과 비교하여 국소 특징 학습, 성능 안정성, 연산 효율성 측

면에서 어떠한 차이를 보이는지 확인하고자 하였다.

ViT-B는 셀프 어텐션을 기반으로 영상 내 전역적 관계를 학습하는 트랜스포머 기반 모델로, ConvNeXtV2-B와 전역 문맥 정보 반영 특성을 비교하기 위한 기준으로 선정하였다. 이를 통해 ConvNeXtV2-B가 트랜스포머 기반 모델과 비교하여 의료영상의 국소 특징과 문맥 정보를 어떻게 반영하는지 분석하고자 하였다.

ConvNeXtV2-B는 기존 CNN의 계층적 특징 추출 구조를 유지하면서 트랜스포머의 설계 요소를 반영한 현대적 CNN 모델로, 본 연구의 중심 분석 모델로 선정하였다. 큰 커널 기반의 깊이별 합성곱을 통해 넓은 공간 정보를 반영하고, 레이어 정규화와 채널 확장 구조를 통해 특징 표현 능력을 향상한다. 또한 전역 응답 정규화(GRN, Global Response Normalization)를 적용하여 채널별 특징 반응의 불균형을 완화하고 특징 표현의 다양성을 높인다. 이를 통해 ConvNeXtV2-B가 의료영상의 경계, 색상 및 질감과 같은 국소 특징과 병변 주변부를 포함한 문맥 정보를 효과적으로 학습하는지를 분석하고, EfficientNetV2-M 및 ViT-B와 비교하여 분류 성능, 성능 안정성 및 연산 효율성 측면의 특성을 확인하고자 하였다.

실험에는 서로 다른 영상 특성과 라벨 구조를 갖는 세 가지 의료영상 데이터셋을 사용하였다.

Kvasir-v2는 위장관 내시경 영상으로 구성된 단일 라벨 다중 클래스 데이터셋으로, 총 8개의 클래스를 포함한다. 해당 데이터셋은 색상, 점막 질감, 병변 경계와 같은 국소적 시각 특징이 분류에 중요한 영향을 미치며, 클래스 간 데이터 수가 비교적 균형적이다.

HAM10000은 7개의 피부 병변 클래스로 구성된 단일 라벨 다중 클래스 데이터셋으로, 클래스 간 시각적 유사성이 높고 클래스 불균형이 존재한다.

ChestXray14는 14개의 흉부 질환 라벨을 포함하는 다중 라벨 데이터셋으로, 하나의 영상에 여러 질환이 동시에 존재할 수 있으며 질환이 미세한 음영 변화나 넓은 영역의 구조적 변화로 나타나는 특징을 가진다.

모든 데이터셋은 학습, 검증, 테스트 세트로 분할하여 사용하였다. Kvasir-v2와 HAM10000은 클래스

분포가 각 분할 세트에서 유사하게 유지되도록 계층적 분할을 적용하였고, ChestXray14는 동일 환자의 영상이 서로 다른 세트에 포함되지 않도록 환자 단위로 분할하였다. 이를 통해 데이터 분할 과정에서 발생할 수 있는 평가 편향을 최소화하였다.

전처리 과정에서는 세 모델이 동일한 입력 조건에서 비교될 수 있도록 이미지를 모델 입력 크기에 맞게 변환하였다. Kvasir-v2와 HAM10000은 224×224 크기로 변환하였으며, ChestXray14는 256×256 크기로 변환한 뒤 학습 단계에서 224×224 크기로 크롭하여 사용하였다. 또한 반복 학습 과정에서 이미지 로딩으로 인한 입출력 병목을 줄이기 위해 램 캐싱을 적용하였다. 모든 이미지는 ImageNet 사전학습 모델의 입력 조건에 맞추어 평균 [0.485, 0.456, 0.406]과 표준편차 [0.229, 0.224, 0.225]를 사용하여 정규화하였다.

데이터 증강은 각 데이터셋의 촬영 방식과 분류에 중요한 영상 특징을 고려하여 다르게 적용하였다. Kvasir-v2는 내시경의 이동과 시야각에 따라 촬영 방향이 달라지고, 광원과 촬영 거리에 따라 색상 및 밝기가 변할 수 있으므로 수평 뒤집기, 회전, 밝기·대비·채도 변화를 적용하였다. HAM10000은 피부 병변의 방향이 분류 기준에 미치는 영향이 상대적으로 작고, 촬영 환경에 따라 색상과 명암이 달라질 수 있으므로 수평·수직 뒤집기, 회전 및 밝기·대비·채도 변화를 적용하였다. 반면 ChestXray14는 좌우 해부학적 위치와 영상의 음영 정보가 진단에 중요한 영향을 미치므로 뒤집기와 색상 변형을 제외하였다. 대신 촬영 과정에서 발생할 수 있는 환자의 자세, 위치 및 크기 차이를 반영하기 위해 무작위 크롭, 소각도 회전 및 아핀 변환을 적용하였다. 검증 및 테스트 세트에는 평가 조건의 일관성을 유지하기 위해 데이터 증강을 적용하지 않고 크기 조정과 정규화만 수행하였다.

모든 모델은 배치 크기 16과 최대 학습 에포크 20의 공통 조건에서 AdamW 옵티마이저를 사용하여 학습하였다. 학습률은 EfficientNetV2-M에는  $1 \times 10^{-4}$ , ViT-B와 ConvNeXtV2-B에는  $3 \times 10^{-5}$ 을 적용하였다. Kvasir-v2와 HAM10000은 단일 라벨 다중 클래스 분류 문제이므로 CrossEntropyLoss를 사용하

였고, ChestXray14는 다중 라벨 분류 문제이므로 BCEWithLogitsLoss를 사용하였다. 또한 ChestXray14의 클래스 불균형을 고려하기 위해 질환별 양성 클래스 가중치를 손실 함수에 반영하였다.

실험은 NVIDIA GeForce RTX 4060 GPU와 64 GB RAM 환경에서 수행하였다. 무작위 초기화와 데이터 순서에 따른 결과 변동을 고려하기 위해 각 모델과 데이터셋 조합에 대해 서로 다른 세 개의 무작위 시드를 적용하여 총 3회 반복 학습하였다. 최종 성능은 세 번의 실험 결과에 대한 평균과 표준편차로 제시하여 결과의 재현성과 안정성을 평가하였다.

Kvasir-v2와 HAM10000은 단일 라벨 다중 클래스 데이터셋이므로 Accuracy와 Macro F1을 평가 지표로 사용하였다. Accuracy는 전체 표본 중 예측 클래스가 실제 클래스와 일치한 비율로 식 (1)과 같이 계산하였다.

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N I(y_i = \hat{y}_i) \quad (1)$$

여기서  $N$ 은 전체 표본 수이며,  $y_i$ 와  $\hat{y}_i$ 는 각각  $i$ 번째 표본의 실제 클래스와 예측 클래스이다.  $I(\cdot)$ 는 조건의 일치 여부를 나타내는 지시 함수이다.

Macro F1은 클래스별 F1 점수를 동일한 비중으로 평균한 값으로 식 (2)와 같이 계산하였다.

$$\text{Macro F1} = \frac{1}{C} \sum_{c=1}^C \frac{2TP_c}{2TP_c + FP_c + FN_c} \quad (2)$$

여기서  $C$ 는 전체 클래스 수이며,  $TP_c$ ,  $FP_c$ ,  $FN_c$ 는 각각 클래스  $c$ 의 참 양성, 거짓 양성 및 거짓 음성 수이다.

ChestXray14는 다중 라벨 데이터셋이므로 각 질환을 독립적인 이진 분류 문제로 간주하고 Macro AUC와 Macro F1을 사용하였다. Macro AUC는 결정 임계값 변화에 따른 질환별 ROC 곡선 아래 면적을 평균하여 식 (3)과 같이 계산하였다.

$$\text{Macro AUC} = \frac{1}{C} \sum_{c=1}^C \int_0^1 \text{TPR}_c(u) du \quad (3)$$

여기서  $u$ 는 거짓 양성률을 나타내며,  $\text{TPR}_c$ 는 질환  $c$ 의 참 양성률이다. ChestXray14의 Macro F1도 식 (2)와 동일하게 질환별 F1 점수를 평균하여 산출하였다. 다만 단일 라벨 데이터셋과 달리, 각 질환에 결정 임계값을 적용하여 얻은 독립적인 이진 예측 결과를 사용하였다.

### 3.2 성능 비교 분석

표 1은 Kvasir-v2, HAM10000 및 ChestXray14에서 EfficientNetV2-M, ViT-B 및 ConvNeXtV2-B의 분류 성능을 비교한 결과이다. 전체적으로 ConvNeXtV2-B는 세 데이터셋의 모든 주요 평가지표에서 가장 높은 평균값을 기록하였다.

표 1. 데이터셋별 모델 성능 비교

Table 1. Comparison of classification performance across medical image datasets

| Dataset | Model          | Accuracy       | AUC            | MacroF1        |
|---------|----------------|----------------|----------------|----------------|
| Kva     | EfficientNetV2 | 93.15<br>±1.23 | -              | 93.25<br>±1.22 |
|         | ViT            | 94.45<br>±0.47 | -              | 94.47<br>±0.47 |
|         | ConvNeXtV2     | 95.72<br>±1.05 | -              | 95.71<br>±1.05 |
| HAM     | EfficientNetV2 | 87.39<br>±1.42 | -              | 80.08<br>±1.96 |
|         | ViT            | 88.72<br>±1.85 | -              | 82.87<br>±1.64 |
|         | ConvNeXtV2     | 89.92<br>±0.59 | -              | 84.87<br>±0.98 |
| CXray   | EfficientNetV2 | -              | 83.13<br>±0.38 | 31.77<br>±1.63 |
|         | ViT            | -              | 82.31<br>±0.36 | 28.66<br>±0.47 |
|         | ConvNeXtV2     | -              | 84.35<br>±0.30 | 33.18<br>±1.11 |

Kvasir-v2에서 ConvNeXtV2-B는 Accuracy 95.72±1.05%, Macro F1 95.71±1.05%로 가장 높은 평균 성능을 보였다. ViT-B는 각각 94.45±0.47%와 94.47±0.47%, EfficientNetV2-M은 93.15±1.23%와 93.25±1.22%를 기록하였다. ConvNeXtV2-B는 ViT-B보다 Accuracy에서 1.27%, Macro F1에서 1.24% 높았으며, EfficientNetV2-M과 비교해서는 각각 2.57%

와 2.46% 높은 결과를 나타냈다. 세 모델 모두 Accuracy와 Macro F1이 유사하게 나타난 것은 전체 표본에 대한 정확도와 클래스별 성능을 동일한 비중으로 반영한 결과 사이에 큰 차이가 없었음을 의미한다. 이는 클래스 분포가 비교적 균형적이고 점막의 질감, 색상 변화 및 병변 경계와 같은 시각적 특징이 비교적 명확한 Kvasir-v2의 특성이 반영된 결과로 볼 수 있다. 이러한 조건에서 ConvNeXtV2-B는 세 모델 중 가장 높은 평균 성능을 기록하였으나, ViT-B와의 차이가 크지 않으므로 유사한 고성능 범위에서 상대적으로 우세한 결과로 해석하는 것이 적절하다.

HAM10000에서도 ConvNeXtV2-B가 Accuracy  $89.92 \pm 0.59\%$ , Macro F1  $84.87 \pm 0.98\%$ 로 두 지표 모두 가장 높은 평균값을 기록하였다. ViT-B는 Accuracy  $88.72 \pm 1.85\%$ , Macro F1  $82.87 \pm 1.64\%$ 를 보였으며, EfficientNetV2-M은 각각  $87.39 \pm 1.42\%$ 와  $80.08 \pm 1.96\%$ 를 기록하였다. ConvNeXtV2-B는 ViT-B보다 Accuracy에서 1.20%, Macro F1에서 2.00% 높았고, EfficientNetV2-M보다 각각 2.53%와 4.79% 높은 성능을 나타냈다.

HAM10000에서는 세 모델 모두 Macro F1이 Accuracy보다 낮게 나타났으며, 두 지표의 차이는 EfficientNetV2-M 7.31%, ViT-B 5.85%, ConvNeXtV2-B 5.05%였다. HAM10000은 클래스별 표본 수의 차이가 크고 일부 피부 병변이 색상, 형태 및 경계 측면에서 유사하기 때문에 다수 클래스와 소수 클래스 간 분류 성능의 차이가 발생할 수 있다. Accuracy는 전체 표본의 정분류 비율을 반영하므로 다수 클래스의 영향을 상대적으로 크게 받는 반면, Macro F1은 각 클래스의 F1 점수를 동일한 비중으로 평균하여 소수 클래스의 성능도 함께 반영한다. 따라서 ConvNeXtV2-B가 가장 높은 Macro F1을 기록하고 Accuracy와 Macro F1 간 차이도 가장 작게 나타난 것은 전체 분류 정확도뿐만 아니라 클래스별 균형 성능에서도 상대적으로 우수한 결과를 보였음을 의미한다. 또한 ConvNeXtV2-B는 두 지표에서 가장 작은 표준편차를 기록하여 반복 실험에 따른 성능 변동도 상대적으로 작게 나타났다.

ChestXray14에서는 ConvNeXtV2-B가 AUC  $84.35 \pm 0.30\%$ , Macro F1  $33.18 \pm 1.11\%$ 로 가장 높은 평균 성능을 기록하였다. EfficientNetV2-M은 AUC  $83.13 \pm 0.38\%$ , Macro F1  $31.77 \pm 1.63\%$ 를 보였으며, ViT-B는 각각  $82.31 \pm 0.36\%$ 와  $28.66 \pm 0.47\%$ 를 기록하였다. ConvNeXtV2-B는 EfficientNetV2-M보다 AUC에서 1.22%, Macro F1에서 1.41% 높았고, ViT-B보다 각각 2.04%와 4.52% 높은 결과를 나타냈다.

ChestXray14에서 세 모델의 AUC는 모두 82% 이상이었으나, Macro F1은  $28.66 - 33.18\%$ 로 상대적으로 낮게 나타났다. ChestXray14는 하나의 영상에 여러 질환이 동시에 존재할 수 있는 다중 라벨 데이터셋이며, 질환별 양성 표본 수도 큰 차이가 있다. AUC는 특정 결정 임계값에 의존하지 않고 양성 표본과 음성 표본을 구분하는 전반적인 순위화 능력을 평가한다. 반면 Macro F1은 결정 임계값이 적용된 최종 예측 결과를 평가하므로 질환별 양성 표본의 희소성과 임계값 설정의 영향을 직접적으로 받는다. 따라서 상대적으로 높은 AUC와 낮은 Macro F1이 함께 나타난 것은 상충하는 결과가 아니며, 전반적인 질환 구분 능력에 비해 일부 희소 질환의 양성 사례를 정확하게 판별하는 것이 어려웠음을 보여준다. 이러한 조건에서도 ConvNeXtV2-B는 AUC와 Macro F1 모두에서 가장 높은 평균값을 기록하여, 질환의 전반적인 구분 능력과 임계값 기반의 최종 예측 성능에서 상대적으로 우수한 결과를 나타냈다.

종합하면, ConvNeXtV2-B는 영상의 시각적 특성과 라벨 구조가 서로 다른 세 데이터셋에서 모든 주요 평가지표의 가장 높은 평균값을 기록하였다. Kvasir-v2에서는 세 모델 모두 높은 성능을 보인 가운데 ConvNeXtV2-B가 소폭 우세하였으며, HAM10000에서는 Accuracy와 Macro F1 모두에서 가장 높은 성능을 나타냈다. ChestXray14에서도 AUC와 Macro F1이 모두 가장 높게 나타나 다중 라벨 분류 환경에서도 상대적인 우위를 보였다. 다만 별도의 통계적 유의성 검정 없이 평균값의 차이만으로 모델 간 우위를 확정할 수는 없다. 따라서 표 1의 결과는 ConvNeXtV2-B가 특정 데이터셋에서 압도적인 성능 향상을 달성했다기보다, 서로 다른

의료영상 도메인과 라벨 구조에서 일관되게 높은 평균 분류 성능을 보였다는 점에 의미가 있다.

이러한 전체 성능 차이가 실제 클래스별 예측에서 어떠한 형태로 나타나는지 확인하기 위해 그림 1에 HAM10000 데이터셋의 혼동행렬을 제시하였다. 그림 1은 세 차례의 반복 실험에서 얻은 테스트 예측 결과를 합산한 것으로, 본 실험의 라벨 순서에서 클래스 2는 benign keratosis-like lesions(bkl), 클래스 4는 melanoma(mel), 클래스 5는 melanocytic nevi(nv)에 해당한다. 세 모델에서 공통적으로 bkl, mel 및 nv 사이의 혼동이 두드러졌으며, 특히 nv와 mel 사이에서 양방향 오분류가 빈번하였다. 이는 두 병변이 모두 색소성 피부 병변으로 관찰되고 색상 분포, 내부 질감 및 경계의 불규칙성과 같은 시각적 특징이 부분적으로 유사하여, 국소 특징이나 전체 형태만으로는 안정적인 구분이 어려운 것으로 해석된다.

모델별로 살펴보면, EfficientNetV2-M은 실제 nv를 mel로 오분류한 사례가 66건으로 가장 많았으며, nv의 정분류 수는 1,909건이었다. EfficientNetV2-M은 연산 효율성과 특징 추출 성능의 균형을 추구하지만, 채널 확장과 압축이 반복되는 과정에서 유사한 병변을 구분하는 데 필요한 세밀한 경계 변화와 병변 전체의 형태적 관계가 충분히 유지되지 못했을 가능성이 있다. 특히 nv와 mel의 차이는 개별 색상이나 질감보다 특징들의 공간적 분포와 결합 양상에서 나타날 수 있으므로, 이러한 오분류는 표 1에서 가장 낮은 Accuracy와 Macro F1을 기록한 결

과와도 연결된다.

ViT-B에서는 nv를 mel로 오분류한 사례가 58건으로 EfficientNetV2-M보다 적었으나, nv의 정분류 수는 1,903건으로 세 모델 중 가장 적었다. 또한 nv를 bkl로 오분류한 사례가 43건으로 가장 많아, nv에 대한 혼동이 mel뿐만 아니라 bkl 방향으로도 분산되었다. ViT-B는 셀프 어텐션을 통해 서로 떨어진 영역 간 관계와 전체 형태를 반영하는 데 유리하지만, 영상을 고정된 크기의 패치로 분할하는 과정에서 미세한 경계와 질감 변화, 병변 주변부의 연속적인 색상 변화가 충분히 보존되지 않았을 가능성이 있다. 따라서 전역적 관계를 반영하는 능력만으로는 유사한 색소성 병변을 안정적으로 구분하기 어려웠던 것으로 해석된다.

ConvNeXtV2-B는 nv를 mel로 오분류한 사례가 51건으로 EfficientNetV2-M의 66건과 ViT-B의 58건보다 적었으며, nv의 정분류 수도 1,919건으로 가장 많았다. 이는 합성곱 기반의 계층적 특징맵을 통해 색상, 질감 및 경계와 같은 국소 정보를 보존하면서, 7×7 깊이별 합성곱으로 비교적 넓은 영역의 형태와 주변 문맥을 함께 반영한 결과로 해석할 수 있다. 또한 GRN은 채널 간 전역 응답의 경쟁을 유도하여 특정 색상이나 질감에 대한 과도한 의존을 완화하고, 여러 시각적 단서를 종합적으로 활용하는 데 기여했을 가능성이 있다. 이러한 구조적 특성은 nv의 분류 성능을 유지하면서 mel과의 혼동을 상대적으로 줄인 결과와 부합한다.

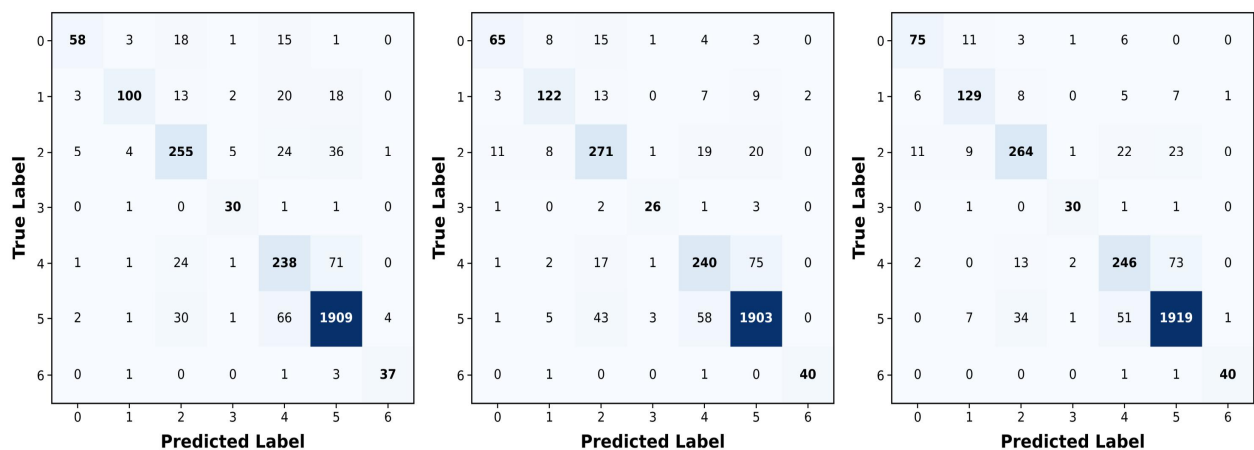


그림 1. HAM10000 혼동행렬 비교

Fig. 1. Confusion matrix comparison on HAM10000

다만 실제 mel을 nv로 오분류한 사례는 EfficientNetV2-M 71건, ViT-B 75건, ConvNeXtV2-B 73건으로 모델 간 차이가 크지 않았다. 따라서 ConvNeXtV2-B는 nv를 mel로 오분류하는 경향을 상대적으로 완화하였지만, mel을 nv와 구분하는 문제까지 충분히 해결한 것은 아니다. 이는 두 클래스의 시각적 유사성이 특정 모델의 구조적 한계라기보다 HAM10000의 공통적인 분류 난제로 작용했음을 의미한다. 또한 혼동행렬만으로 모델 구조를 성능 차이의 직접적인 원인으로 단정할 수 없으므로, 구조적 특성과 관찰된 오분류 양상이 부합하는 수준으로 해석하는 것이 적절하다.

그림 2는 HAM10000 데이터셋에서 nv에 대한 Grad-CAM 결과를 비교한 것이다. 이를 통해 그림 1의 혼동행렬에서 확인된 유사 색소성 병변 클래스 간 오분류 양상을 보다 구체적으로 해석하고자 하였다. 해당 영상의 실제 클래스는 nv이지만, EfficientNetV2-M은 이를 mel로 오분류하였고, ViT-B와 ConvNeXtV2-B는 nv로 정분류하였다.

EfficientNetV2-M의 Grad-CAM 결과에서는 병변 중앙의 짙은 색소 영역과 병변 내부에 강한 활성화가 집중되었다. 이는 모델이 병변 자체를 인식하지 못한 것이 아니라, 중앙의 짙은 색조와 내부의 색상 및 질감 변화를 주요 판단 근거로 활용했음을 의미한다. 그러나 이러한 특징은 nv와 mel에서 공통적으로 나타날 수 있으므로, 병변 내부의 특정 색소 영역만으로 두 클래스를 명확하게 구분하기는 어렵다. 특히 병변 외곽과 주변 피부 사이의 관계에 대한 활성화가 상대적으로 제한적으로 나타난 점을 고려하면, EfficientNetV2-M은 중앙의 짙은 색소와 내부 질감에 과도하게 의존하여 해당 영상을 mel로 오분류했을 가능성이 있다.

ViT-B에서는 중앙의 짙은 색소 영역보다 병변의 외곽과 주변 피부를 포함한 비교적 넓은 영역에 활성화가 분산되어 나타났다. 이는 병변 내부의 특정 색소 특징에만 의존하지 않고 병변의 전체적인 형태와 주변 영역 사이의 공간적 관계를 함께 활용했음을 시사한다. 이러한 활성화 양상은 중앙의 짙은 색소를 mel의 결정적인 단서로 해석하는 것을 완화하여 해당 영상을 nv로 정분류하는 데 기여했을 수 있다. 다만 활성화가 병변 외부까지 넓게 분산되어

있으므로, 병변 경계를 정밀하게 포착했다기보다는 병변과 주변부를 포함한 전반적인 문맥을 활용한 것으로 해석하는 것이 적절하다.

ConvNeXtV2-B에서는 중앙의 짙은 색소 영역에도 비교적 약한 활성화가 나타났으나, 해당 영역에만 집중하지 않고 병변 내부의 색상 및 질감 변화와 외곽 경계, 주변 피부와 맞닿는 영역까지 활성화가 분포하였다. 이는 중앙의 색소 특징을 일부 판단 근거로 반영하면서도 병변 내부와 외곽에 분포한 여러 시각적 단서를 함께 활용했음을 보여준다. 이처럼 특정 색소 영역에 대한 의존을 줄이고 병변 내부의 변화와 주변 경계의 관계를 종합적으로 반영한 점이 해당 영상을 nv로 정분류하는 데 기여했을 가능성이 있다.

결과적으로 그림 2는 EfficientNetV2-M의 오분류가 병변 중앙의 짙은 색소와 내부 질감에 대한 높은 의존 및 병변 외곽과 주변 피부 사이의 관계에 대한 제한적인 반영과 관련될 수 있음을 보여준다. 반면 ViT-B는 병변과 주변 영역에 걸친 넓은 공간적 문맥을 활용하였으며, ConvNeXtV2-B는 중앙의 색소 특징을 일부 반영하면서 병변 내부의 변화와 외곽 경계 및 주변 문맥을 함께 활용하여 nv를 정분류한 것으로 해석된다.

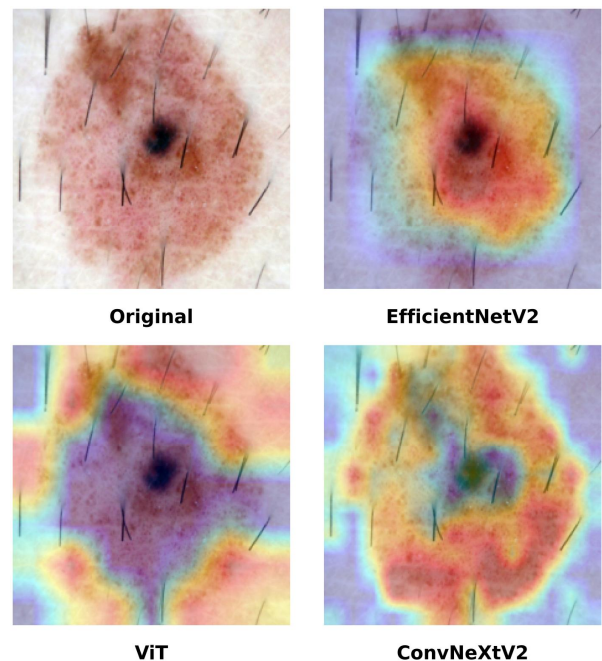


그림 2. 클래스 5(nv) 오분류 사례에 대한 Grad-CAM 비교  
Fig. 2. Grad-CAM comparison for a misclassification case of class 5(nv)

그림 3을 보면 ConvNeXtV2-B는 대부분의 질환 클래스에서 EfficientNetV2-M 및 ViT-B와 비교하여 같거나 더 높은 F1 점수를 보였다. 특히 Atelectasis, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pneumothorax 등 여러 클래스에서 ConvNeXtV2-B가 가장 높은 성능을 나타냈다. 이는 표 1에서 ConvNeXtV2-B가 ChestXray14 데이터셋의 전체 F1 점수와 AUC에서 가장 높은 성능을 보인 결과와 일치한다.

ChestXray14의 질환들은 단순한 색상이나 명확한 객체 경계로 구분되는 것이 아니라, 흉부 영역 내의 미세한 음영 변화, 병변 위치, 주변 조직과의 관계, 넓은 영역의 구조적 변화로 나타나는 경우가 많다. 따라서 모델은 국소적인 질환 특징뿐만 아니라 흉부 영상 전체의 문맥 정보도 함께 고려해야 한다. ConvNeXtV2-B는 합성곱 신경망 기반의 계층적 구조를 통해 초기 단계에서 세밀한 국소 특징을 추출하고, 후반 단계에서 더 넓은 영역의 특징을 통합할 수 있다. 또한 7×7 깊이별 합성곱은 넓은 수용영역을 제공하므로, 병변 주변의 구조적 정보를 함께 반영하는 데 유리하다. 이러한 구조적 특징이 ChestXray14에서 ConvNeXtV2-B가 여러 질환 클래스에 대해 비교적 높은 F1 점수를 보인 원인으로 해석된다.

특히 Hernia 클래스에서 모델 간 차이가 뚜렷하게 나타났다. ConvNeXtV2-B와 EfficientNetV2-M은 Hernia에서 상대적으로 높은 F1 점수를 보인 반면,

ViT-B는 매우 낮은 F1 점수를 기록하였다. Hernia는 ChestXray14에서 상대적으로 샘플 수가 적고, 영상 내에서 질환 특징이 넓게 분포하기보다 특정 위치의 구조적 변화로 나타나는 경우가 많다. 이처럼 데이터 수가 적고 국소적 위치 정보가 중요한 클래스에서는 패치 단위로 영상을 분할하여 전역 관계를 학습하는 ViT-B가 세밀한 공간적 단서를 충분히 보존하지 못했을 가능성이 있다. 반면 ConvNeXtV2-B는 합성곱 기반 구조를 유지하기 때문에 국소적인 공간 특징을 단계적으로 보존하며 학습할 수 있고, 큰 커널 기반의 깊이별 합성곱을 통해 병변 주변의 문맥까지 함께 반영할 수 있어 Hernia와 같이 국소 구조 변화가 중요한 클래스에서 상대적으로 높은 성능을 보인 것으로 해석할 수 있다.

Effusion, Infiltration, Pneumothorax와 같이 F1 점수가 비교적 높게 나타난 클래스에서도 ConvNeXtV2-B는 가장 우수하거나 상위 수준의 성능을 보였다. 이러한 질환들은 흉부 X선에서 일정 영역의 음영 변화나 폐 영역의 구조적 변화를 동반하는 경우가 많다. ConvNeXtV2-B의 계층적 단계 구조는 이러한 질환의 국소적 패턴과 넓은 영역의 구조적 변화를 함께 학습하는 데 적합하며, GRN은 특징맵의 전역 응답을 조절하여 특정 채널이나 일부 특징에 과도하게 의존하는 것을 완화한다. 이로 인해 ConvNeXtV2-B는 다중 라벨 환경에서도 여러 질환 특징을 비교적 균형 있게 반영할 수 있었던 것으로 볼 수 있다.

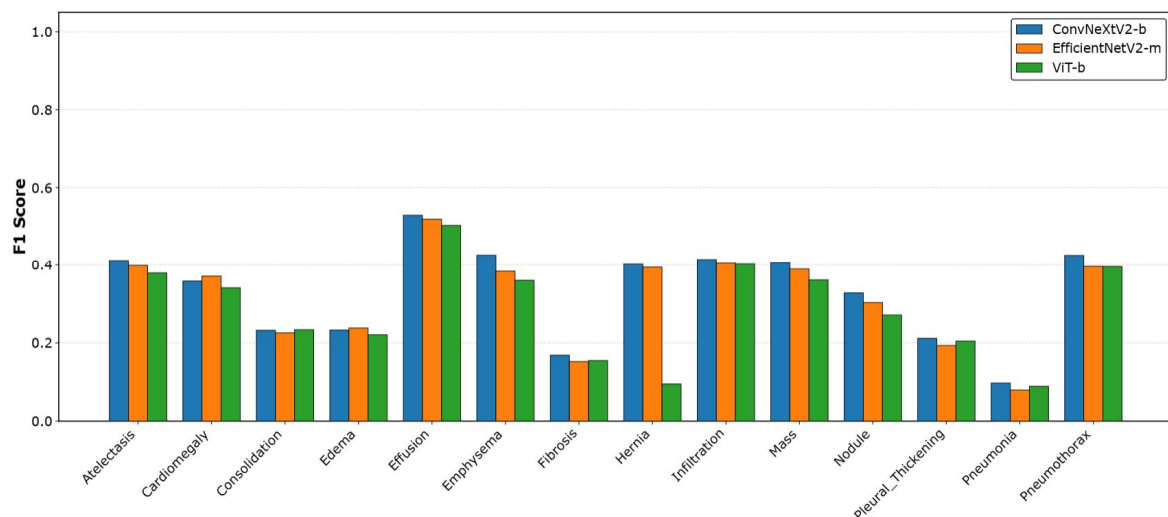


그림 3. ChestXray14 데이터셋의 질환별 F1 점수 비교  
Fig. 3. Class-wise F1 score comparison on ChestXray14 dataset

반면 Pneumonia, Fibrosis, Pleural Thickening과 같은 일부 클래스에서는 세 모델 모두 상대적으로 낮은 F1 점수를 보였다. 이는 해당 질환들이 영상 내에서 뚜렷한 형태로 나타나기보다 미세한 음영 차이나 다른 질환과 유사한 패턴으로 표현되기 때문으로 해석할 수 있다. 또한 ChestXray14는 질환별 샘플 수가 불균형한 데이터셋이므로, 일부 소수 클래스에서는 충분한 학습이 이루어지기 어렵다. 따라서 ConvNeXtV2-B가 전체적으로 가장 우수한 성능을 보였더라도, 모든 질환 클래스에서 높은 F1 점수를 확보한 것은 아니며, 소수 클래스와 시각적 구분이 어려운 질환에서는 여전히 한계가 존재한다.

표 2는 EfficientNetV2-M, ViT-B, ConvNeXtV2-B의 연산 효율성을 비교한 결과이다. EfficientNetV2-M은 세 모델 중 가장 경량적인 구조로, 제한된 연산 자원 환경에서 장점을 가질 수 있다. 그러나 앞선 성능 분석에서 EfficientNetV2-M은 세 데이터셋 전반에서 ConvNeXtV2-B보다 낮은 성능을 보였고, 특히 HAM10000과 ChestXray14처럼 클래스 간 구분이 어렵거나 다중 라벨 구조를 갖는 데이터셋에서 한계가 나타났다. 따라서 EfficientNetV2-M은 계산 비용 측면에서는 유리하지만, 복잡한 의료영상 특징을 충분히 표현하는 데에는 상대적으로 제한이 있었던 것으로 해석할 수 있다.

표 2. 모델별 연산 효율성 비교 결과  
Table 2. Model efficiency comparison

| Model            | Params (M) | FLOPs (G) | Inference (ms) | Memory (MB) |
|------------------|------------|-----------|----------------|-------------|
| EfficientNetV2-M | 52.87      | 5.22      | 19.74          | 898.85      |
| ViT-B            | 85.80      | 16.85     | 5.39           | 1260.01     |
| ConvNeXtV2-B     | 87.80      | 15.35     | 11.77          | 1301.86     |

ViT-B는 추론 속도 측면에서 장점을 보였으나, ChestXray14에서 ConvNeXtV2-B보다 낮은 AUC와 F1 점수를 보였고, HAM10000에서도 유사 클래스 간 오분류가 나타났다. 이는 ViT-B가 패치 간 전역 관계를 빠르게 처리할 수 있는 구조적 장점을 가지지만, 제한된 의료영상 데이터 환경이나 세밀한 국소 특징이 중요한 분류 문제에서는 성능 안정성에 한계가 있을 수 있음을 보여준다.

ConvNeXtV2-B는 EfficientNetV2-M보다 더 많은

계산 자원을 요구하지만, 세 데이터셋에서 가장 높은 성능을 보였다. 또한 ViT-B보다 빠른 모델은 아니지만, 국소 특징과 넓은 문맥 정보가 동시에 요구되는 의료영상 분류에서 더 균형 잡힌 성능을 나타냈다. 따라서 ConvNeXtV2-B는 가장 경량적인 모델은 아니지만, 정확도와 일반화 안정성이 중요한 의료영상 분류 환경에서 성능과 계산 비용 사이의 균형을 갖춘 모델로 해석할 수 있다.

#### IV. 결론 및 향후 연구

본 연구에서는 ConvNeXtV2-B의 의료영상 분류 적용 가능성을 분석하기 위해 Kvasir-v2, HAM10000, ChestXray14 데이터셋을 대상으로 EfficientNetV2-M 및 ViT-B와의 비교 실험을 수행하였다. 세 데이터셋은 각각 내시경 영상, 피부 병변 영상, 흉부 X선 영상으로 구성되어 있으며, 서로 다른 영상 특성과 라벨 구조를 가진다. 이를 통해 ConvNeXtV2-B가 서로 다른 영상 특성과 라벨 구조를 갖는 의료영상 분류 환경에서 어떠한 성능적·구조적 특성을 보이는지 확인하고자 하였다.

실험 결과, ConvNeXtV2-B는 Kvasir-v2와 HAM10000에서 가장 높은 Accuracy와 Macro F1을 나타냈으며, ChestXray14에서도 가장 높은 AUC와 Macro F1을 기록하였다. 이는 ConvNeXtV2-B가 모든 의료영상 분류 문제에서 절대적인 우위를 가진다는 의미라기보다는, 본 연구에서 사용한 서로 다른 영상 특성과 라벨 구조의 세 데이터셋에서 전반적으로 우수하고 안정적인 성능을 보였음을 의미한다. 특히 ChestXray14와 같은 다중 라벨 흉부 X선 분류 환경에서는 AUC와 Macro F1 모두에서 가장 높은 성능을 보여, 복합적인 질환 특징이 나타나는 데이터셋에서 상대적으로 뚜렷한 장점을 확인할 수 있었다.

또한 혼동행렬, Grad-CAM 및 클래스별 F1 점수 분석을 통해 평균 성능만으로 확인하기 어려운 모델별 분류 특성을 살펴보았다. 분석 결과, ConvNeXtV2-B는 일부 사례에서 병변의 경계와 주변 문맥 정보를 함께 반영하는 경향을 보였으며, 이는 CNN 기반의 계층적 특징 추출 구조와 넓은 수용영역을 갖는 구조적 특성과 관련되는 것으로 해

석된다. 다만 이러한 결과는 특정 사례와 데이터셋에 기반한 해석이므로, ConvNeXtV2-B가 모든 의료영상 분류 문제에서 항상 우수하다고 일반화하기는 어렵다.

연산 효율성 측면에서 ConvNeXtV2-B는 EfficientNetV2-M보다 많은 계산 자원과 메모리를 요구하였다. 그러나 다양한 데이터셋에서 높은 분류 성능을 안정적으로 보였다는 점에서, 단순히 가장 정량적인 모델이라기보다는 정확도와 성능 안정성이 중요한 의료영상 분류 환경에서 고려할 수 있는 모델로 볼 수 있다. 따라서 ConvNeXtV2-B는 의료영상 분류에서 국소 특징과 문맥 정보를 함께 반영할 수 있는 현대적 CNN 기반 모델로서 활용 가능성을 가진다.

본 연구의 한계는 실험 대상이 내시경 영상, 피부 병변 영상, 흉부 X선 영상에 한정되어 CT, MRI, 초음파 등 다른 의료영상 도메인까지 검증하지 못했다는 점이다. 또한 Grad-CAM 분석은 일부 사례에 대한 정성적 해석에 머물렀으며, 병변 위치 주석이나 전문가 판독 결과를 활용한 정량적 설명 가능성 평가는 수행하지 못하였다. 향후 연구에서는 다양한 의료영상 도메인으로 실험 범위를 확장하고, 설명 가능성 분석을 정량적으로 검증할 필요가 있다. 아울러 클래스 불균형을 완화하기 위한 손실 함수, 데이터 증강 및 샘플링 전략을 적용하여 소수 클래스의 분류 성능을 개선하는 연구가 필요하다.

### Acknowledgment

본 논문은 2026년도 한국정보기술학회 하계종합 학술대회에서 발표한 논문(ConvNeXt 기반 의료영상 분류 및 기존 딥러닝 모델과의 성능 비교 분석[14])을 확장한 것임.

### References

- [1] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders", Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, pp. 16133-16142, Jun. 2023. <https://doi.org/10.1109/CVPR52729.2023.01548>.
- [2] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection", Proc. of the 8th ACM on Multimedia Systems Conference, Taipei, Taiwan, pp. 164-169, Jun. 2017. <https://doi.org/10.1145/3083187.3083212>.
- [3] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions", Scientific Data, Vol. 5, Art. no. 180161, Aug. 2018. <https://doi.org/10.1038/sdata.2018.161>.
- [4] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 3462-3471, Jul. 2017. <https://doi.org/10.1109/CVPR.2017.369>.
- [5] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training", Proc. of the 38th International Conference on Machine Learning, Online, Vol. 139, pp. 10096-10106, Jul. 2021. <https://doi.org/10.48550/arXiv.2104.00298>.
- [6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", International Conference on Learning Representations (ICLR), Online, May 2021. <https://doi.org/10.48550/arXiv.2010.11929>.

[1] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders", Proc. of the IEEE/CVF Conference on Computer

- [7] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning", arXiv Preprint, arXiv:1711.05225, Nov. 2017. <https://doi.org/10.48550/arXiv.1711.05225>.
- [8] S. Jain, U. Singhania, B. K. Tripathy, E. A. Nasr, M. K. Aboudaif, and A. K. Kamrani, "Deep Learning-Based Transfer Learning for Classification of Skin Cancer", Sensors, Vol. 21, No. 23, Art. no. 8142, Dec. 2021. <https://doi.org/10.3390/s21238142>.
- [9] M.-L. Huang and Y.-C. Liao, "Stacking Ensemble and ECA-EfficientNetV2 Convolutional Neural Networks on Classification of Multiple Chest Diseases Including COVID-19", Academic Radiology, Vol. 30, No. 9, pp. 1915-1935, Sep. 2023. <https://doi.org/10.1016/j.acra.2022.11.027>.
- [10] A. A. Waskita, J. M. Amda, D. S. K. Sihono, and H. Prasetyo, "EfficientNetV2 Based for MRI Brain Tumor Image Classification", Proc. of the 2023 10th International Conference on Computer, Control, Informatics and its Applications (IC3INA), Bandung, Indonesia, pp. 171-176, Oct. 2023. <https://doi.org/10.1109/IC3INA60834.2023.10285782>.
- [11] S. Regmi, A. Subedi, U. Bagci, and D. Jha, "Vision Transformer for Efficient Chest X-ray and Gastrointestinal Image Classification", arXiv Preprint, arXiv:2304.11529, Apr. 2023. <https://doi.org/10.48550/arXiv.2304.11529>.
- [12] A. Halder, S. Gharami, P. Sadhu, P. K. Singh, M. Woźniak, and M. F. Ijaz, "Implementing Vision Transformer for Classifying 2D Biomedical Images", Scientific Reports, Vol. 14, No. 1, Art. no. 12567, May 2024. <https://doi.org/10.1038/s41598-024-63094-9>.
- [13] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),

New Orleans, Louisiana, USA, pp. 11976-11986, Jun. 2022. <https://doi.org/10.1109/CVPR52688.2022.01167>.

- [14] M.-S. Kim, S.-W. Ko, and H.-G. Kim, "ConvNeXt-Based Medical Image Classification and Comparative Performance Analysis with Conventional Deep Learning Models", Proc. 2026 Summer Conference of the Korean Institute of Information Technology, Jeju, Korea, pp. 1705-1709, Jun. 2026.

### 저자소개

김민서 (Min-Seo Kim)



2020년 3월 ~ 현재 : 광운대학교  
전자융합공학과 학사과정  
관심분야 : 신호처리, 인공지능

고승완 (Seung-Wan Ko)



2022년 3월 ~ 현재 : 광운대학교  
동북아문화산업학부 학사과정  
관심분야 : 신호처리, 인공지능

김형국 (Hyoung-Gook Kim)



2002년 2월 :  
독일베를린공과대학교  
전자공학과  
컴퓨터과학(공학박사)  
2007년 3월 ~ 현재 : 광운대학교  
전자융합공학과 교수  
관심분야 : 신호처리, 인공지능