

두피 이미지 분석을 위한 다중 라벨, 다중 클래스 분류 모델에 대한 예비 연구

유현선, 고동환, 이세영, 김관*

연세대학교 의공학과, *연세대학교 AI융합과학원

dbgustjs0106@naver.com, donghwan22@yonsei.ac.kr, slee0895@yonsei.ac.kr, *kimkwan@yonsei.ac.kr

A Preliminary Study on a Multi-label, Multi-class Classification Model for Scalp Image Analysis

Yu Hyunseon, Ko Dong Hwan, Lee Se Young, Kim Kwan*

Department of Biomedical Engineering, Yonsei Univ.,

*Institute of AI Convergence Science, Yonsei Univ.

요약

본 연구는 단일 두피 영상에서 미세각질, 피지과다, 모낭사이홍반, 모낭홍반/농포, 비듬, 탈모 등 6개 지표의 질환 수준(양호:0 ~ 심각:3)을 동시에 추정할 수 있는 다중 라벨·다중 클래스 분류 모델을 구현하는데 목적이 있다. 640×480 해상도의 두피 이미지 및 질환 수준 정보가 포함된 라벨링 파일 등 각각 17,785개를 사용하며 데이터 학습, 점검, 평가를 위해 전체 데이터를 70:15:15 비율로 구분했다. ImageNet에서 사전학습된 EfficientNet-B4를 공유 백본으로 사용했으며, 질환 수준 별 4-class softmax 헤드 6개를 연결하였다. 이미지는 비율을 유지한 채 448×448로 정규화하고, AdamW를 최적화 함수로 사용했으며 학습률 0.0001, 배치 크기 16, 최대 30 epoch로 설정했다. 평가 데이터셋에서 overall macro-F1=0.518, Hamming accuracy=0.658을 확인했으며, 지표 별 macro-F1은 미세각질 0.458, 피지과다 0.517, 모낭사이홍반 0.661, 모낭홍반/농포 0.311, 비듬 0.624, 탈모 0.526으로 확인되었다. 제안한 모델은 한 장의 영상으로 6개 지표를 동시에 추정할 수 있는 가능성을 보였으나, 지표별 성능 차이는 해당 질환의 영상학적 특징과 연관된 것으로 판단된다. 향후 모델·세부 파라미터 비교 실험을 통해 성능을 개선하고, 실시간 수준의 추정 가능성을 검토할 예정이다.

I. 서론

건선, 탈모 등의 모발 및 두피 질환은 개인의 건강과 웰빙에 중대한 영향을 미칠 수 있어 이를 조기에 발견하고 이를 예방, 관리하는 것이 중요하다[1]. 헬스케어 산업 및 기술의 발달과 함께, 두피 영상 장치를 이용한 질환 분류 영상 처리 기법이 연구되고 있으며, 최근에는 CNN 기반 자동 분류·검출이 도입되어 진단 성능을 개선하는 연구 또한 진행되고 있다[1,2].

하지만, 두피 질환은 단일 증상만 존재하는 것이 아닌 여러 증상이 복합적으로 존재하는 경우가 많아 임상 환경을 고려한 두피 질환 분류 알고리즘을 개발하는 것이 중요하다.

본 연구는 단일 또는 다중 질환이 복합적으로 존재하는 두피 이미지에서 여러 질환 지표의 수준을 한 번에 예측하기 위한 다중 라벨, 다중 클래스 분류 모델을 설계하는 것을 목표로 한다.

II. 본론

본 연구에서 활용한 데이터셋은 640×480 해상도의 두피 이미지와, 두피 질환 관련 6개 지표(미세각질, 피지과다, 모낭사이홍반, 모낭홍반/농포, 비듬, 탈모)에 대해 0(정상)부터 3(심각)까지의 수치가 표기된 라벨링 파일로 구성된다. 이미지 및 라벨링 파일은 각각 17,785개를 사용했으며, 모델 학습 (Train), 파라미터 점검 (Validation), 그리고 성능 평가 (Test)를 위해 전체 데이터를 70:15:15 비율로 구분했다.

학습은 파이토치 기반으로 구현되었다. ImageNet에서 사전학습된 EfficientNet-B4를 공유 백본으로 사용하고, 그 위에 6개 독립 헤드(각 4-class softmax)를 연결하여 한 장의 영상에서 6개 지표의 등급을 동시

에 추론하도록 설계했다. 전처리에는 이미지 원본 비율을 유지하면서 축소 및 패딩 과정을 거쳐 448×448로 정규화하여 이미지 내 텍스트 정보를 유지하면서 연산량을 감소시켰다. 최적화는 AdamW를 사용, 학습률은 0.0001로 고정했으며, 배치 크기는 16으로 설정했다. 최대 30 epoch까지 학습하되, 검증 macro-F1이 5 epoch 이상 개선되지 않으면 학습을 조기 종료하여 불필요한 과적합과 계산을 방지했다.

클래스 불균형과 샘플 편향을 완화하기 위해 Focal loss를 사용하였고, $\gamma=2.0$ 으로 고정하였다. 학습 시 수평 반전, 밝기, 대조도, 색조, 채도, 회전, 스케일 조절 등 데이터 증강을 적용했다.

표 1. 데이터 라벨 분포

총 데이터 수 (=17,785)		질환 수준			
		양호 (0)	경증 (1)	중증 (2)	심각 (3)
질환 지표	미세각질	14758	1125	1345	557
	피지과다	3678	6879	6216	1012
	모낭사이홍반	5916	7247	3411	1211
	모낭홍반/농포	16963	552	196	74
	비듬	10585	4113	2483	604
	탈모	13003	3435	1075	272

모델의 성능 평가를 위한 지표는 6개의 질환 지표 각각, 0~3까지 4개의 질환 수준을 고려하여 TP (True Positive), FP (False Positive), FN (False Negative), TN (True Negative)를 구한 뒤 지표 별 macro-F1과 이를 평균계산한 overall macro-F1을 계산했다. 또한, 이미지 내 6개 지표 중 몇 개 지표에 대해 정확히 예측했는지 확인하기 위해 Hamming

Accuracy를 계산했다.

표 2는 각 질환 별 macro-F1 및 overall macro-F1를 정리한 것으로, 평가 데이터셋의 경우 overall macro-F1은 0.518, Hamming accuracy는 0.658로 확인되었다. 각 질환 지표 별 macro-F1은 미세각질 0.458, 피지 0.517, 모낭사이 홍반 0.661, 모낭 홍반/농포 0.311, 비듬 0.624, 탈모 0.526으로, 모낭사이 홍반과 비듬에서 상대적으로 높은 macro-F1을 보였고 모낭 홍반/농포가 가장 낮았다.

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}, \quad \text{Recall}_k = \frac{TP_k}{TP_k + FN_k}$$
$$F1_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}$$
$$\text{macro-F1}_l = \frac{1}{K} \sum_{k=1}^K F1_{l,k}$$
$$\text{overall macro-F1} = \frac{1}{L} \sum_{l=1}^L \left(\frac{1}{K} \sum_{k=1}^K F1_{l,k} \right), \quad L=6, K=4.$$
$$\text{Hamming Accuracy} = \frac{1}{NL} \sum_{i=1}^N \sum_{l=1}^L \mathbf{1}\{\hat{y}_{i,l} = y_{i,l}\}$$

그림 1. macro-F1 등 평가 지표 계산 방법

표 2. 질환 지표 별 macro-F1 및 overall macro-F1

기준	검점 (Validation)	평가 (Test)
미세각질	0.490	0.458
피지과다	0.506	0.517
모낭사이홍반	0.679	0.661
모낭홍반/농포	0.343	0.311
비듬	0.652	0.624
탈모	0.491	0.526
평균	0.527	0.518

그림 2는 각 질환 지표 별 Confusion matrix로, 질환 별 각 수준에 대해 예측된 비율을 표현한 것이다. 대각선 집중된 형태가 나타나지만 질환 별 차이가 관찰되었다. macro-F1이 상대적으로 높은 모낭사이홍반과 비듬 지표는 비율 추세가 비교적 안정적이지만, 모낭홍반/농포는 심각(3) 수준을 중증(2) 수준으로 과소추정하는 패턴이 관찰되었다.

학습 과정 중, 검점 데이터셋에 대한 macro-F1은 초기 6epoch 부근까지 상승한 뒤 서서히 줄어드는 추세를 보였다 (그림 3). 학습에 대한 loss는 epoch가 반복됨에 따라 지속적으로 줄어들었으나, 검점에 대한 loss는 상승하는 추세를 보였다. 이는 전형적인 과적합 양상으로, 개선을 위해 학습 모델 관련 파라미터의 최적화 과정이 필요하다.

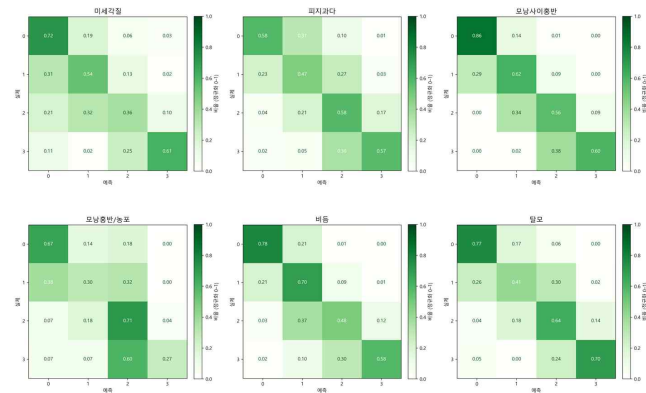


그림 2. 질환 별 예측 비율 (Confusion matrix)

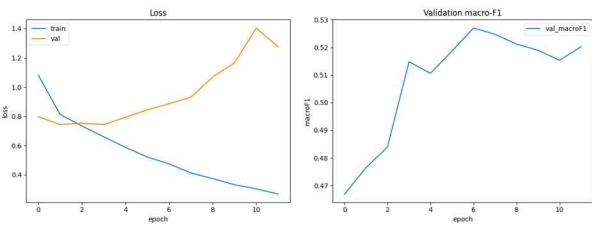


그림 3. 질환 별 예측 비율 (Confusion matrix)

III. 결론

본 연구는 두피 이미지를 통해 다양한 두피 질환의 수준을 복합적으로, 동시에 추정하는 것을 목표로 한다. 질환 별 macro-F1가 차이가 나는 것은 각 질환 별 영상학적 특징과 관련이 있을 수 있다. 모낭사이홍반은 두피 내 모세혈관 확장 등으로 적색 성분이 증가하며, 비듬은 기존 피부와 비교하여 흰색 수준의 차이가 나타날 수 있다. 모낭홍반/농포의 경우 두피의 홍반 뿐만 아니라 염증까지 고려된 것으로, 질환 수준을 구별하기에 상대적으로 복잡할 수 있다. 향후 모델 별, 세부 파라미터 별 학습 및 성능 비교를 통해 성능 개선 및 모델의 실시간 수준의 질환 추정 가능성을 검토할 예정이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2025년도 의료 AI반도체 교육·연구 플랫폼 구축사업의 결과로 수행되었음 (2024-0-0097).

참 고 문 헌

[1] B. M. K. Shetty et al., "Hair and Scalp Inspection System Using Deep Learning Based Analysis," 2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS), pp. 1-9, 2025.

[2] L. Bi et al., "Multi-Label classification of multi-modality skin lesion via hyper-connected convolutional neural network," Pattern Recognition, 107, 107502, 2020.