Journal of KIIT. Vol. 23, No. 11, pp. 57-66, Nov. 30, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.11.57

당뇨망막병증 등급 분류를 위한 글로컬 특징 융합 기반 하이브리드 모델

도지현*, 손창환**

Hybrid Model with Glocal Feature Fusion for Diabetic Retinopathy Grading

Ji-Hyun Do*, Chang-Hwan Son**

요 약

당뇨망막병증은 실명의 주요 원인 중 하나로, 조기 진단과 정확한 등급 분류가 매우 중요하다. 기존 합성곱 신경망 계열 모델은 미세한 병변의 국소 정보를 효과적으로 추출할 수 있으나, 전체 영상의 구조적·의미적 정보를 이해하는 데에는 한계가 있다. 반면 트랜스포머 계열 모델은 전역 표현 학습에 강점이 있지만, 국소 특징을 정교하게 포착하기는 어렵다. 본 논문에서는 이러한 각 계열의 장점을 결합하기 위해, 합성곱 신경망 기반의 국소 특징 추출 분기와 트랜스포머 기반의 전역 특징 추출 분기를 병렬로 구성한 하이브리드 모델을 제안한다. 두 분기에서 추출된 특징은 특징 융합 모듈을 통해 차원 정합과 비선형 변환을 거쳐 통합되며, 이를 기반으로 최종 당뇨망막병증 등급을 예측한다. 데이터셋 실험 결과, 제안한 모델은 정확도 83.26%와 제곱 가중치 카파 지수 86.62%를 기록하였으며, 기존 단일 구조 및 하이브리드 모델보다 더 우수한 성능을 달성하였다.

Abstract

Diabetic retinopathy is one of the leading causes of blindness, making early diagnosis and accurate grading critically important. Conventional convolutional neural network-based models effectively extract local features of subtle lesions but have limitations in understanding the structural and semantic information of the entire image. In contrast, transformer-based models excel at learning global representations but struggle to precisely capture local features. To combine the strengths of both architectures, this paper proposes a hybrid model that parallelly integrates a convolutional neural network branch for local feature extraction and a transformer branch for global feature extraction. The features extracted from both branches are fused through a feature fusion module involving dimensional alignment and nonlinear transformation, based on which the final diabetic retinopathy grade is predicted. Experimental results on the dataset show that the proposed model achieved an accuracy of 83.26% and a quadratic weighted kappa score of 86.62%, outperforming conventional single-architecture and hybrid models.

Keywords
diabetic retinopathy grading, vision transformer, hybrid architecture

* 국립군산대학교 소프트웨어학과 학사과정
- ORCID: https://orcid.org/0009-0008-5314-3247
** 국립군산대학교 소프트웨어학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0001-7077-3074

· Received: Sep. 30, 2025, Revised: Oct. 30, 2025, Accepted: Nov. 02, 2025
· Corresponding Author: Chang-Hwan Son
Department of Software Science and Engineering, Kunsan National University, Republic of Korea
Tel.: +82-63-469-8915, Email: cson@kunsan.ac.kr

Ⅰ. 서 론

당뇨망막병증(DR, Diabetic Retinopathy)은 당뇨병으로 인한 망막 혈관 손상에 의해 발생하는 주요 합병증으로, 시력 저하와 실명의 주요 원인 중 하나이다. DR은 미세출혈, 삼출물, 신생혈관 등 다양한 병변이 안저 이미지에 나타나며, 병변의 분포와 정도에 따라 등급이 구분된다. 그러나 이러한 병변들은 크기가 작고 혈관과 유사한 형태로 퍼져 있어 정밀한 판독이 어렵고, 진단에는 높은 전문성이 요구된다. 기존의 DR 진단은 숙련된 전문가의 육안 판독에 의존하기 때문에 시간과 인력 소모가 크며, 이를 대체할 자동화 기술 개발의 필요성이 꾸준히 제기되고 있다.

이러한 문제를 해결하기 위해, 안저 이미지 기반 DR 등급 분류 연구가 활발히 진행되었고 딥러닝 모델이 주요 방법론으로 자리 잡았다. 특히, 합성곱 신경망(CNN, Convolutional Neural Networks)은 국소 영역의 세밀한 시각 특징을 효과적으로 추출하는 강점을 지니고 있으며, 대표적인 모델로는 ResNet[1], Inception V3[2], EfficientNet[3] 등이 있다. 한편, DR의 질환 진행은 미세혈관 손상, 출혈, 삼출물, 신생혈관 등 망막 전역에서 나타나는 병리적 변화와 밀접하게 연관되므로, 전역 문맥 정보의 반영이 중요하다. 이를 위해, 어텐션 메커니즘을 활용하여 장기 문맥 정보를 효과적으로 학습할 수 있는 비전 트랜스포머(ViT, Vision Transformer) 아키텍처가 제안되었다. 대표적인 ViT 모델로는 CrossViT[4], MViT[5] 등이 있다. 그러나 DR 관련 병변은 크기뿐만 아니라 공간적 분포에서도 큰 다양성을 보이므로, CNN이나 ViT 계열의 단일 모델만으로는 DR 등급 분류 정확도를 충분히 향상하는 데 한계를 지닐 수 밖에 없다.

따라서 본 논문에서는 DR 등급별로 나타나는 다양한 병변의 크기와 공간적 분포를 효과적으로 반영하기 위해, 국소 및 전역 특징을 융합·상호 보완할 수 있는 하이브리드 모델을 제안한다. 구체적으로, 국소 특징 추출에 강점을 지닌 CNN 계열 ResNet-50과 멀티스케일 기반 전역 특징 추출이 가능한 MViT를 병렬 구조로 결합하여 하이브리드 아키텍처를 설계하였다. 실험 결과, 제안한 하이브리드 모델은 중증도별 병변을 효과적으로 포착할 수 있었으며, 분류 정확도 측면에서도 기존의 CNN 및 ViT 계열보다 더 우수한 성능을 획득하였다.

본 논문의 기여도는 다음과 같다. 먼저, 기존 단일 모델의 한계점을 극복하고자, CNN과 ViT 모델을 병렬 구조로 결합하여 DR 분류 예측에 적합한 하이브리드 모델을 제시하였다. 둘째, DDR 데이터셋 실험을 통해, 제안한 하이브리드 모델이 DR 등급 분류에서 기존의 SOTA 모델 대비 가장 우수한 정확도 성능을 달성하였다. 셋째, 제안한 병렬 구조 기반의 하이브리드 접근 방식이 국소적 단서와 전역적 문맥을 동시에 고려하여 DR 자동 판독의 정확성과 신뢰성을 높이는 데 효과적임을 검증하였다. 이러한 결과는 향후 DR 조기 진단과 임상적 활용에 기여할 것으로 기대된다.

Ⅱ. 관련 연구

2.1 CNN 계열 모델

CNN은 합성곱 계층을 사용하는 모델로서, 학습 가능한 필터를 통해서 이미지에서 텍스처 및 에지와 같이 객체 특성을 표현할 수 있는 국소 특징을 추출할 수 있다. 이를 통해, 기존 수작업 기반의 분류기보다 이미지넷과 같은 대규모의 이미지 데이터셋에서 뛰어난 성능을 달성하였다.

초기 대표 모델로는 스킵 연결을 도입하여 기울기 소실 문제를 해결한 ResNet[1]이 있으며, 이후 지속적인 튜닝과 변형을 통해 다양한 이미지 분류 분야에서 여전히 우수한 성능을 보이고 있다. ResNet 이후에는 아키텍처 설계의 중요성이 강조되면서, 스킵 연결을 강화한 DenseNet[6], 멀티스케일 특징 추출을 가능하게 한 FPN(Feature Pyramid Network)[7], 리소스 제한 환경에 최적화된 경량화 모델 MobileNet[8] 등이 개발되었으며, 이를 통해 CNN은 다양한 환경과 목적에 맞춰 지속적으로 발전해왔다. 특히, DR 등급 분류와 같은 안저 이미지 진단 분야에서 CNN은 국소 병변(미세혈관류, 출혈, 삼출물 등)을 효과적으로 탐지하는 데 널리 활용되고 있다. 기존 연구에서는 CNN 계열 모델이 DR 등급 분류와 병변 검출에서 우수한 성능을 보였으나, 전역 문

맥 정보를 효과적으로 반영하지 못하는 한계가 지적되었다[9].

2.2 ViT 계열 모델

CNN 계열에 어텐션 메커니즘을 통합한 모델의 등장 이후[10], 자연어 처리 분야에서는 장기 문맥 이해 능력을 강화한 트랜스포머 아키텍처가 개발되었다[11]. 이러한 트랜스포머 구조는 이미지와 같은 고차원 벡터에 적용될 수 있도록 변형되어 ViT 모델로 발전하였다. ViT는 입력 이미지를 패치 단위로 분할하고, 각 패치를 선형 임베딩하여 트랜스포머의 자기 어텐션(Self-attention) 메커니즘을 적용함으로써, 이미지의 전역 문맥 정보를 효과적으로 학습할 수 있으며, CNN에 버금가거나 그 이상의 성능을 달성하였다. 또한, ViT 모델은 고차원 벡터 연산과 전역 정보 학습을 위해 멀티스케일 아키텍처 설계가 강조되었으며, 대표적인 관련 모델로 CrossViT[4], MViT[5], PVT[12], ROI-ViT[13] 등이 있다.

ViT 계열 모델도 CNN과 함께 최근 DR 등급 분류와 같은 안저 이미지 진단 분야에서 활발히 활용되고 있다[14]. 특히, ViT는 국소 병변뿐만 아니라 망막 전체 구조에 나타나는 병리적 변화를 효과적으로 반영할 수 있어, CNN이 가진 국소 특징 추출의 한계를 보완할 수 있다[15]. 이러한 연구를 기반으로, 최근에는 CNN과 ViT의 장점을 상호 보완적으로 결합한 하이브리드 접근 방식이 제안되었다. 대표적인 모델로는 CNN과 ViT를 결합한 CoAtNet[16]과 연산 속도를 개선한 FastViT[17] 모델 등이 있지만, DR 등급 분류와 같은 의료 데이터셋에 최적화된 모델이 아니다. 즉, 본 연구의 대상인 DR 등급 분류를 위한, 두 계열에 대한 연결 방식 및 각 계열의 적합한 모델 구조에 대한 분석도 미흡한 상태이다.

III. 당뇨망막병증 등급 분류를 위한 국소·전역 특징 융합 기반 하이브리드 모델

본 연구에서는 DR 병변의 다양한 크기와 모양에 적응적으로 학습하기 위해, CNN과 ViT 계열을 결합한 하이브리드 모델을 제안한다. 특히, 국소 및 전역 특징을 상호 보완하여 DR 등급 분류에서 특징 구별력을 강화할 수 있는 아키텍처를 설계한다. 특히, DR 병변의 다양한 크기와 공간적 분포를 효과적으로 포착하기 위해 멀티스케일 기반의 ViT 모델을 활용하여 전역 특징을 추출하고, 이러한 전역 특징을 보완하기 위해 CNN 모델과 병렬 구조로 결합하여 최종 하이브리드 모델을 완성한다.

3.1 제안한 접근 방법

그림 1은 제안한 DR 등급 분류용 하이브리드 아키텍처를 보여준다. 제안한 모델은 ViT 분기와 CNN 분기가 병렬로 결합된 이중 분기 구조로 구현된다. ViT 분기는 DR 병변의 다양한 크기와 공간적 분포를 반영한 장기 문맥 정보를 학습하기 위해 멀티스케일 방식으로 설계된다. 계층이 증가함에 따라, 멀티헤드 풀링 어텐션(MHPA, Multi-head Pooling Attention)을 통해 공간 해상도를 줄이고 대신 채널 해상도를 확장하는 전략을 적용한다. CNN 분기에서는 합성곱 계층과 완전 연결 계층(FC, Fully Connected Layer)을 통해 저수준에서 고수준까지의 이미지 특징을 추출한다. ViT와 CNN 분기에서 추출된 특징을 상호 보완하기 위해, 글로컬 특징 융합 모듈(Glocal feature fusion module)을 도입한다. 전역 및 국소 정보를 융합한 최종 글로컬 특징은 분류 헤드를 거쳐 DR 등급 분류를 위한 최종 예측값을 결정한다.

3.2 국소 정보 추출을 위한 CNN 분기

미세출혈 및 삼출물 등과 같은 작은 병변을 포착하기 위해서는 학습 가능한 합성곱 필터를 적용하는 것이 효과적이다. 따라서 제안한 모델에서는 CNN 분기를 위해, ResNet-50 모델을 사용한다. ResNet-50[1]은 기울기 소실 문제를 해결하기 위해, 잔차 학습 구조를 도입하였으며 스킵 연결을 통해 학습의 안정성과 효율성을 보장한다. ResNet-50의 기본 구조는 합성곱 계층, 배치 정규화, 활성화 함수로 이루어진 잔차 블록을 연속적으로 쌓은 형태이다. 각 잔차 블록에서 입력 x는 식 (1)과 같이 처리된다.

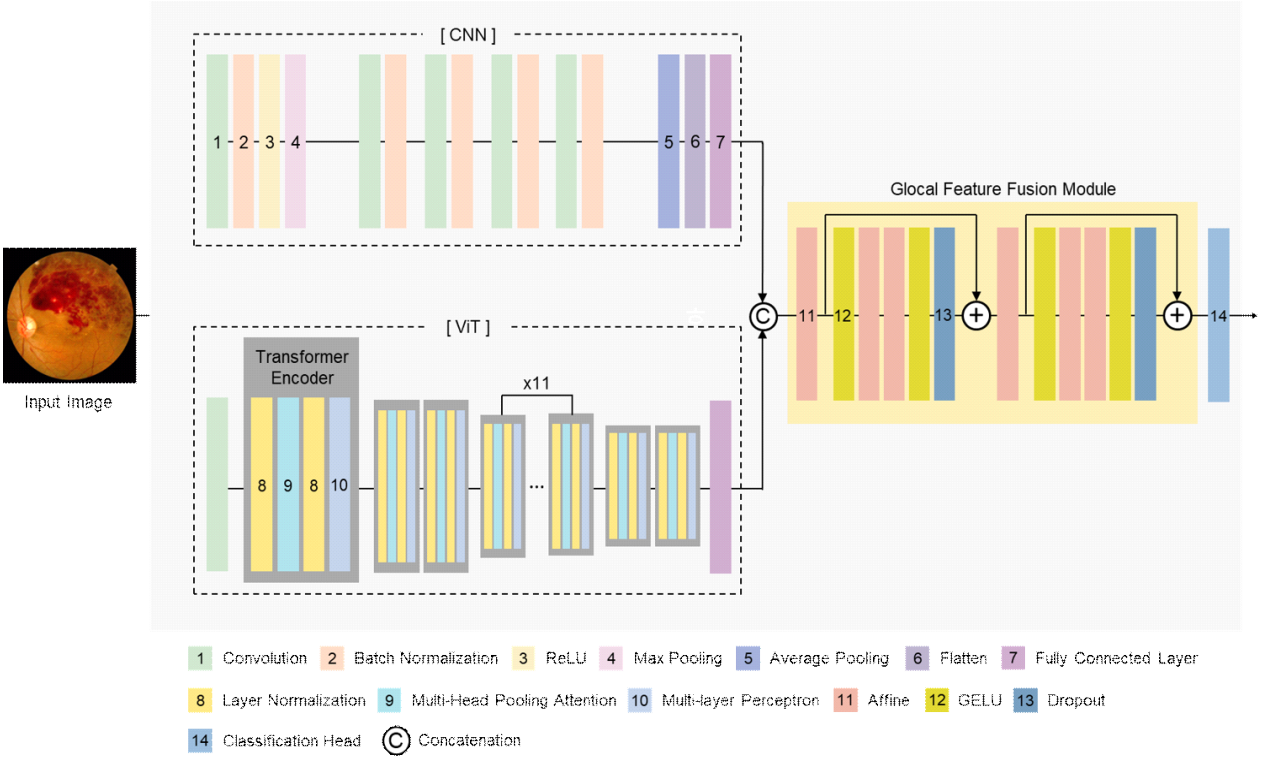

그림 1. 제안한 DR 등급 분류를 위한 하이브리드 모델
Fig. 1. Proposed hybrid model for DR grading

$$y = \sigma(BN(Conv(x))) + x \tag{1}$$

여기서 첫 번째 항목의 σ와 BN은 각각 ReLU 활성화 함수와 배치 정규화 함수를 나타낸다. 그리고 두 번째 항목은 스킵 연결을 의미한다.

그림 1의 CNN 분기에서는 다양한 입력 이미지 크기에 적응적으로 대응하고 동일한 크기의 특징 벡터를 생성하기 위해, 잔차 블록의 마지막 단계에 적응적 평균 풀링(Adaptive average pooling)을 적용한다. 이후 완전 연결 계층을 통해 병변 분류에 필요한 고수준 시맨틱 정보를 추출한다.

$$F_{cnn} = fc(pool(f_{local}(x)) \tag{2}$$

식 (2)의 fc와 $pool$은 각각 완전 연결 계층과 적응적 평균 풀링을 나타낸다. 그리고 f_{local}는 ResNet-50에서 마지막 잔차 블록까지 잘린 모델을 의미한다. 즉, 국소 영역의 공간적인 특징 추출을 담당하는 특징 추출기에 해당한다.

3.3 전역 정보 추출을 위한 ViT 분기

DR 등급 분류는 국소적 병변 특징뿐만 아니라, 망막 전체의 구조적 맥락과 병변 간 상호관계를 반영해야 높은 정확도를 달성할 수 있다. 이를 위해, 제안한 하이브리드 모델의 ViT 분기에는 멀티스케일 비전 트랜스포머(MViT)를 적용한다. 기존의 비전 트랜스포머와 달리, MViT는 멀티헤드 풀링 어텐션을 통해 다양한 스케일에서 전역적 특징 간 유사도를 효과적으로 모델링할 수 있어, DR 병변의 공간적 분포를 고려한 문맥 정보 학습에 적합하다.

MViT는 기존의 ViT와 유사하게 입력 이미지를 패치 단위로 분할한 후, 각 패치를 선형 임베딩하여 토큰 시퀀스로 변환한다.

$$X = \{x_1, x_2, ..., x_N\},\ x_i \in \mathbb{R}^{P^2 \cdot C},\ N = \frac{H \times W}{P^2} \tag{3}$$

$$Z_0 = [x_1 E; \cdots; x_N E] \in \mathbb{R}^{N \times D},\ D = 96 \tag{4}$$

식 (3)에서 x_i는 입력 영상에서 분할된 각각의 패치에 해당하며, P^2는 패치 크기, N은 패치 개수를 의미한다. 그리고 식 (4)의 $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$는 임베딩 공간으로 사상하기 위한 선형 변환 행렬을 나타낸다.

멀티스케일 표현을 학습하기 위해, MViT는 기존의 어텐션과는 달리 풀링 연산(Pooling)을 적용한다. 구체적으로, 각 계층 l에서 Query, Key, Value는 수식 (5)~(7)과 같이 풀링 연산을 적용해서 생성된다.

$$Q^{(l)} = Z^{(l)} W_Q \tag{5}$$

$$K^{(l)} = pool(Z^{(l)} W_k) \tag{6}$$

$$V^{(l)} = pool(Z^{(l)} W_V) \tag{7}$$

여기서 $pool$은 공간 해상도를 점진적으로 축소하는 풀링 연산을 의미한다. 따라서, 멀티헤드 풀링 어텐션은 다음과 같이 계산된다.

$$head_h = Softmax\left(\frac{Q_h^{(l)}(K_h^{(l)})^T}{\sqrt{d_k}}\right) V_h^{(l)} \tag{8}$$

$$MHPA(Z^{(l)}) = Concat(head_1, \cdots, head_H)\, W_O \tag{9}$$

여기서 $MHPA$는 식 (8)과 (9)로 정의되는 최종 멀티헤드 풀링 어텐션의 결과로, 각 헤드 $head_h$는 $Q_h^{(l)}$, $K_h^{(l)}$, $V_h^{(l)}$ 행렬 간의 내적 연산을 통해 계산된다. 이때 d_k는 Key 벡터의 차원 수를 의미하며, W_O는 모든 헤드의 어텐션 출력을 융합하기 위한 선형 변환 행렬이다. MViT는 $MHPA$의 풀링 연산을 적용함으로써, 계층이 깊어질수록 토큰 수 N을 줄이고 채널 차원 D를 확장함으로써 다양한 스케일의 전역 문맥 정보를 계층적으로 학습할 수 있게 된다. 최종 단계에서는, 클래스 토큰(Class token)을 통해 입력 전체의 전역 정보를 요약하며, 이렇게 얻어진 전역 특징 벡터는 국소 정보 분기에서 추출된 특징과 함께 글로컬 특징 융합 모듈로 전달된다.

3.4 글로컬 특징 융합 모듈

제안한 하이브리드 모델의 마지막 단계에서는 CNN 분기에서 추출된 국소 특징과 ViT 분기에서 얻은 전역 특징을 결합하여 최종적으로 DR 등급을 예측한다. 본 연구에서는 두 특징 정보를 융합하기 위해, 먼저 연결 계층을 통해 입력을 구성한 뒤, 비선형 변환과 스킵 연결을 적용하여 최종 글로컬 특징을 생성한다. 그림 2는 제안한 글로컬 특징 융합 모듈의 전체 구조를 보여준다.

먼저, 식 (10)과 같이 CNN 분기에서 추출한 국소 특징과 ViT 분기에서 추출한 전역 특징을 결합하여 하나의 입력 벡터 F_{concat}을 생성한다.

$$F_{concat} = [F_{cnn} \| F_{vit}] \tag{10}$$

여기서 F_{cnn}과 F_{vit}는 각각 CNN 및 ViT 분기에서 추출한 768차원 특징을 말한다. 그리고 기호 ‖는 채널 방향으로 두 특징을 쌓는 연산을 의미한다.

둘째, 국소 및 전역 특징 정보에 가중치를 부여하고 비선형 연산 과정을 거쳐 새로운 글로컬 특징 벡터로 표현한다.

$$H_0 = W_1 F_{concat} + b_1 \tag{11}$$

여기서 W_1은 1536차원의 결합 벡터를 768차원으로 투영하는 선형 변환 가중치이고 b_1는 편향(bias) 가중치에 해당한다. 즉, 식 (11)은 어파인 변환을 나타내며, 글로컬 벡터는 저차원의 공간으로 사상되어 정보가 압축된 시맨틱한 정보로 변환된다. 그리고 비선형 변환을 위해, 그림 2에서 보듯이 비선형 활성화 GELU 함수, 두 개의 선형 변환, 드롭아웃으로 구성된 잔차 블록을 통과하게 된다.

$$H_1 = f_{res}(H_0) \tag{12}$$

여기서 f_{res}는 그림 2의 비선형 변환을 위한 잔차 블록에 해당하며, 식 (12)를 통해 글로컬 벡터의 특징 구별력은 강화된다. 잔차 블록은 제안한 글로컬 특징 융합 모듈에서 2회 반복하도록 설계한다.

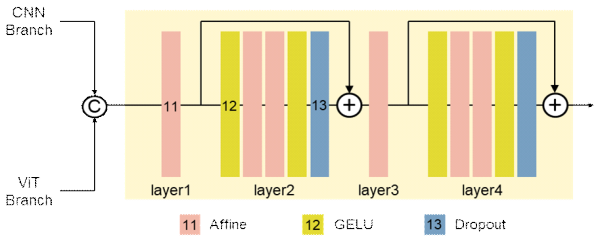

그림 2. 글로컬 특징 융합 모듈
Fig. 2. Glocal feature fusion module

3.5 분류 헤드

본 연구에서는 DR 등급 예측을 위해, 글로컬 특징 융합 모듈 이후에 분류 헤드를 추가한다. 분류 헤드는 식 (13)과 같이 하나의 선형 계층과 소프트맥스 함수로 구성된다.

$$p = Softmax(W_{proj}F_{glocal} + b_{proj}) \tag{13}$$

여기서 F_{glocal}는 글로컬 특징 융합 모듈의 출력 결과이며 W_{proj}와 b_{proj}는 어파인 변환을 위한 가중치와 편향 행렬이다. $Softmax$는 [0,1] 구간의 확률값으로 변환하기 위한 소프트맥스 함수이며, p는 최종 DR 등급 예측치에 해당한다.

IV. 실험 및 결과

4.1 실험 환경

DR 등급 분류 실험을 위해 DDR(Dataset for Diabetic Retinopathy)[18] 데이터셋을 사용하였다. 전체 13,673장의 안저 이미지 중 저품질 및 비정상 라벨 데이터를 제외한 12,522장을 실험에 활용하였으며, 총 5개 등급(정상, 경도, 중등도, 중증, 증식성)에 대해 다중 클래스 분류를 수행하였다.

그림 3은 각 등급에 해당하는 안저 이미지 예시를 나타낸다. 데이터는 총 12,522장을 기준으로 학습용 6,261장, 검증용 2,508장, 테스트용 3,753장으로 구성하였으며, 학습은 PyTorch 프레임워크를 기반으로 NVIDIA RTX 3060 GPU 환경에서 수행되었다. 입력 크기는 224×224로 설정하였고, 배치 크기는 16, 에폭은 100, 학습률은 0.0001로 설정하였으며, 최적화 기법은 Adam 옵티마이저를 사용하였다.

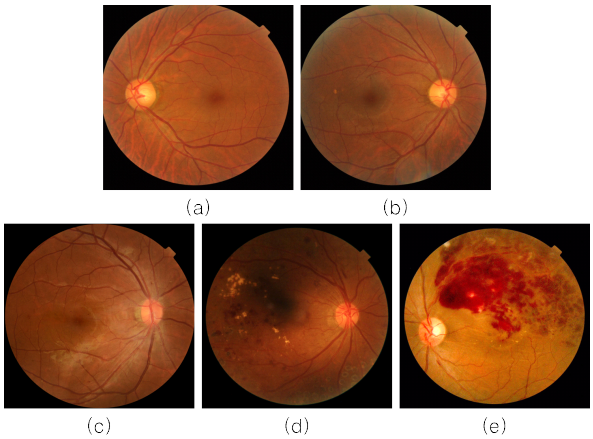

그림 3. DR 등급별 안저 이미지 예시
(a) 정상, (b) 경도, (c) 중등도, (d) 중증, (e) 증식성
Fig. 3. Sample fundus images by DR severity level (a) No DR, (b) Mild, (c) Moderate, (d) Severe, (e) Proliferative

4.2 정성적 평가

제안한 모델의 정성적 평가를 위해, 국제 임상 당뇨망막병증 분류 척도(ICDR, International Clinical DR Scale)[19]를 기준으로 CAM 시각화 결과와 병변별 정답 분할 맵을 비교하였다. ICDR은 당뇨망막병증을 경도, 중등도, 중증, 증식성 단계로 구분하며, 각 단계는 주요 병변 유형에 따라 정의된다. 경도 단계에서는 미세동맥류가 주로 관찰되고, 중등도 단계에서는 경성 삼출물, 출혈, 미세동맥류가 함께 나타난다. 중증 단계는 출혈이 두드러지며, 증식성 단계는 이미지 전반에 걸친 광범위한 출혈이 특징적이다. 그림 4는 중증, 증식성 단계의 안저 이미지에 대한 CAM 시각화 결과와 정답 병변 분할맵을 보여준다. 첫 번째 열은 원본 안저 이미지이고, 두 번째부터 네 번째 열은 각각 ResNet-50, MViT, 그리고 제안한 모델의 CAM 시각화 결과이다. 이후 각 열은 순서대로 경성 삼출물, 출혈, 미세동맥류, 연성 삼출물의 정답 분할맵을 나타낸다. 또한, 첫 번째와 두 번째 행은 중증, 세 번째와 마지막 행은 증식성 단계를 의미한다. 참고로, CAM 시각화에서 빨간색 계통은 모델이 병변이 존재할 확률이 높은 관심 영역으로 예측한 부분이고, 파란색 계통은 모델이 정보가 없는 배경 영역으로 간주한 부분을 나타낸다. 그림 4의 실험 결과에서, 제안한 모델은 중증 및 증식성 단계 모두에서 ResNet-50보다 병변을 더 연속적으로 포착해 단편적인 활성화 경향을 줄였으며,

MViT보다 작은 병변에 대한 표현이 더 뚜렷하게 나타났다. 특히, 국소적·전역적 정보를 동시에 활용함으로써 큰 병변과 세부 병변을 모두 보존하고, DR 진행 단계에 따른 병변의 특성을 균형 있게 반영하였다. 이러한 시각화 결과는 정답 분할맵과의 높은 일치도를 통해 확인되며, 이는 병변의 분포와 확산 정도를 명확히 드러내어 DR 단계 구분과 진행 정도를 시각적으로 해석하는 데 임상적으로도 유의미한 근거를 제공한다.

4.3 정량적 평가

정량적 평가를 위해, 이미지 분류 분야에서 널리 활용되는 대표적 모델들과 성능을 비교하였다. 비교 대상에는 CNN 계열, ViT 계열, 그리고 하이브리드 계열 모델이 포함된다. 구체적으로는 ResNet-50[1], Inception V3[2], CrossViT[4], MViT[5], CoAtNet[16], FastViT[17], NextViT[20]을 선정하였다. 평가 척도로는 전체 테스트 이미지 중 올바르게 분류한 샘플의 비율을 나타내는 정인식률(Correct Recognition Rate)과 클래스 문제에서 예측 등급과 실제 등급 간의 거리차를 제곱 가중치로 반영한 통계적 지표인 QWK(Quadratic Weighted Kappa)를 사용하였다.

정인식률과 QWK는 다음과 같이 계산된다.

$$CRR = \frac{c}{N} \tag{14}$$

여기서 N은 전체 샘플 수이며 c는 예측 등급과 정답 등급이 일치하는 개수이다.

$$QWK = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \tag{15}$$

여기서 i와 j는 각각 실제 등급과 예측 등급을 의미한다. 그리고 $w_{i,j}$는 가중치 값으로서, 예측 등급과 실제 등급 차이의 제곱 값에 비례한다. 그리고 $O_{i,j}$는 실제 등급 i를 예측 등급 j로 분류한 개수이며 $E_{i,j}$는 실제 등급의 주변합(Marginal)을 예측 등급의 주변합을 곱해서 전체 샘플 개수로 나눈 값이다[21]. 즉, QWK는 DR 예측과 같이 순서형 클래스 문제에서 예측 등급과 실제 등급 간의 거리차를 반영한 평가 척도라 해석할 수 있다.

표 1은 각 모델의 두 지표에 대한 평가 결과를 보여준다. 표에서 확인할 수 있듯이, 제안한 모델은 정확도 83.26%와 QWK 86.62%를 기록하며 모든 비교 모델 중 가장 우수한 성능을 보였다. 이는 병변의 미세한 국소 특징을 정밀하게 추출하는 CNN 분기와, 병변의 공간적 분포와 단계적 진행 양상을 포괄적으로 학습하는 ViT 분기를 병렬로 결합함으로써, DR 등급 분류에 필요한 국소 및 전역 정보가 유기적으로 통합된 결과로 해석된다.

표 1. 정량적 평가
Table 1. Quantitative evaluation

Models	Correct recognition rate	QWK
ResNet-50 [1]	73.76%	73.31%
Inception V3 [2]	76.88%	80.55%
CrossViT [4]	72.29%	71.43%
MViT [5]	82.07%	85.22%
CoAtNet [16]	79.28%	76.69%
FastViT [17]	74.89%	74.62%
NextViT [20]	76.64%	76.69%
Proposed model	**83.26%**	**86.62%**

특히, 기존 하이브리드 구조(CoAtNet, FastViT, NextViT)보다 높은 정인식률과 QWK를 보인 것은, 멀티스케일 기반 병렬 구조와 글로컬 특징 융합 설계가 DR 병변의 다양한 스케일 표현에 효과적으로 작용한 결과로 판단된다. 그림 4에서 확인할 수 있듯이, DR 병변은 매우 미세한 혈관 변화부터 비교적 큰 출혈·삼출 병변까지 다양한 크기와 형태를 갖기 때문에, 멀티스케일 표현 능력을 갖춘 MViT 구조의 적용은 DR 등급 분류에 매우 적합하다. 또한, MViT와 CNN 계열인 ResNet의 특징 맵을 병렬로 추출한 뒤 융합하는 전략은 서로 다른 수용영역(Receptive field)과 표현적 강점을 결합함으로써 성능을 추가적으로 향상시킨 것으로 분석된다. 특히, 본 연구에서 제안한 모델의 QWK가 80% 이상을 기록했다는 점은, 예측 등급과 실제 등급 간 일치도가 매우 강한 수준에 해당함을 의미한다. 다시 말해, 제안한 모델은 단순한 정인식률 향상을 넘어, 순서형 클래스 분류에서의 신뢰성과 일관성 측면에서도 우수한 성능을 입증한 것으로 해석할 수 있다.

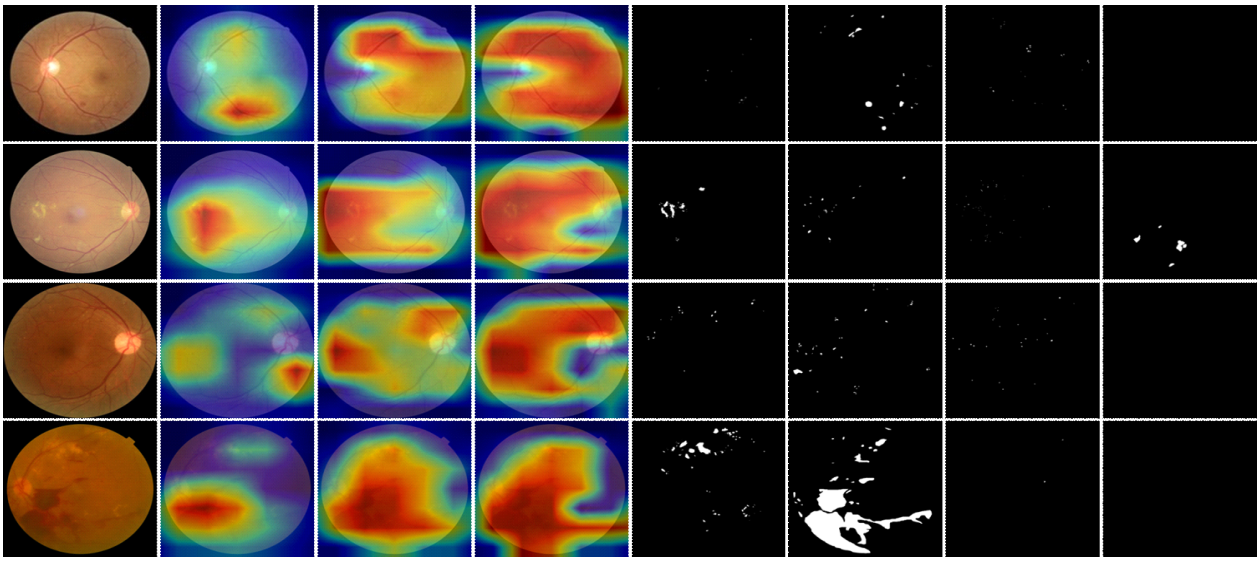

그림 4. 실험 결과 : 클래스별 원본 이미지 (첫 번째 열), ResNet-50의 CAM 이미지 (두 번째 열), MViT의 CAM 이미지 (세 번째 열), 제안한 모델의 CAM 이미지 (네 번째 열), 경성 삼출물 분할맵 (다섯 번째 열), 출혈 분할맵 (여섯 번째 열), 미세동맥류 분할맵 (일곱 번째 열), 연성 삼출물 분할맵 (마지막 열)

Fig. 4. Experimental results : Original images for each class (first column), CAM images of ResNet-50 (second column), CAM images of MViT (third column), CAM images of the proposed model (fourth column), Hard exudates segmentation images (fifth column), Hemorrhage segmentation images (sixth column), Microaneurysm segmentation images (seventh column), and Soft exudates segmentation images (last column)

V. 결 론

본 논문에서는 당뇨망막병증 등급 분류의 정확도를 향상시키기 위해, 합성곱 신경망과 비전 트랜스포머의 장점을 결합한 병렬 하이브리드 모델을 제안하였다. 제안한 모델은 ResNet-50 기반의 국소 정보 학습 분기와 MViT 기반의 전역 정보 학습 분기를 병렬로 구성하고, 두 분기에서 추출된 특징을 글로컬 특징 융합 모듈에서 결합하여 등급을 예측한다. 실험은 DDR 데이터셋을 활용하여 수행되었으며, 성능 평가는 정인식률과 QWK 지표를 기준으로 진행하였다. 그 결과, 제안한 모델은 정인식률 83.26%, QWK 86.62%를 기록하며 기존 합성곱 신경망 기반 모델과 비전 트랜스포머 기반 모델을 모두 상회하는 성능을 보였다. 특히, 기존 하이브리드 모델들과 비교했을 때에도 가장 높은 성능을 달성함으로써, 국소적 세부 정보와 전역적 문맥 정보를 병렬적으로 학습하는 구조의 효과성을 확인할 수 있었다. 또한, 제안한 모델은 입력 크기 224×224 기준으로 이미지당 약 14.33 GFLOPs의 연산량과 평균 74.8 ms의 추론 시간을 기록하였으며, 이는 초당 약 13장의 이미지를 처리할 수 있는 속도에 해당한다. 이러한 결과는 단일 GPU에서도 원활한 실시간 추론이 가능함을 보여주며, 정확도와 연산 효율성의 균형을 갖춘 모델의 실용적 가능성을 시사한다.

Acknowledgement

2025년도 한국정보기술학회 하계종합학술대회에서 발표한 논문(당뇨망막병증 등급 분류를 위한 컴퓨터 비전 모델 성능 비교)[22]을 확장한 것임

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", Proc. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 770-778, Jun. 2016. https://doi.org/10.1109/CVPR.2016.90.

[2] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision", Proc. IEEE

Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 2818-2826, Jun. 2016. https://doi.org/10.1109/CVPR.2016.308.

[3] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks", Proc. International Conference on Machine Learning, Long Beach, CA, USA, pp. 6105-6114, Jun. 2019. https://doi.org/10.48550/arXiv.1905.11946.

[4] C.-F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification", Proc. IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, pp. 347-356, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.00041.

[5] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, and J. Malik, "Multiscale vision transformers", Proc. IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, pp. 6804-6815, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.00675.

[6] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks", Proc. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 2261-2269, Jul. 2017. https://doi.org/10.1109/CVPR.2017.243.

[7] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection", Proc. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 2117-2125, Jul. 2017. https://doi.org/10.1109/CVPR.2017.106.

[8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications", arXiv preprint arXiv:1704.04861, Apr. 2017. https://doi.org/10.48550/arXiv.1704.04861.

[9] F. Zang and H. Ma, "CRA-Net: Transformer guided category-relation attention network for diabetic retinopathy grading", Computers in Biology and Medicine, Vol. 170, pp. 107993, Mar. 2024. https://doi.org/10.1016/j.compbiomed.2024.107993.

[10] H.-J. Yu, C.-H. Son, and D. H. Lee, "Apple leaf disease identification through region-of-interest aware deep convolutional neural network", Journal of Imaging Science and Technology, Vol. 64, No. 2, pp. 20507-1-20507-10, Jan. 2020. https://doi.org/10.2352/j.imagingsci.technol.2020.64.2.020507.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", arXiv preprint arXiv:1706.03762, Aug. 2023. https://doi.org/10.48550/arXiv.1706.03762.

[12] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions", Proc. IEEE/CVF International Conference on Computer Vision, Montreal, Canada, pp. 568-578, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.00061.

[13] G.-E. Kim, C.-H. Son, and S. Lee, "ROI-aware multiscale cross-attention vision transformer for pest image identification", Computers and Electronics in Agriculture, Vol. 218, pp. 110732, Sep. 2025. https://doi.org/10.1016/j.compag.2025.110732.

[14] M. Z. Atwany, A. H. Sahyoun, and M. Yaqub, "Deep learning techniques for diabetic retinopathy classification: A survey", IEEE Access, Vol. 10, pp. 28642-28655, Mar. 2022. https://doi.org/10.1109/ACCESS.2022.3157632.

[15] 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 16×16 words: Transformers for image recognition at scale", arXiv preprint arXiv:2010.11929, Sep. 2021. https://doi.org/10.48550/arXiv.2010.11929.

[16] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for

all data sizes", Proc. Advances in Neural Information Processing Systems, Online, pp. 3965-3977, Dec. 2021. https://doi.org/10.48550/arXiv.2106.04803.

[17] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "FastViT: A fast hybrid vision transformer using structural reparameterization", Proc. IEEE/CVF International Conference on Computer Vision, Paris, France, pp. 5762-5772, Oct. 2023. https://doi.org/10.1109/ICCV51070.2023.00532.

[18] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening", Information Sciences, Vol. 501, pp. 511-522, Oct. 2019. https://doi.org/10.1016/j.ins.2019.06.011.

[19] C. P. Wilkinson, F. L. Ferris III, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, and J. T. Verdaguer, "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales", Ophthalmology, Vol. 110, No. 9, pp. 1677-1682, Sep. 2003. https://doi.org/10.1016/S0161-6420(03)00475-5.

[20] J. Li, X. Xia, W. Li, H. Li, X. Wang, X. Xiao, R. Wang, M. Zheng, and X. Pan, "Next-ViT: Next generation vision transformer for efficient deployment in realistic industrial scenarios", arXiv preprint arXiv:2207.05501, Aug. 2022. https://doi.org/10.48550/arXiv.2207.05501.

[21] M. J. Warrens, A. de Raadt, R. J. Bosker, and H. A. L. Kiers, "Weighted Kappa for Interobserver Agreement and Missing Data", Machine Learning & Knowledge Extraction, Vol. 7, Feb. 2025. https://doi.org/10.3390/make7010018.

[22] J. H. Do and C. H. Son, "Performance Comparison of Computer Vision Models for Diabetic Retinopathy Grading Classification", Proc. Korea Information Technology Conference, Jeju, Korea, pp. 678-681, Jun. 2025.

저자소개

도 지 현 (Ji-Hyun Do)

2022년 3월 ~ 현재 : 국립군산대학교 소프트웨어학과 학사과정

관심분야 : 컴퓨터 비전, 영상처리, 기계학습, 딥 러닝

손 창 환 (Chang-Hwan Son)

2002년 2월 : 경북대학교 전자전기공학부(공학사)

2004년 2월 : 경북대학교 전자공학과(공학석사)

2008년 8월 : 경북대학교 전자공학과(공학박사)

2017년 4월 ~ 현재 : 국립군산대학교 소프트웨어학과 교수

관심분야 : 컴퓨터 비전, 영상처리, 기계학습, 딥 러닝