JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 11, pp. 523-530, November 2026
https://doi.org/10.9708/jksci.2026.31.11.523

A Deep Learning-Based Learning Engagement Analysis Framework Using Multimodal Observation Data

Young-Bok Cho*

*Professor, Dept. of Computer Education, Gyeongkok National University, Andong, Korea

[Abstract]

In this paper, we propose a multimodal deep learning-based learning engagement analysis framework that combines learners' facial video with learning-environment acoustics, utilizing the "Learning Attitude and Disposition Observation Data" released on AI-Hub. The proposed framework integrates a visual perception stream composed of MobileNetV3 and Bi-LSTM with an MFCC-based 1D-CNN acoustic perception stream through a late fusion structure, and classifies learner states into five classes. In experiments, the proposed model achieved an accuracy of 92.4% and a macro F1-score of 0.919, outperforming the vision-only single-modality model (85.7%) by 6.7 percentage points. In particular, the misclassification between "unfocused-bored" and "drowsy"—which is difficult to distinguish from visual information alone—was substantially reduced through the fusion of acoustic features. In future work, this study can serve as a core technology for providing personalized feedback in AI Digital Textbook (AIDT) and smart learning systems.

▸ **Key words**: Learning Attitude Analysis, Multimodal Deep Learning, Learning Engagement, AI-Hub, Educational AI, AI Digital Textbook

[요 약]

본 논문에서는 AI-Hub에 공개된 학습태도 및 성향 관찰 데이터를 활용하여 학습자의 안면 영상과 학습 환경 음향을 결합한 멀티모달 딥러닝 기반 학습태도·행동 몰입 상태 분류 프레임워크를 제안한다. 제안 프레임워크는 MobileNetV3와 Bi-LSTM으로 구성된 시각인지 스트림과 MFCC 기반 1D-CNN 청각 인지 스트림을 학습태도·행동 몰입 상태 분류 프레임워크 구조로 결합하고 학습 상태를 5개 클래스로 분류해 실험한 결과 제안 모델의 정확도는 92.4%, 매크로 F1-score는 0.919으로 영상 단일 모달리티 모델(85.7%) 대비 6.7%p 향상된 성능을 보였다. 특히 시각 정보만으로는 구분이 어려운 '비집중-지루함'과 '졸음' 간 오분류가 음향 특징의 결합을 통해 크게 감소하였다. 향후 본 연구는 AIDT 및 스마트 학습 시스템에서 맞춤형 피드백을 제공하기 위한 원천 기술로 활용될 수 있을 것이다.

▸ **주제어**: 학습태도분석, 멀티모달딥러닝, 학습태도·행동 몰입 상태 분류 프레임워크, 행동몰입, AI-Hub, 교육인공지능

• First Author: Young-Bok Cho, Corresponding Author: Young-Bok Cho
*Young-Bok Cho (ybcho@gknu.ac.kr), Dept. of Computer Education, Gyeongkok National University
• Received: 2026. 07. 23, Revised: 2026. 10. 20, Accepted: 2026. 10. 22.

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

교육 현장은 2025년 인공지능 디지털교과서(AIDT)의 도입을 기점으로 학습자 맞춤형 교육 체계로 전환되고 있다. 그러나 기존 스마트교육 플랫폼은 진도율이나 정답률 같은 사후적 정량 지표에 의존하여, 학습 중 발생하는 집중력 저하와 인지적 과부하를 실시간으로 포착하기 어렵다. 학업 성취의 핵심 변인인 학습 몰입도 역시 전통적으로 설문이나 정성적 관찰에 의존해 왔으나[1,2] 실시간 피드백 연동에 한계가 있다. 이를 극복하기 위해 시·청각 관찰 신호 기반의 자동 측정 연구가 시도되고 있으나, 초기 시각 중심 연구는 조명이나 착용물에 취약할 뿐만 아니라 외형이 유사한 '지루함'과 '졸음'을 명확히 구분하지 못하는 한계가 있다[3]. 음향 신호는 신체 마찰음이나 호흡에 따른 에너지 변동 등 시각적으로 포착하기 어려운 보완적 단서를 제공할 수 있다. 다만 본 연구는 이러한 음향 단서가 특정 상태를 결정적으로'입증'한다고 주장하지 않으며, 시각 정보에 대한 상보적 단서로서의 가능성에 초점을 둔다.

본 연구의 대상은 학습 성취나 내적·정서적 몰입 전체가 아니라, 안면 영상과 환경 음향으로 외형적으로 관찰 가능한 학습태도 및 행동적 몰입 관련 상태의 분류로 한정한다. 따라서 분류 결과는 학습자의 내적 상태에 대한 확정적 판단이 아니라 보조적 참고 정보로 해석되어야 한다.

본 연구는 AI-Hub에 공개된 71,715개의 학습태도 및 성향 관찰 데이터를 활용하여 다음 세 가지를 기여한다. 첫째, 해당 데이터에 대한 재현 가능한 멀티모달 전처리·분할 파이프라인을 정의하고 피험자 단위 분할로 평가 독립성을 확보한다. 둘째, 경량 시각·음향 스트림을 게이팅 특징수준 융합으로 결합한 분류 프레임워크를 제안하고, 단순 연결·평균·고정 가중합·게이트 제거 등과의 어블레이션으로 게이팅의 기여를 분리 검증한다. 셋째, 음향 에너지 포락선의 저주파 변동을 정량 분석하여'비집중-지루함'과'졸음' 판별에 대한 음향의 상보적 효과를 통계적으로 검토한다.

본 논문의 구성은 다음과 같다. 2장에서는 멀티모달 학습의 개념과 연구 동향, 학습 몰입도 분석에 관한 기존 연구를 고찰한다. 3장에서는 제안하는 프레임워크의 데이터 개요, 전처리 절차, 모델 구조를 기술한다. 4장에서는 실험 환경과 결과를 분석하고, 5장에서 결론과 향후 연구 방향을 제시한다.

II. Related Works

1. Concepts and Research Trends in Multimodal Learning

멀티모달 학습은 서로 다른 감각 채널이나 이질적 신호원에서 얻은 데이터를 공동으로 표현·추론하는 기계학습 패러다임이다[5]. 초기의 오디오-비주얼 공유 표현 학습[6]에서 출발하여 트랜스포머와 교차 어텐션의 도입으로 모달리티 간 장거리 의존성 모델링이 발전하였다[7]. Baltrušaitis 등은 멀티모달 연구의 과제를 표현·변환·정렬·융합·공동학습으로 체계화하고[5], 이 중 융합은 과업 성능에 직접적 영향을 미친다.

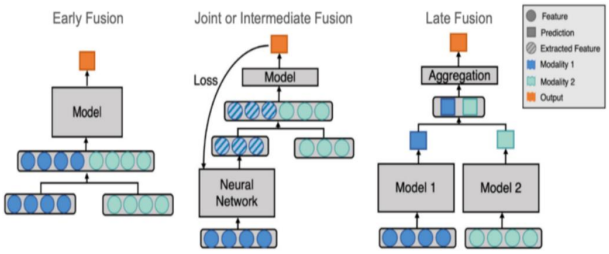


Fig. 1. Type of Multimodal Deep Learning

[그림1]과 같이 융합은 결합 시점에 따라 구분된다. 조기 융합(early fusion)은 입력·저수준 특징을 결합하여 미세한 상호작용을 학습하나 동기화에 민감하고 결측에 취약하다. 반면 결정 단계에서 결합하는 후기 융합(late fusion)은 설계가 자유롭고 결측에 강건하다. 본 연구가 제안하는 방식은 각 모달리티를 독립 인코딩한 뒤 결정 직전이 아니라 공통 특징 공간(128차원)에서 게이팅 가중으로 결합하는 특징수준 융합에 해당하므로, 이를 게이팅 특징수준 융합으로 명명한다.교육 현장의 저사양 단말을 고려하면 대규모 모델보다 MobileNet 계열 경량 백본이 현실적 대안으로 평가된다[8].

2. Prior Studies on Learning Engagement Analysis

학습 몰입도 자동 측정 연구는 Whitehill 등이 안면 영상 기반의 4단계 척도를 정의한 이후[3], D'Mello와 Graesser가 정서 순환 모형[9]으로 확장되었다. 데이터셋 측면에서는 Gupta 등의 DAiSEE[10]가 널리 쓰이나라 라벨 불균형과 온라인 시청 상황에 국한된다는 한계가 있다. 음향을 활용한 접근으로 Abedi와 Khan은ResNet-TCN으로 시간 문맥을 강화했으나[11,12] 단일 모달리티 중심이라는 제약이 있다. 국내에서는 로그 기반 지식추적에 편중

www.kci.go.kr

되어 행동 관찰 데이터 활용이 초기 단계이다. 이에 본 연구는 국내 교육 환경에서 수집된 멀티모달 관찰 데이터[4]를 게이팅 특징수준 융합으로 결합하여 관찰 가능한 학습 태도·행동 상태를 분류한다.

III. Proposed Deep Learning-Based Learning Engagement Analysis Framework

1. Dataset Overview

본 연구에서 활용한 「학습태도 및 성향 관찰 데이터」는 실제 학습 과정의 멀티모달 데이터를AI-Hub에 구축한 공개 데이터셋이다[4]. 원시 소스와 메타데이터 쌍으로 구성되며 안면 바운딩 박스, 프레임 단위 상태 레이블, 소음 변동, 인구통계 정보를 포함한다. 데이터셋은 영상·음성·생체신호·로그를 포함하나, 본 연구는 그중 안면 영상과 환경음향 두 모달리티만사용하며 나머지는 사용하지 않았다. 본 연구에서 학습몰입은 상위 개념이며, 5개 클래스는 인지 축(집중/비집중)과 정서·생리 축(흥미/차분/지루함/졸음)이 결합된2차원 관찰 상태 체계로 정의한다. 학습태도, 정서 상태 등의 용어는 이 관찰 상태를 가리키는 것으로 통일하며, 학습 성취나 내적 몰입 전체를 지칭하지 않는다. 학습자 상태는 [표 1]과 같이 총 5개의 레이블로 정의된다.

Table 1. Definition of five target classes for learning attitude analysis

Category	Class	State and Observed Characteristics
Focused	Focused-Interested	Face oriented toward the screen with an animated expression; smiling and raised eyebrows indicating positive affect
Focused	Focused-Calm	Steady gaze with stable breathing and posture; few facial changes and regular movement
Unfocused	Unfocused-Calm	Gaze deviates from the screen while body movement remains low; early stage of attention drift
Unfocused	Unfocused-Bored	Frequent posture changes, yawning and dissatisfied expressions; accompanied by irregular friction noise
Sleep	Drowsy	Prolonged eye closure and head drooping; reduced acoustic energy with regular breathing sounds

본 연구에서 . 세그먼트 추출 및 피험자 단위 분할은 영상과 오디오가 동기화되고 레이블 검수를 통과한 5초 길이의 멀티모달 페어를 추출하였다. 각 페어는 프레임 시퀀스와 음향 파형이 쌍을 이룬다. 모델이 개인의 외형·배경·촬영 환경을 암기하여 성능이 과대평가되는 것을 방지하기 위해, 세그먼트 생성 이전 단계에서 피험자 단위로 학습·검증·시험 세트를 분할하였다.동일 피험자의 서로 다른 세션이나 동일 원본 영상에서 생성된 인접 구간이였다. [표 2]는 피험자 단위 분할 결과와 세트별 클래스 분포를 나타낸다. 실제 학습 환경의 자연스러운 분포를 유지하기 위해 인위적 균등화는 지양하고, 학습 시 클래스 가중치로 불균형을 보정하였다.

Table 2. Class-wise data distribution and stratified split

Class	Total	Train (80%)	Valid (10%)	Test (10%)	Ratio
Focused-Interested	12,000	9,600	1,200	1,200	24.0%
Focused-Calm	13,500	10,800	1,350	1,350	27.0%
Unfocused-Calm	9,000	7,200	900	900	18.0%
Unfocused-Bored	9,000	7,200	900	900	18.0%
Drowsy	6,500	5,200	650	650	13.0%
Total	50,000	40,000	5,000	5,000	100.0%

또한 누출 검증 결과 피험자, 세션, 원본영상 세 수준 모두에서 학습/검증/시험 세트 간 교집합이 존재하지 않음을 확인하였다.

2. Data Preprocessing

원시 멀티모달 데이터는 해상도, 프레임률, 샘플링 주파수가 상이하므로, 모델 입력을 위한 표준화된 정규화 과정이 필수적이다. [그림2]는 시각 및 청각 스트림에 대해 각각 적용한 전처리 파이프라인을 도식화한 것이다.

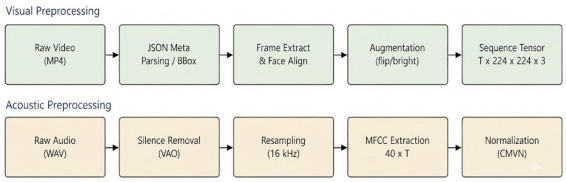


Fig. 2. Multimodal data preprocessing pipeline

2-1. Visual Stream Preprocessing

JSON 메타데이터의 안면 바운딩 박스 좌표를 기반으로 얼굴 영역을 추출한다. 좌표가 누락되거나 신뢰도가 임계값 미만인 프레임은 MTCNN 검출기로 보완하고[14], 미검출 시 직전 유효 프레임으로 대체한다. 추출된 영역은 양

안 기준 정렬 및 224×224 크기로 리사이즈한 후, ImageNet 통계를 이용해 채널별 정규화를 수행한다. 시퀀스는 태블릿 단말의 연산 부하를 고려해 5초 세그먼트에서 5fps로 균등 샘플링하여 T=25 프레임으로 구성한다. 학습 시에는 동일한 증강 파라미터 적용을 위해 좌우 반전, 밝기·대비 지터링(±20%), 랜덤 크롭(스케일 0.9~1.0)을 적용하였다

2-2. Acoustic Stream Preprocessing

오디오 신호는 16kHz 단채널로 리샘플링하고 계수 0.97의 프리엠퍼시스 필터로 고주파 성분을 보강한 뒤, 25ms 윈도우와 10ms 홉 길이의 단시간 푸리에 변환(STFT) 및 40개 멜 필터뱅크를 거쳐 MFCC를 산출한다[14]. 정적 계수 40차에 1·2차 델타 계수를 추가하여 최종 120차원 특징 맵을 구성하고 무음 구간이 많은 학습 환경 특성을 고려해 에너지 기반 VAD(Voice Activity Detection)로 압축하되, 졸음 판별 단서인 저에너지 호흡음의 소실을 방지하고자 임계값을 보수적으로 설정했다. 마지막으로 화자 및 환경에 따른 편차를 제거하기 위해 켑스트럼 평균·분산 정규화(CMVN)를 적용했다. [표3]은 전처리 결과 확정된 입력 사양을 나타낸 것이다.

Table 3. Model input specification after preprocessing

Modality	Item	Setting
Visual	Segment length / frame rate	5s / 5fps (T = 25)
Visual	Input resolution	224 × 224 × 3
Visual	Normalization	ImageNet per-channel mean/std
Acoustic	Sampling rate / channel	16kHz / mono
Acoustic	Window / hop length	25ms / 10ms(Hamming)
Acoustic	Feature dimension	MFCC 40 + Δ 40 + ΔΔ 40 = 120
Common	Label format	5-class one-hot

3. Gating Feature Level Convergence Framework

실제 교육 현장에서는 마이크 미사용·카메라 차단 등으로 모달리티 결측이 빈번하므로 단일 스트림만으로도 추론 가능한 구조가 요구된다. 영상과 음향은 시간 해상도와 통계적 성질이 달라 저수준 직접 결합이 어렵다. 이에 각 스트림을 독립 인코딩한 뒤 공통 특징 공간에서 게이팅 가중으로 결합한다. [그림3]은 제안 구조이다.

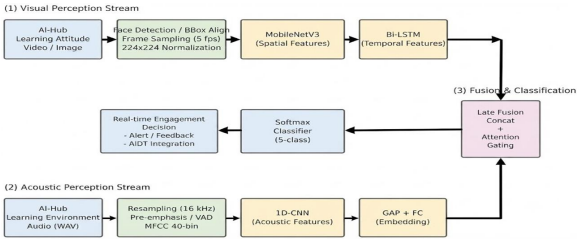


Fig. 3. Proposed multimodal learning engagement analysis framework

3-1. Visual Perception Stream

학습자의 안면 프레임 시퀀스는 MobileNetV3-Small[8]을 공간 특징 추출기로 사용하여 프레임 단위 임베딩으로 변환된다. t번째 프레임의 공간 특징은 식(1)과 같다.

$$f_t = \phi_{CNN}(x_t;\theta_v),\ f_t \in R^{576} \quad (1)$$

프레임별 특징은 시간 문맥 반영을 위해 2계층 양방향 LSTM[17]에 입력된다. 최종 시각 특징은 식(2)와 같이 순방향·역방향 은닉 상태의 연결을 시간 평균하여 얻는다.

$$h_t = [LSTM\rightarrow(f_t);\ LSTM\leftarrow(f_t)] \quad (2)$$
$$z_v = \frac{1}{T}\sum_{t=1}^{T} h_t,\ z_v \in R^{512}$$

3-2. Acoustic Perception Stream

120차원 MFCC 특징 맵은 4개의 1차원 합성곱 블록으로 구성된 1D-CNN에 입력된다. 각 블록은 커널 크기 5의 합성곱, 배치 정규화, ReLU 활성화, 최대 풀링으로 구성하고, 채널 수는 64→128→256→256으로 확장된다. 마지막 블록의 출력에 전역 평균 풀링을 적용하여 시간 축을 소거함으로써 식(3)의 고정 길이 음향 특징 벡터를 얻는다.

$$z_a = GAP(\psi_{1DCNN}(MFCC(a);\theta_a)),\ z_a \in R^{256} \quad (3)$$

3-3. Gating Feature Level Convergence and Classification

두 특징은 차원과 스케일이 다르므로, 식(4)와 같이 각각 완전연결 계층으로 동일한 128차원 잠재 공간에 사영한다.

$$z_v' = W_v z_v + b_v,\ \ z_a' = W_a z_a + b_a\ \ z_v', z_a' \in R^{128} \quad (4)$$

모달리티별 기여는 조명 조건이나 주변 소음에 따라 매 순간 달라지기 때문에 이를 반영하기 위해 식(5)와 같은 두 사영 특징을 연결한 벡터로부터 게이팅 어텐션 가중치 α를 산출한다. α는 두 사영 특징을 연결한 벡터로부터 산출되며 시그모이드 함수를 거쳐 [0, 1] 범위의 원소별 가중치로 정규화된다. 이는 특정 모달리티의 품질이 낮을 때

www.kci.go.kr

해당 기여도를 자동으로 축소하는 역할을 하며 ⊙는 원소별 곱을 의미한다.

$$\alpha = \sigma\left(W_g\left[z_v^{'}; z_a^{'}\right] + b_g\right), \alpha \in R^{128} \quad (5)$$
$$z_f = \alpha \odot z_v^{'} + (1-\alpha) \odot z_\alpha^{'}$$
$$\hat{y} = softmax(W_o z_f + b_o)$$

본 연구의 게이트α는 별도의 품질 레이블로 감독되지 않으며, 분류 손실을 최소화하도록 학습된 모달리티별 특징 가중이다. 따라서 (식5)의 α가 신호 품질이나 신뢰도 자체를 측정한다고 단정하지 않는다. 조명·소음 수준, 얼굴 검출 신뢰도, 음향 SNR과 게이트 값의 관계 분석은 향후 과제로 남긴다. 손실 함수는 클래스 불균형 보정을 위해 표본 수의 역수에 비례하는 가중치 w_c를 적용한 가중 교차 엔트로피를 사용한다(식6). y_c는 정답 원-핫 성분, $\hat{y}_c$는 클래스 c의 예측 확률이다.

$$L = -\sum_{c=1}^{5} w_c \cdot y_c \cdot \log \hat{y}_c \quad (6)$$

융합 특징 z_f는 드롭아웃(p = 0.3)을 거쳐 최종 완전연결 계층에 입력되고, 소프트맥스 활성화 함수를 통해 5개 클래스에 대한 확률 분포로 변환된다. 손실 함수는 클래스 불균형을 보정하기 위해 표본 수의 역수에 비례하는 가중치 w_c를 적용한 가중 교차 엔트로피를 사용한다. 식(6)의 y_c는 정답 레이블의 원-핫 성분, $\hat{y}_c$는 클래스 c에 대한 예측 확률을 의미한다. 추론 단계에서는 5초 세그먼트 단위로 산출된 확률 분포를 3개 세그먼트 이동 평균으로 평활화하여 순간적인 오탐을 억제하고, 동일 상태가 연속 2회 이상 관측될 때에만 교수자 대시보드에 경고를 전달한다.

4. Ethical Considerations and Privacy Protection

안면 영상과 음향은 민감 개인정보이므로 본 연구는 다음 원칙을 전제한다. 첫째, 원본 영상·음성은 단말(엣지)에서 특징 벡터로 변환한 뒤 즉시 파기하고 서버로 원본을 전송하지 않는 처리 구조를 지향한다. 둘째, 미성년 학습자를 포함하므로 학생 및 보호자의 사전 동의 절차를 준수하고 수집 목적·범위를 고지한다. 셋째, 분류 결과는 학습자에 대한 확정적 판단이 아니라 보조적 참고 정보로만 활용하며, 오분류로 인한 낙인 효과를 방지하기 위해 개별 학습자에 대한 단독 평가·공개를 지양하고 교수자 판단을 보완하는 용도로 제한한다. 넷째, 실제 적용 시 「개인정보 보호법」상의 규정과 정보주체 동의 요건을 따른다.

IV. Experiment and Analysis

1. Experimental Setup and Hyperparameters

실험은 Intel Core i9-12900K CPU, NVIDIA RTX 3080 GPU(10GB), 64GB RAM 환경에서, Python 3.10과 PyTorch 2.1을 사용하였다. 학습에 적용한 주요 하이퍼파라미터는 [표 4]와 같이 최적 설정은 검증셋에 대한 매크로 F1-score를 기준으로 탐색하고, 조기 종료(patience=5)를 적용해 과적합을 방지하였고 모든 실험은 서로 다른 난수 시드(42, 1, 2)로 3회 반복하였으며, 결과는 평균±표준편차로 나타낸다. 비교 대상 모델은 동일한 시드·에폭·최적화기·데이터 분할을 사용하여 공정 비교를 보장하였다.

Table 4. Training hyperparameter settings

Hyperparameter	Value	Hyperparameter	Value
Optimizer	AdamW	Batch size	32
Initial learning rate	1×10^{-4}	LR scheduler	Cosine Annealing
Weight decay	1×10^{-5}	Max epochs	30 (early stop)
Dropout	0.3	Loss function	Weighted Cross-Entropy
LSTM hidden dim	256 (2-layer, Bi)	Fusion dim	128
Repeating Seed	42, 1, 2 (3rd Time)	Modality dropout	0.3 (Missing data experiment)

평가 지표는 정확도와 함께 클래스 불균형을 고려한 매크로 정밀도, 재현율, F1-score로 모든 실험은 서로 다른 난수 시드로 3회 반복한 뒤 평균값을 측정한 결과 [그림4]와 같다.

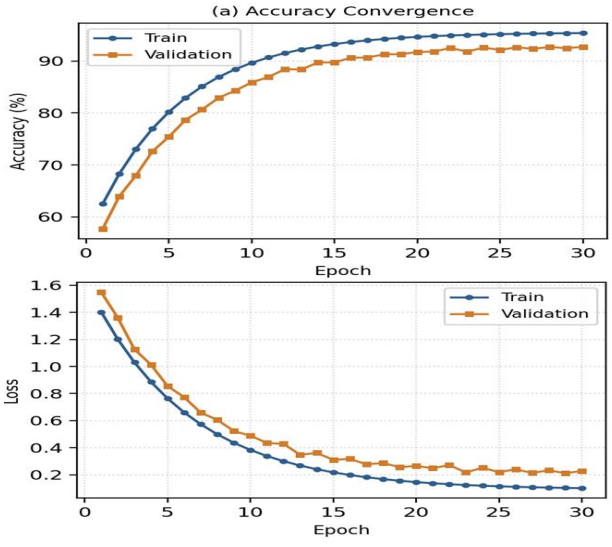


Fig. 4. Training convergence curves of the proposed model

www.kci.go.kr

[그림 4]에서 제안 모델의 에폭별 정확도 및 손실 변화를 도식화한 것으로 검증 정확도는 약 15에포크 이후 완만하게 수렴하고, 학습 손실과 검증 손실의 간격이 0.12 이내로 유지되어 심각한 과적합은 나타나지 않았다.

2. Performance Comparison by Modality

제안하는 게이팅의 기여를 분리 검증하기 위해, 단일 모달리티와 조기 융합에 더하여 단순 특징 연결, 평균, 로짓 평균, 고정 가중합, 게이트 제거를 포함한 어블레이션을 [표 5]에 제시한다. 모달리티 및 융합 전략별 성능을 비교한 결과이다. 단일 모달리티에서 음성 기반 1D-CNN은 Accuracy 64.2%, macro F1 0.604로 가장 낮은 성능을 보였다. 영상 기반에서는 MobileNetV3 단독(Accuracy 82.4%, F1 0.809)보다 Bi-LSTM을 결합한 모델(85.7%, F1 0.844)이 높아, 프레임 간 시간적 문맥 정보가 학습 성향 판별에 유효함을 확인하였다.

Table 5. Performance comparison by modality and fusion strategy

Model	Accuracy	Precision	Recall	F1-score
Audio-Only (1D-CNN)	64.2%	0.611	0.598	0.604
Vision-Only (MobileNetV3)	82.4%	0.813	0.805	0.809
Vision-Only (MobileNetV3+Bi-LSTM)	85.7%	0.845	0.844	0.844
Early Fusion (feature concat)	89.1%	0.884	0.879	0.881
Mean fusion	89.4	0.873	0.868	0.870
Logit mean	88.7	0.881	0.879	0.880
Fixed weight (α=0.5)	92.1	0.887	0.875	0.881
Gated w/o gate	91.4	0.916	0.912	0.914
Proposed (Late Fusion + Attention)	92.4%	0.918	0.920	0.919

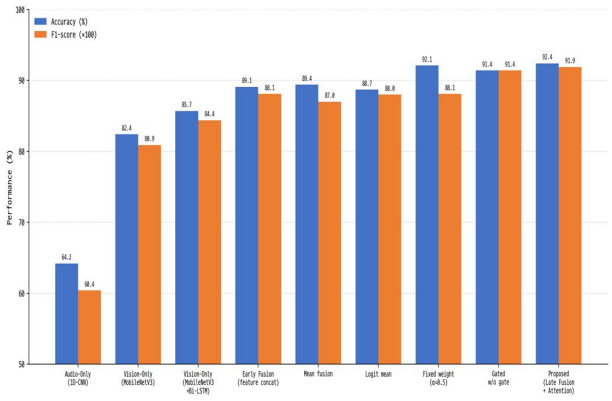


Fig. 4. Accuracy and F1-score comparison across model configurations

조기 융합(feature concat)은 Accuracy 89.1%, F1 0.881로 단일 모달리티보다 향상되었다. 후기 융합 계열의 단순 결합 방식 중 평균 융합(Accuracy 89.4%, F1 0.870)과 로짓 평균(88.7%, F1 0.880)은 조기 융합과 유사하거나 낮은 수준이었다. 이는 단순 평균이 입력에 따라 모달리티별 신뢰도 차이를 반영하지 못하기 때문으로 해석할 수 있다. 고정 가중합(α=0.5)은 Accuracy 92.1%, F1 0.881을 기록하였다. 게이트 구조를 제거한 모델(gated w/o gate)은 Accuracy 91.4%, F1 0.914로 나타났다. 제안 기법(Late Fusion + Attention 기반 게이트)은 Accuracy 92.4%, Precision 0.918, Recall 0.920, F1 0.919로 비교한 모든 방식 중 가장 높은 값을 보였다.

3. Class-wise Analysis and Confusion Matrix

[그림 5]는 제안 모델의 평가셋에 대한 정규화 혼동행렬로 대각 성분이 전반적으로 87% 이상으로 모든 클래스에 대해 안정적인 판별이 이루어졌음을 확인할 수 있다.



Fig. 5. Normalized confusion matrix of the proposed model (%)

[그림 5]와 같이 오분류의 대부분이 인접한 의미 범주 사이에서 발생하였다. '집중-흥미로움'과 '집중-차분함' 사이의 상호 오분류(3.8%, 3.7%)는 두 클래스 모두 '집중' 상태에 해당하여 실제 교수 개입 측면에서 영향이 적다. 그러나 '비집중-지루함'과 '졸음' 사이의 오분류되는 경우(6.7%)와 그 반대의 경우(4.6%)는 개입 전략이 상이해야 하므로, 이러한 오분류는 모델의 실질적인 적용 시 더 유의미한 분석 대상이 된다.

[표 6]과 [그림 6]은 시각 단일 모달리티 모델과 제안 모델의 클래스별 F1-score를 비교한 것이다.

www.kci.go.kr

Table 6. Class-wise F1-score comparison

Class	Vision-Only	Proposed	Gain
Focused-Interested	0.915	0.950	+0.035
Focused-Calm	0.905	0.947	+0.042
Unfocused-Calm	0.834	0.902	+0.068
Unfocused-Bored	0.776	0.888	+0.112
Drowsy	0.793	0.909	+0.116
Macro average	0.844	0.919	+0.075

[그림 6]에서와 같이 음향 데이터 융합에 따른 성능 향상은 클래스별로 뚜렷한 편차를 보였고, 특히 시각 정보만으로는 구분이 어려운 비집중-지루함과 졸음 클래스에서 F1-score가 각각 0.112, 0.116만큼 상승하며 가장 큰 개선폭을 기록하였다. 이는 두 상태가 외형적으로는 유사한 저각성 상태이나, 음향적으로는 명확히 구별 가능하다는 본 연구의 가설을 실증적으로 뒷받침한다.

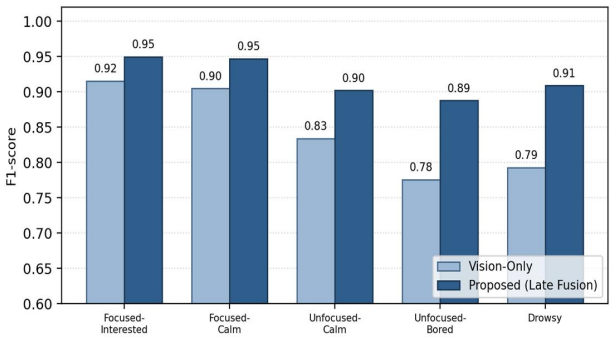


Fig. 6. Class-wise F1-score of the vision-only baseline and the proposed model

4. Quantitative analysis of acoustic cues

음향의 상보적 효과를 검토하기 위해, 각 클래스의 음향 에너지 포락선(RMS envelope)에서 0.2~0.4Hz 대역의 상대파워를 산출하였다. 이때 0.2~0.4Hz는 음향 신호의 고유 주파수가 아니라, 호흡 주기에 대응하는 에너지의 시간적 변동 성분임을 명확히 한다. 5초 단일 구간은 0.2Hz의 한 주기에 해당하므로 단일 세그먼트만으로 주기성을 확정하기 어렵다는 한계가 있으며, 따라서 다수 표본에 대한 분포와 비모수 검정으로 경향을 검토하였다. 또한 다중 화자·배경 소음 환경에서는 검출된 성분이 특정 학습자에게서 기인한다고 단정하기 어렵다.

V. Conclusions

본 연구는 AI-Hub「학습태도 및 성향 관찰 데이터」를 활용하여 안면 영상과 환경 음향을 결합한 게이팅 특징수준 융합 기반의 관찰 가능한 학습태도·행동 몰입 상태 분류 프레임워크를 제안하였다. 피험자 단위 분할로 평가 독립성을 확보하고, 3-seed 반복의 평균±표준편차와 융합 어블레이션·통계 검정을 통해 게이팅 구조의 기여를 분리 검증하였다. 특히 시각 정보만으로 구분이 모호한 비집중-지루함과 '졸음'의 변별에서 음향의 상보적 효과가 관찰되었다.

본 연구의 한계는 통제된 환경 데이터로 학습되어 다인 수업의 배경 소음·다중 화자 상황에 대한 추가 검증이 필요하다. 또한 외형 기반 관찰 레이블의 특성상 학습자의 내적 인지 상태를 온전히 반영하지 못하며, 본 연구의 결과는 관찰 가능한 상태의 분류로 한정된다. 마지막으로, 게이트는 품질 레이블로 감독되지 않으므로 신호 신뢰도 측정 장치로 해석할 수 없고 결측·현장 적용 결과는 예비적 수준이다.

향후 연구로는 ResNet-TCN·Transformer 등 강한 기준 모델과의 확장 비교와 품질 레이블 기반 게이트 감독 및 조명·SNR과 게이트 값의 상관 분석 및 실제 교육용 단말에서의 온디바이스 성능 검증과 원본 전송을 배제한 연합학습 기반 프라이버시 보호 구조로의 확장할 필요가 있다.

REFERENCES

[1] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School Engagement: Potential of the Concept, State of the Evidence," Review of Educational Research, Vol. 74, No. 1, pp. 59-109, 2004.

[2] M. Csikszentmihalyi, "Flow: The Psychology of Optimal Experience," Harper & Row, New York, 1990.

[3] J. Whitehill, Z. Serpell, Y. C. Lin, A. Foster, and J. R. Movellan, "The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions," IEEE Trans. on Affective Computing, Vol. 5, No. 1, pp. 86-98, 2014.

[4] National Information Society Agency (NIA), "Learning Attitude and Disposition Observation Data (DataSetSn: 71715)," AI-Hub, 2023.

[5] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 41, No. 2, pp. 423-443, 2019.

[6] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," Proc. of the 28th ICML, pp. 689-696, 2011.

[7] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), Vol. 30, pp. 5998-6008, 2017.

[8] A. Howard et al., "Searching for MobileNetV3," Proc. of the

www.kci.go.kr

IEEE/CVF ICCV, pp. 1314-1324, 2019.
[9] S. K. D'Mello and A. Graesser, "Dynamics of Affective States during Complex Learning," Learning and Instruction, Vol. 22, No. 2, pp. 145-157, 2012.
[10] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards User Engagement Recognition in the Wild," arXiv:1609.01885, 2016.
[11] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and Localization of Student Engagement in the Wild," Proc. of DICTA, pp. 1-8, 2018.
[12] A. Abedi and S. S. Khan, "Improving State-of-the-Art in Detecting Student Engagement with ResNet and TCN Hybrid Network," Proc. of the 18th CRV, pp. 151-157, 2021.
[13] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," IEEE Signal Processing Letters, Vol. 23, No. 10, pp. 1499-1503, 2016.
[14] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 28, No. 4, pp. 357-366, 1980.
[15] Ministry of Education, "Plan for the Implementation of AI Digital Textbooks," Press Release, Republic of Korea, 2023.
[16] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, Vol. 9, No. 8, pp. 1735-1780, 1997.

Authors



Young-Bok Cho received the M.S., and Ph.D. degrees in Computer Science from Chungbuk National University, Korea, in 2003 and 2012, respectively. also Dr. Cho received more Ph.D degrees in Medical and Law from Chungbuk National University and Chungnam National University , Korea, in 2019 and 2024, respectively. She has Professor of Information Security at Daejeon University, Daejeon, Korea , in 2018 to 2024, She is currently a Professor in the Computer Education at Gyeongkuk National University, Andong, Korea, in 2024. Her research interests include AI medical image processing, AI security and medical information protection, mobile security.

www.kci.go.kr