Journal of KIIT. Vol. 24, No. 5, pp. 201-214, May 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.5.201

멀티모달 기반 딥페이크 비디오 및 과장 광고 통합 탐지 프레임워크 구현

등덕호*, 한주은**, 임채림***, 박현호****, 안수빈*****, 이다은******, 유길상*******

Implementation of a Multimodal Integrated Detection Framework for Deepfake Videos and Exaggerated Advertisements

Tehhao Teng*, Jueun Han**, Chaerim Lim***, Hyunho Park****, Subin Ahn*****, Daeun Lee******, and Gilsang Yoo*******

본 연구성과물은 2023년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임
(No. RS-2023-00246191)

요 약

최근 생성형 AI의 발전과 함께 온라인 플랫폼에서 딥페이크를 악용한 과장 광고로 인한 소비자 피해가 증가하고 있다. 기존 검증 방식은 영상 조작과 광고 문구의 과장성을 개별적으로 분석하여 멀티미디어 콘텐츠의 종합적 위험 판단에 한계가 있다. 이에 본 논문에서는 영상과 텍스트를 결합한 멀티모달 기반 딥페이크 및 과장 광고 통합 탐지 시스템을 제안한다. 제안 시스템은 영상의 시공간적 특징을 학습하기 위해 MobileNetV3와 LSTM을 결합한 하이브리드 모델로 딥페이크를 탐지하고, 텍스트 분석에는 KcELECTRA 모델을 활용하여 광고 문구의 과장성을 판별한다. 두 분석 결과를 통합하여 광고 위험 수준을 분류하였다. 실험 결과, 딥페이크 탐지 정확도 0.8833, 과장 광고 판별 정확도 0.9748을 달성하였다. 본 연구는 지능화되는 온라인 과장 광고 위협에 대응하여 소비자 신뢰를 제고하고 실질적 기술 가이드라인으로 활용될 것으로 기대한다.

Abstract

As generative AI advances, consumer harm from deepfake-based exaggerated advertising on online platforms has increased. Conventional methods analyze video manipulation and textual exaggeration independently, limiting comprehensive risk assessment. To address this issue, this paper proposes a multimodal detection system for deepfake and exaggerated advertising. The system employs a MobileNetV3-LSTM hybrid model to capture spatiotemporal features for deepfake detection and a KcELECTRA-based language model to identify exaggerated expressions in advertising text. Their outputs are fused to classify advertising risk levels. Experimental results show detection accuracies of 0.8833 for deepfakes and 0.9748 for exaggerated advertising. This study is expected to enhance consumer trust and provide practical technical guidance against increasingly sophisticated online advertising threats.

Keywords
multimodal deep learning, deepfake detection, exaggerated advertising, video-text integration

* 성균관대학교 지능형로봇학과
- ORCID: https://orcid.org/0009-0002-8153-7901
** 고려대학교 보건정책관리학부/뇌인지과학융합
- ORCID: https://orcid.org/0009-0001-1415-8648
*** 전북대학교 바이오메디컬공학부
- ORCID: https://orcid.org/0009-0008-5770-2544
**** The Hong Kong Polytechnic University, School of Accounting and Finance
- ORCID: https://orcid.org/0009-0004-7543-1945
***** 숙명여자대학교 글로벌서비스학부
- ORCID: https://orcid.org/0009-0008-5852-0021
****** 고려대학교 교육대학원 컴퓨터교육학과
- ORCID: https://orcid.org/0009-0008-3119-4545
******* 고려대학교 정보대학 정보창의교육연구소(교신저자)
- ORCID: https://orcid.org/0009-0002-1085-5355

· Received: Feb. 12, 2026, Revised: Mar. 13, 2026, Accepted: Mar. 16, 2026
· Corresponding Author: Gilsang Yoo
Creative Informatics & Computing Institute, 145 Anam-ro, Seongbuk-gu, Seoul, S.Korea
Tel.: +82-2-3290-1674, Email: ksyoo@korea.ac.kr

Ⅰ. 서 론

디지털 미디어 환경의 고도화와 함께 온라인 동영상 플랫폼의 이용이 급증하고 있다. 이에 따라 광고 콘텐츠의 전달 방식 또한 정적인 텍스트 중심에서 동적 영상 중심으로 빠르게 재편되고 있다. 그러나 인공지능, 특히 딥러닝 기술의 비약적인 발전은 영상, 음성, 이미지를 정교하게 조작할 수 있는 딥페이크(Deepfake) 기술의 확산을 초래하였으며, 이는 디지털 콘텐츠 생태계의 새로운 위협 요소로 부상하였다. 딥페이크로 생성된 콘텐츠는 육안으로는 실재와 구분하기 어려워 소비자의 인지적 판단을 왜곡할 위험이 크며, 이는 단순한 정보의 오염을 넘어 정치적 선전, 명예훼손, 금융 사기 등 심각한 사회적 문제로 전이될 가능성을 내포하고 있다[1]. 특히 딥페이크 생성 기술과 이를 유통·소비하는 미디어 환경이 동시에 진화함에 따라, 이에 대응하기 위한 실효성 있는 기술적·정책적 방안 마련이 시급한 실정이다.

이러한 위협은 온라인 광고 시장, 그중에서도 국민 건강과 직결되는 건강기능식품 광고 영역에서 더욱 심각하게 나타난다. 현행 「식품 등의 표시·광고에 관한 법률」 제8조는 거짓·과장 광고 및 소비자를 기만하는 행위를 엄격히 금지하고 있으며, 특히 의료 전문가의 추천을 가장한 광고 또한 규제 대상이다. 또한 최근 전문직 종사자의 광고 출연 금지 조항을 우회하기 위해, 딥페이크 기술로 생성된 '가상의 의사'를 등장시켜 법망을 교묘히 회피하는 사례가 발생하고 있다. 이러한 기만적 광고는 소비자가 거짓 정보를 전문적인 의학 정보로 오인하게 유도하며, 합리적인 구매 의사결정을 심각하게 저해하고 있다[2].

실제로 이러한 문제는 통계적으로도 심화되는 양상을 보인다. 식품의약품안전처 자료에 따르면, 최근 5년간 건강기능식품 부당광고 적발 건수는 총 22,948건에 달하며, 2025년 기준 연간 적발 건수는 5,214건으로 역대 최고치를 기록하였다[3]. 1981~1996년 출생한 소비자를 대상으로 한 심층 인터뷰 연구에서도 소비자들은 허위·과장 광고에 대해 강한 부정적 인식을 드러내며, 이에 대한 적극적인 규제와 기술적 필터링의 필요성을 강조하고 있다[4].

기존의 허위·과대광고 대응 체계는 주로 전문 모니터링 요원에 의한 육안 검사와 사후적 적발 방식에 의존해 왔다. 그러나 이러한 인력 의존적 방식은 판단의 주관성으로 인해 결과의 편차가 발생할 수 있으며, 숏폼(Short-form) 영상이나 라이브 커머스 등 대량으로 생성되고 휘발성이 강한 최신 광고 포맷에 실시간으로 대응하기에는 구조적인 한계가 존재한다[5]. 또한, 기존의 딥러닝 기반 광고 분석 연구들은 주로 텍스트 감성 분석이나 단순 분류에 치중되어 있어, 딥페이크와 과장 광고가 결합된 복합적인 위험도를 정량적으로 탐지하는 데에는 어려움이 있다. 이에 본 연구에서는 영상과 텍스트 정보를 상호보완적으로 활용하는 멀티모달(Multi-modal) 기반의 통합 탐지 프레임워크를 제안한다. 제안하는 프레임워크는 영상 내 인물의 얼굴 조작 여부를 탐지하는 MobileNetV3(Mobile Networks Version 3) 및 LSTM(Long Short-Term Memory) 기반의 딥페이크 탐지 모듈과, 광고 문구의 과대·허위성을 판별하는 KcELECTRA(Korean comments Efficiently Learning an Encoder that Classifies Token Replacements Accurately) 기반의 텍스트 분석 모듈을 병렬적으로 구성하여 광고의 종합적인 위험도를 평가한다. 본 시스템은 광고가 노출되는 시점에 멀티모달 정보를 실시간으로 분석함으로써, 사후 처벌 위주의 규제 한계를 넘어 소비자의 구매 행위 이전에 잠재적 위험을 경고하는 예방적 기술 대안을 제시한다는 점에서 의의를 지닌다.

본 논문의 구성은 다음과 같다. 제2장에서는 시스템에 적용된 기반 모델들과 관련 선행 연구를 고찰한다. 제3장에서는 본 연구가 제안하는 멀티모달 딥페이크 및 과장 광고 판별 시스템의 설계 원리와 구체적인 구현 방법론을 기술한다. 특히 식품의약품안전처 가이드라인을 반영한 KcELECTRA 기반 과장 광고 탐지 알고리즘과 시계열적 특징을 반영한 MobileNetV3 및 LSTM 결합 모델의 세부 구조를 다룬다. 제4장에서는 제안 시스템의 성능 평가를 위한 실험 환경 및 결과를 비교·분석하고, 실시간 온라인 환경에서의 적용 결과를 제시한다. 마지막으로 제5장에서는 본 연구의 결론 및 향후 연구 방향에 대해 기술하였다.

II. 관련 연구

온라인 광고의 신뢰성을 판단하기 위해서는 텍스트의 의미적·맥락적 분석뿐만 아니라, 영상의 공간적·시간적 특징을 통합적으로 고려하는 멀티모달 접근이 요구된다. 특히 건강기능식품 광고와 같이 과장된 표현과 조작 영상이 결합될 가능성이 있는 콘텐츠의 경우, 단일 모달 분석만으로는 기만적 요소를 충분히 탐지하기 어렵다. 이에 따라 본 장에서는 텍스트 기반 과장 표현 판별을 위한 KcELECTRA 모델, 영상 내 딥페이크 탐지를 위한 MobileNetV3 기반 구조, 그리고 프레임 간 시계열적 특징을 반영하기 위한 LSTM 기반 시공간 분석 기법에 대해 차례로 고찰한다.

2.1 KcELECTRA

ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)는 기존 BERT(Bidirectional Encoder Representations from Transformers)의 학습 효율성 문제를 개선하기 위해 RTD(Replaced Token Detection) 방식을 도입한 모델이다[6]. 본 연구에서 활용하는 KcELECTRA는 대규모 한국어 뉴스 댓글 및 사용자 생성 콘텐츠(UGC)를 포함한 말뭉치를 통해 사전 학습된 한국어 특화 언어모델이다. 학습 데이터에 오탈자, 은어, 이모지 등 비정형 표현이 다수 포함되어 있어, 구어체와 비표준적 표현이 빈번한 온라인 텍스트 분석에 탁월한 성능을 보인다. 특히 감성 분석 및 텍스트 분류 태스크에서 높은 성능을 입증하고 있다. J. Kim et al.[7]는 감성 분류를 위해 파인튜닝된 KcELECTRA 모델이 정확도 0.9800, F1 점수 0.9804의 우수한 성능을 기록하였음을 보고하였다. 해당 연구는 단일 문장뿐 아니라 질문-응답 쌍을 통합 분석하여 대화의 맥락을 파악하는 기법을 제안하였으며, 이는 텍스트의 표면적 의미를 넘어 문맥적 의도를 해석하는 데 이바지하였다. 또한, S. Jun et al.[8]는 홈쇼핑 및 SNS의 패션 리뷰 데이터를 활용하여 속성 기반 감성 분류에 KcELECTRA를 활용하였다. 해당 연구는 속성별 평가에 최대 0.8728의 정확도와 0.8526의 F1 점수를 기록하였으며, 이를 통해 마케팅 전략 수립 및 기업의 의사결정에 기여할 수 있는 유의미한 정보를 제공하였다. 이러한 선행 연구들은 KcELECTRA가 비정형성이 강한 한국어 텍스트 환경에서도 문맥적 의미를 정확히 포착함을 보여준다. 이는 본 연구의 목표인 온라인 광고 문구 내의 교묘한 과장이나 기만적 표현을 판별하는 데 있어서도 해당 모델이 효과적인 백본(Backbone)으로 작용할 수 있음을 시사한다.

2.2 MobileNetV3

MobileNet은 모바일 기기 및 임베디드 시스템과 같이 제한된 컴퓨팅 자원 환경에서도 효율적인 추론이 가능하도록 설계된 경량 합성곱 신경망(CNN, Convolutional Neural Network)이다. 그중 MobileNetV3는 신경망 구조 탐색(NAS, Neural Architecture Search)과 NetAdapt 알고리즘을 결합하여, 경량화와 정확도 간의 최적 균형을 달성하도록 설계되었다[9].

Y. Han et al.[10]는 MobileNetV3를 기반으로 모바일 애플리케이션 환경에서 실시간 이미지 분류를 수행하여 F1 점수 0.9949, 정확도 0.9984를 기록하였으며, 이를 통해 저사양 환경에서의 실시간 추론이 가능함을 실험적으로 검증하였다. 딥페이크 탐지 분야에서도 경량 모델의 활용 연구가 활발하다. S. Jang[11]는 MobileNet 기반 모델에 양자화 및 가지치기(Pruning) 등의 최적화 기법을 적용하여, 모바일 환경에서도 실시간 딥페이크 탐지가 가능함을 실험적으로 검증하였다. 최적화된 MobileNet의 정확도는 XceptionNet의 정확도 0.9580, F1 점수 0.9300보다 낮은 0.9310, 0.9000을 기록하였으나, 메모리 사용량은 XceptionNet이 512MB, MobileNet이 128MB로 크게 감소하였다. 이를 통해 모바일 환경에서도 실시간 딥페이크 탐지가 가능함을 실험적으로 검증하였다. 아울러 S. Choi et al.[12]는 동일한 모델 구조라도 데이터셋의 구성과 전처리 방식에 따라 탐지 성능이 크게 달라질 수 있음을 지적하며, 모델의 효율성뿐만 아니라 데이터 중심의 설계 전략이 중요함을 강조하였다. 해당 연구에서는 실험을 위해 VGG16, VGG19, MobileNetV2, ResNet50V2, EfficientNetB7을 활용해 전처리만 다르게 적용한 데

이터셋을 비교했으며, 이 중 MobileNetV2에 얼굴 부분을 크롭한 데이터셋을 학습시켰을 때 정확도 0.8872, F1 점수 0.8914로 가장 높은 성능을 보였다.

이러한 연구 결과들은 MobileNetV3가 높은 실시간성과 정확도를 동시에 요구하는 온라인 광고 영상 분석에 적합한 모델임을 뒷받침한다. 본 연구에서는 이러한 MobileNetV3의 특징을 활용하여 대량의 광고 콘텐츠를 신속하게 처리하고, 딥페이크로 의심되는 영상을 효과적으로 선별하는 영상 특징 추출기로 활용하고자 한다.

2.3 LSTM 기반의 시공간적 특징 분석

딥페이크 영상은 단일 프레임 내의 이미지 조작뿐만 아니라, 프레임과 프레임 사이의 시간적 흐름에서 발생하는 부자연스러운 움직임, 깜빡임(Flickering), 입술 모양 불일치(Lip-sync mismatch) 등의 시계열적 이상 징후를 동반한다. 따라서 정지 이미지 분석에 특화된 CNN 모델만으로는 이러한 시간적 불일치(Temporal inconsistency)를 포착하는 데 한계가 존재한다. 반면, LSTM은 순환 신경망의 장기 의존성(Long-term dependency) 문제를 해결하기 위해 고안된 모델로, 시계열 데이터의 시간적 패턴을 학습하는 데 탁월한 성능을 보인다. 즉, 영상 내 공간적 특징(Spatial features)과 시간적 특징(Temporal features)을 동시에 고려하는 시공간적(Spatiotemporal) 분석은 딥페이크 탐지의 정확도를 높이는 핵심 요인임을 시사한다. 본 연구에서는 이러한 점에 착안하여, 앞서 기술한 MobileNetV3를 통해 프레임별 특징을 효율적으로 추출하고, 이를 LSTM 네트워크에 입력하여 광고 영상 내 인물의 연속적인 움직임 정보를 통합 분석하는 구조를 채택하였다. 이는 정적인 이미지 조작뿐만 아니라, 동영상 광고에서 발생할 수 있는 교묘한 시계열적 조작을 효과적으로 탐지하도록 한다.

III. 제안한 과장 광고 판별 시스템

유튜브 동영상 광고의 딥페이크 여부 및 과장 광고 여부를 탐지하기 위한 멀티모달 시스템의 전체 구조는 그림 1과 같다.

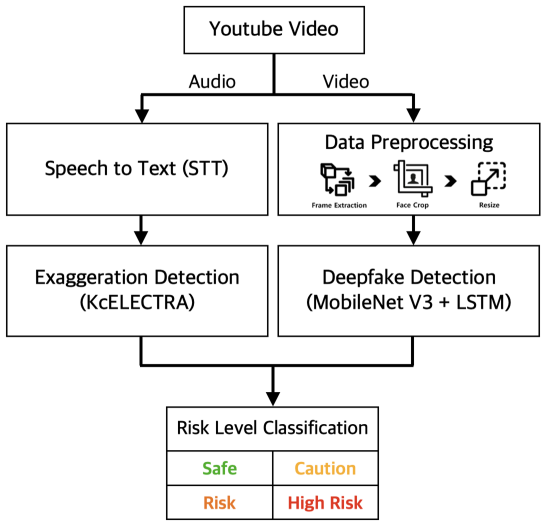

그림 1. 전체 시스템 구조도
Fig. 1. Overall system architecture

본 시스템은 입력된 동영상으로부터 음성(Audio)과 영상(Frame) 정보를 각각 추출한 후, 두 모달리티를 독립적으로 분석하고 그 결과를 통합하여 최종적인 광고 위험 수준을 산출하는 프레임워크로 설계되었다.

텍스트 분석 단계에서는 동영상으로부터 음성 트랙을 추출한 뒤, 자동 음성 인식 모델인 Whisper를 활용하여 발화 내용을 텍스트로 변환한다. Whisper는 다국어 음성 데이터를 기반으로 학습된 OpenAI의 ASR(Automatic Speech Recognition) 모델로, 온라인 영상 광고와 같이 배경 소음이 포함된 환경에서도 비교적 안정적인 전사 성능을 보이는 모델로 알려져 있다[13]. 변환된 텍스트는 문장 단위로 분할되어 과장·허위 광고 탐지 모델을 통해 광고 내 언어적 주장, 효능 표현, 오인 가능성이 있는 문장 구조를 분석하여 과장 여부가 판별된다.

영상 분석 단계에서는 입력 동영상으로부터 프레임을 추출한 후, 영상 분석 모듈로 전달된다. 딥페이크 탐지 모델은 합성곱 신경망 모델을 통해 전달된 프레임에서 공간적 특징을 추출하고, 시계열 모델을 통해 프레임 간 시간적 일관성까지 반영함으로써 딥페이크를 판별한다.

최종적으로 텍스트 기반 과장 광고 탐지 결과와 영상 기반 딥페이크 탐지 결과를 결합하여 광고 위험 수준을 4단계로 판별한다.

3.1 과장 광고 판별

음성 기반 텍스트 분석은 과장되거나 허위 가능성이 높은 광고 문장을 탐지한다. 본 시스템에서는 의미 단위를 최대한 보존하면서 과장 표현을 정밀하게 탐지하기 위해 문장 단위 분석을 수행하였다. 데이터셋은 유튜브 건강기능식품 광고 아카이브, KHFF TV(한국건강기능식품협회), 페이스북 메타 광고 라이브러리 등 주요 온라인 동영상 광고 아카이브 및 광고 라이브러리를 중심으로 수집하여 실 환경 분포를 반영하였다. 수집된 광고 문장은 과장 광고 1,090문장과 일반 광고 1,090문장을 이진 범주로 라벨링 하였다. 라벨링 과정은 단일 기준에 의존하지 않고 복수 기준에 따른 교차 검증을 수행하여 라벨 신뢰도를 확보하였다. 건강기능식품 광고 관련 법규에 명시된 금지 항목을 기준으로, 효과 확정 표현, 비교·우월 표현, 전문가·권위 인용 표현, 감정 과장 표현, 체험담 조작 표현의 다섯 가지 유형으로 정의하였다. 이 중 하나 이상의 유형이 포함된 경우 해당 문장을 과장 광고 문장으로 분류하였다[14].

광고 텍스트는 구어체, 은어, 맥락 의존적 표현이 빈번하다는 특성을 고려하여, 한국어 댓글 데이터로 사전학습된 KcELECTRA를 백본 모델로 채택하고 파인튜닝하였다. KcELECTRA는 Generator와 Discriminator를 동시에 학습하는 RTD(Replaced Token Detection) 방식을 사용하며, 전체 손실 함수 Ltotal은 식 (1)과 같이 정의된다.

$$L_{total} = L_{MLM} + \lambda L_{RTD} \quad (1)$$

여기서 L_{MLM}은 Generator가 입력 문장의 일부 토큰을 예측하는 과정에서 발생하는 마스크드 언어 모델링 손실이며, L_{RTD}는 Discriminator가 Generator에 의해 치환된 토큰이 실제 토큰인지 여부를 판별하는 과정에서 발생하는 대체 토큰 탐지 손실을 의미한다. λ는 RTD 손실이 전체 학습에 미치는 영향을 조절하는 가중치 계수이다.

이와 같은 학습 구조를 통해 모델은 문맥적 의미 표현을 유지하면서도, 단어 수준의 미세한 치환이나 과장 표현과 같은 광고 문구 내 조작 패턴을 효과적으로 학습할 수 있다. 본 연구에서는 입력 문장의 최대 길이를 256 토큰으로 설정하였으며, 의미 왜곡을 방지하기 위해 과도한 전처리는 지양하고 원문 형태를 최대한 보존하여 학습에 활용하였다.

3.2 딥페이크 판별

영상 기반 딥페이크 탐지 모델 학습을 위해 KoDF(Korean DeepFake Detection, AI Hub)[15]와 FF++(FaceForensics++)[16] 데이터셋을 7:3 비율로 혼합하여 총 1,500개의 영상 데이터를 구축하였다. 이는 국내 광고 환경에 주를 이루는 한국인의 얼굴 특성을 모델에 충분히 반영하기 위함이다. 이후 학습 및 검증 데이터는 8:2 비율로 분할하였다. 성별에 따른 편향을 최소화하기 위해 남성과 여성의 비율을 1:1로 균등하게 맞추고 다양한 해상도(FHD 등)와 프레임률(FPS)을 포함하여 모델의 일반화 성능을 높였다.

전처리 단계에서는 각 영상에서 같은 간격으로 10개의 프레임을 추출하여 시퀀스를 구성하였다. 얼굴 검출에는 조명 변화와 다양한 각도에 강건한 MTCNN(Multi-task Cascaded Convolutional Networks)[17]을 사용하여 얼굴 ROI(Region of Interest)를 추출하고, 256×256 크기로 리사이즈하여 모델 입력으로 사용하였다. MTCNN은 P-Net, R-Net, O-Net으로 구성된 단계적 구조를 통해 얼굴 후보 영역을 점진적으로 정제함으로써, 다양한 해상도 및 조명 조건에서도 안정적인 얼굴 검출 성능을 제공한다. 그림 2는 MTCNN을 통한 얼굴 검출 및 전처리 결과이다.

그림 2. MTCNN을 이용한 얼굴 영역 검출 및 전처리 결과
Fig. 2. Face detection and preprocessing results using MTCNN

또한, 과적합 방지를 위해 학습 데이터에 한해 좌우 반전(Horizontal flip), ±10° 범위의 회전(Rotation), 밝기·대비·채도·색조 조절(Color jitter), 그리고 전체 이미지의 약 15% 영역을 무작위로 가리는 cutout 증강 기법을 적용하였으며, 그림 3의 t-SNE 시각화 결과와 같이 데이터 증강이 특징 공간의 클래스 분리도를 향상됨을 확인하였다[18].

정적 이미지 기반 CNN 모델은 개별 프레임의 공간적 특징을 효과적으로 추출할 수 있으나, 프레임 간의 미세한 경계 변화, 표정 전이의 비연속성, 움직임의 부자연스러움과 같은 시간적 특성을 충분히 반영하는 데에는 한계가 존재한다[19]. 따라서 프레임 간의 시간적 불일치를 포착하기 위해 MobileNetV3와 LSTM을 결합한 하이브리드 모델을 적용하였다. MobileNetV3-Large는 경량화를 위해 깊이별 분리 합성곱(Depthwise separable convolution)을 사용하며, 식 (2)와 같은 연산 효율성을 갖는다.

$$DSConv(x) = PWConv(DWConv(x)) \tag{2}$$

식 (2)에서 x는 공간적 차원(Height × Width)과 채널(Channel) 정보를 포함하는 입력 특징 맵을 의미한다. DWConv(·)는 각 채널에 대해 독립적으로 공간 필터링을 수행하는 깊이별 합성곱 연산으로, 채널 간 상호작용(Inter-channel mixing)을 수행하지 않음으로써 연산 복잡도와 파라미터 수를 감소시킨다. 이후 PWConv(·)는 1×1 합성곱을 통해 채널 간 정보를 선형 결합하여 특징 표현을 재구성하는 역할을 한다.

또한, 비선형 활성화 함수로 h-swish를 사용하여 연산 속도와 정확도를 최적화하였으며, 이는 식 (3)과 같다[9].

$$h-swish(x) = x \cdot \frac{ReLU6(x+3)}{6} \tag{3}$$

추출된 프레임별 특징 벡터 시퀀스는 LSTM에 순차적으로 입력되어 시간적 의존성을 학습한다. LSTM은 순환 신경망의 장기 의존성 문제를 해결하기 위해 Sepp Hochreiter와 Jürgen Schmidhuber가 제안한 표준 LSTM 구조에 기반하여 LSTM의 t시점에서의 은닉 상태 ht는 식 (4)와 같이 이전 상태 ht-1와 현재 입력 x_t를 통해 갱신된다[20].

$$h_t = LSTM(x_t, h_{t-1}) \tag{4}$$

$$\hat{y} = \sigma(Wh_t + b) \tag{5}$$

식 (5)는 LSTM 기반 분류기의 최종 출력 계산 과정을 나타낸다. 여기서 h_t는 LSTM의 마지막 시점에서의 은닉 상태(Hidden state)를 의미하며, W와 b는 각각 학습 가능한 가중치 행렬(Weight matrix)과 편향 벡터(Bias vector)를 나타낸다. 식 (5)에 의해 계산된 출력값 $\hat{y}$는 선형 변환을 통해 계산된 시그모이드(Sigmoid) 함수를 통과하여 0과 1 사이의 값으로 변환되며, 이는 입력 영상이 딥페이크일 확률로 해석된다.

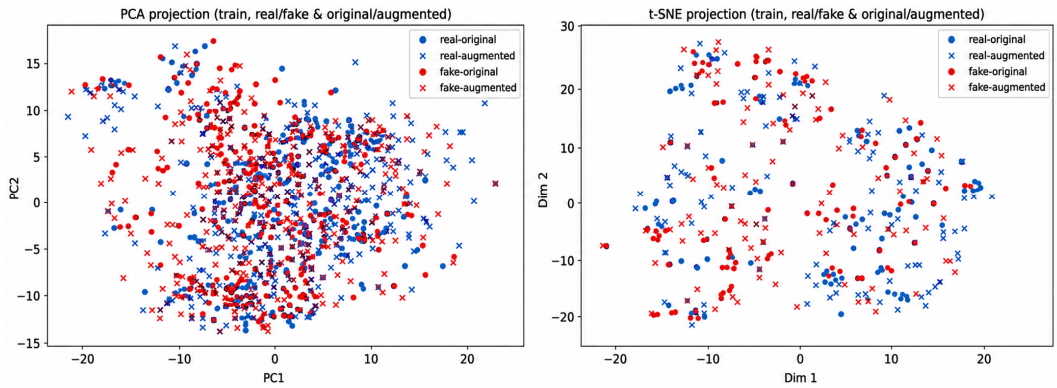

그림 3. 데이터 증강 적용에 따른 특징 공간 분포의 t-SNE 및 PCA 시각화 결과
Fig. 3. t-SNE and PCA visualization of feature distributions with data augmentation

최종적으로 해당 확률값은 사전에 설정된 임계값(Threshold)과 비교되어 딥페이크 여부에 대한 이진 분류 결과가 결정된다. 이와 같이 시공간적 통합 분석 구조는, 단일 프레임 분석으로는 포착하기 어려운 시간적 불연속성과 조작 패턴을 효과적으로 반영할 수 있도록 설계하였다.

Ⅳ. 실험 결과

본 연구에서는 광고 문장의 과장 여부 판별 성능을 검증하기 위해 KoBERT, KoELECTRA, KcELECTRA 등 세 가지 사전학습 언어모델을 비교 분석하였다[21]. 또한, 선정된 모델에 대해 파인튜닝 전략에 따른 성능 변화를 고찰하였다. 아울러 제안하는 MobileNetV3-LSTM 구조의 타당성을 입증하기 위해, 다양한 CNN 백본과 시계열 모델 조합을 동일한 실험 조건에서 비교 평가하였다.

4.1 과장 광고 판별 모델 비교 평가

본 연구에서는 한국어 광고 문장의 과장 여부 판별 성능을 비교하기 위해 KoBERT, KoELECTRA, KcELECTRA의 세 가지 사전학습 언어모델을 활용하였다[21]. 세 모델은 과장 광고 문장과 일반 광고 문장으로 구성된 동일한 데이터셋을 사용하였으며, 8:1:1(훈련/검증/테스트)의 비율로 분할하였다. 학습률은 3e-5~5e-5 범위에서 탐색하였고, 가중치 감쇠(Weight decay)는 0.01~0.1 범위로 설정하였다. 또한 조기 종료(Early stopping)의 patience 값은 3~10 사이에서 조정하였다. 모든 모델은 동일한 데이터 분할과 하이퍼파라미터 탐색 범위를 적용하여 학습함으로써, 모델 간 성능 비교의 공정성을 확보하였다.

그림 4와 같이, KcELECTRA는 검증 데이터셋에서 가장 높은 정확도 0.968을 기록하였으며, 특히 과장 광고(Positive)와 일반 광고(Negative)를 구분함에 있어 정밀도(Precision)와 재현율(Recall) 간의 균형 척도인 F1-score는 0.9748로 세 모델 중 가장 안정적인 것으로 나타났다.

또한 그림 5의 학습 및 검증 손실 곡선을 분석한 결과, KcELECTRA는 학습 초기 단계부터 손실이 빠르게 감소하며 안정적으로 수렴하는 경향을 보였다. 검증 손실 역시 학습 후반부까지 과적합 징후 없이 꾸준히 감소하여, 실사용 환경에서의 일반화 성능을 기대할 수 있다. 이러한 결과를 바탕으로 본 연구에서는 텍스트 분석을 위한 기본 모델로 KcELECTRA를 채택하고, 모델 학습에 적용된 주요 하이퍼파라미터는 표 1과 같다.

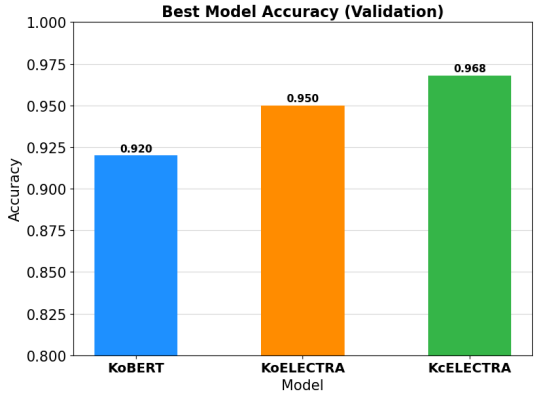

그림 4. 여러 모델들간의 실험 비교
Fig. 4. Experimental comparison between different models

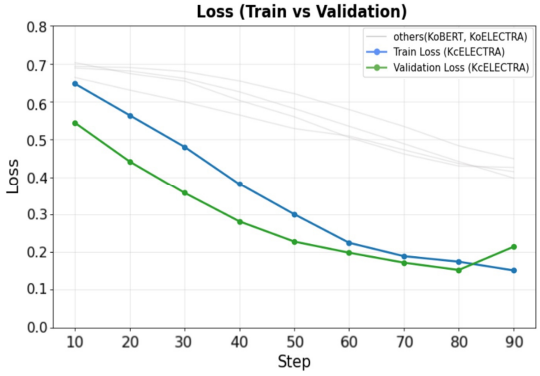

그림 5. 학습 단계에 따른 훈련 및 검증 손실 변화
Fig. 5. Training and validation loss curves over training steps

표 1. KcELECTRA-base-v2022 모델 학습 파라미터
Table 1. KcELECTRA-base-v2022 model training parameters

Maximum length	256
Epoch size	10
Batch size	32
Learning rate	5e-5
Weight decay	0.01
Early stopping (Patience)	3 (Accuracy)

4.2 KcELECTRA 모델 파인튜닝

선정된 모델의 최적화된 파인튜닝을 위해, 학습 가능한 레이어(Trainable layers)의 범위를 달리하여 총 6가지 실험을 수행하였다. 그림 6은 각 전략에 따른 검증 손실 변화를 나타내며, 분석 결과 모델의 마지막 4개 레이어(Last 4 layers)만을 학습시키는 전략이 가장 안정적인 수렴 특성과 우수한 성능을 보였다.

최종 선정된 모델(Last 4 layers 학습)의 성능은 표 2와 같이 정확도 0.9748, F1-score 0.9753을 달성하였다. 또한 표 3의 혼동 행렬 분석 결과, 과장 광고 문장에 대한 재현율이 매우 높게 유지되어, 실제 모니터링 환경에서 소비자 피해를 유발할 수 있는 위험 문장을 효과적으로 탐지할 수 있음을 확인하였다.

표 2. 모델 성능 평가 결과
Table 2. Evaluation results of model performance

Trainable layer	Accuracy	Precision	Recall	F1-Score
Last 4 layer	0.9748	0.9559	0.9954	0.9753

표 3. 혼동행렬 결과
Table 3. Results of confusion matrix

	Hyperbolic (1)	Non-hyperbolic (0)
Hyperbolic (1)	217	1
Non-hyperbolic (0)	10	208

4.3 딥페이크 판별

영상 기반 딥페이크 탐지 모델의 최적 구조를 탐색하기 위해, CNN 백본(ResNeXt-50[22], EfficientNet-B0 [23], MobileNetV3[24])과 시계열 모델(LSTM, Bi-LSTM, ConvLSTM)의 다양한 조합을 단계적으로 비교하였다. 모든 비교 실험은 공정한 평가를 위해 학습률, 시퀀스 길이 등 주요 하이퍼파라미터를 통제한 상태에서 진행되었다.

첫째, 표 4에서와 같이 시계열 모델을 기본 LSTM으로 고정한 상태에서 CNN 백본별 성능을 비교하였다. 실험 결과, MobileNetV3를 백본으로 사용한 모델이 다른 고성능 CNN 모델(ResNeXt, EfficientNet) 대비 대등하거나 더 우수한 성능을 보였다.

표 4. LSTM 고정 조건에서 CNN 백본별 딥페이크 탐지 성능 비교
Table 4. Deepfake detection performance by CNN backbone(LSTM fixed)

Spatial feature extraction	Temporal feature extraction	Method	TP	FN	FP	TN	Best val acc.
ResNext	LSTM	Freeze	124	26	41	109	0.8333
		Unfreeze	105	45	13	137	0.8333
MobilenetV3		Freeze	140	10	36	114	0.8660
		Unfreeze	135	15	33	117	0.8600
Efficientnet-B0		Freeze	135	15	44	106	0.8400
		Unfreeze	134	16	30	120	0.8600

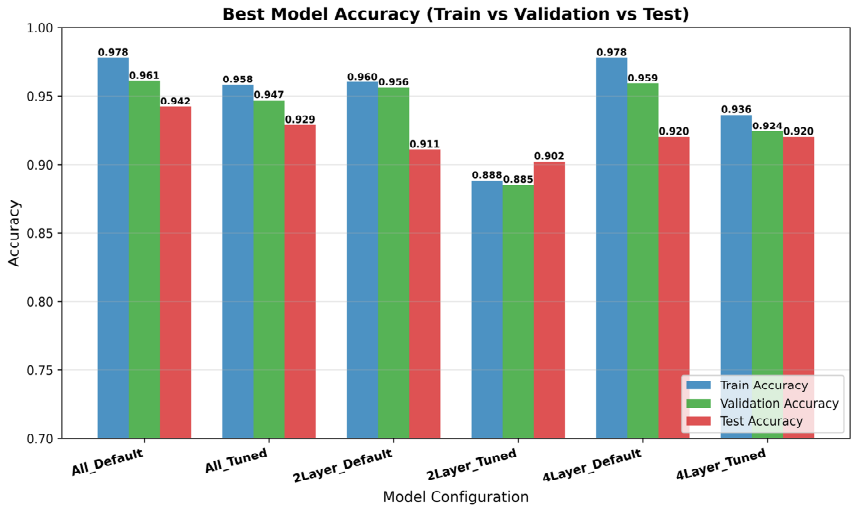

그림 6. KcELECTRA 모델 학습 반복 횟수에 따른 검증 및 검증 손실 변화
Fig. 6. Changes in validation and validation loss according to the number of KcELECTRA model training iterations

특히 ResNeXt-50과 EfficientNet-B0는 일부 조건에서 FP(False Positive) 또는 FN(False Negative) 비율이 증가하며 성능 변동 폭이 컸던 반면, MobileNetV3는 경량 구조임에도 불구하고 얼굴의 미세한 조작 패턴을 안정적으로 포착하였다.

둘째, 표 5에서와 같이 MobileNetV3 백본을 고정한 상태에서 시계열 모델 구조에 따른 성능 차이를 비교하였다. 실험 결과, 복잡한 구조인 ConvLSTM이나 Bi-LSTM보다 기본 LSTM 구조가 본 연구의 데이터셋과 시퀀스 길이 조건에서 가장 높은 정확도와 안정성을 보였다. 기본 LSTM에 비해 정확도가 ConvLSTM은 0.0293, 0.0476 감소하였으며 BiLSTM은 0.016, 0.0393 감소하였다. 종합적으로 본 연구의 데이터 규모와 시퀀스 길이 조건에서는 기본 LSTM 구조가 성능 안정성과 일반화 측면에서 가장 적절한 시계열 모델로 확인되었다.

이상의 단계적 실험을 통해 본 연구에서는 MobileNetV3(Unfreeze) + LSTM 구조를 최종 모델로 선정하였다. 특히 백본의 가중치를 고정하지 않고 일부 학습시키는 Unfreeze 설정이 Freeze 설정 대비 검증 정확도와 F1-score 측면에서 유리한 것으로 확인되었다. 최적화 도구인 Optuna를 활용하여 도출된 최종 하이퍼파라미터 값은 표 6과 같다.

표 5. MobileNetV3 고정 조건에서 시계열 모델별 딥페이크 탐지 성능 비교
Table 5. Deepfake detection performance by temporal model (MobileNetV3 fixed)

Temporal model	Method	TP	FN	FP	TN	Best val acc.
Conv LSTM	Freeze	128	22	27	123	0.8367
	Unfreeze	128	22	33	117	0.8184
Bi-LSTM	Freeze	117	33	12	138	0.8500
	Unfreeze	123	27	35	125	0.8267

표 6. Optuna 도구를 활용한 하이퍼파라미터 튜닝 조합
Table 6. Hyperparameter-tuning for optuna

Hyperparameter	Values
Epoch	50
Learning rte	1e-3
Batch size	32
Sequence length	12
Early stopping (Patience)	5 (Loss)

해당 설정을 적용한 학습 과정에서 훈련 및 검증 손실은 반복 횟수 증가에 따라 그림 7에서와 같이 전반적으로 안정적인 감소 추이를 보였으며, 학습 후반부에서도 과적합 없이 수렴 특성이 유지됨을 확인하였다. 이러한 결과는 제안한 MobileNetV3 - LSTM 구조가 본 연구의 데이터 구성과 시계열 길이 조건에서 효과적인 일반화 성능을 확보하였다.

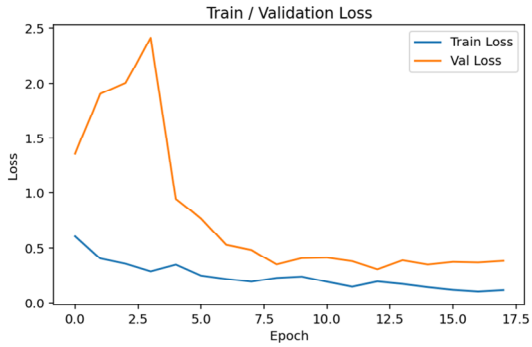

그림 7. 딥페이크 학습 및 검증 손실 변화
Fig. 7. Training and validation loss of the deepfake detection model

검증 데이터셋에 대한 혼동행렬(표 7)과 분류 리포트(표 8)를 분석한 결과, 딥페이크(Positive)와 정상(Negative) 클래스 모두에서 균형 잡힌 분류 성능을 보였으며, 특히 딥페이크를 정상으로 오분류하는 사례를 최소화하여 탐지 시스템으로서의 신뢰성을 확보하였다. 이는 본 모델이 실제 광고 영상 환경에서 위험 콘텐츠 누락 가능성을 낮추는 방향의 예측 특성을 갖추고 있음을 의미한다.

표 7. 혼동행렬 결과
Table 7. Results for confusion matrix

Class	Deepfake(1)	Normal(0)
Deepfake(1)	132	18
Normal(0)	17	133

표 8. 검증 데이터에 대한 분류 성능 평가 결과
Table 8. Classification performance on the validation dataset

Class	Precision	Recall	F1-Score	Accuracy	Support
Deepfake(1)	0.8859	0.8800	0.8829	0.8833	150
Normal(0)	0.8808	0.8867	0.8837		150

4.4 멀티모달 기반 판별 시스템 구현 결과

제안된 멀티모달 프레임워크의 실효성을 검증하기 위해 웹 기반의 자동 탐지 서비스를 구현하였다. 본 시스템은 사용자 개입이 필요 없는 제로 인터랙션(Zero-interaction) 방식을 채택하여, 사용자가 영상을 시청하는 동안 백그라운드에서 분석이 수행되도록 크롬 확장팩으로 개발되었다.

서버 측에서는 영상 기반 딥페이크 분석과 텍스트·음성 기반 과장 광고 분석을 비동기 병렬 방식으로 수행한다. 분석 결과는 사용자가 직관적으로 인지할 수 있도록 Safe, Caution, Risk, High Risk의 4단계 위험 등급으로 제공된다. 위험 등급은 각 분석 모듈에서 산출된 확률값과 임계값 0.5를 기준으로 정의하였다.

각 분석 모듈은 이진 분류 구조로 설계되었으며, 단일 출력 노드를 통해 로짓(Logit) 값을 산출한다. 학습에는 BCEWithLogitsLoss(Binary Cross Entropy with Logits)를 적용하였고, 추론 단계에서는 로짓 값에 sigmoid 함수를 적용하여 0~1 범위의 확률값으로 변환하였다.

이진 분류 문제에서 두 클래스의 사전확률과 오분류 비용이 동일하다고 가정할 경우, 0.5는 합리적인 기본 결정 경계로 간주된다. 본 연구에서는 텍스트 기반 과장 판별 모델과 영상 기반 딥페이크 탐지 모델이 동일한 이진 분류 구조를 갖는 점을 고려하여, 모듈 간 출력 확률의 일관성을 유지하기 위해 임계값을 0.5로 설정하였다.

영상 기반 딥페이크 탐지 확률과 음성·텍스트 기반 과장 광고 판별 확률이 모두 0.5 미만인 경우를 Safe로 분류하였다. 딥페이크 탐지 확률만 0.5 이상인 경우는 Caution, 과장 광고 판별 확률만 0.5 이상인 경우는 Risk로 정의하였다. 두 확률값이 모두 0.5 이상인 경우는 High Risk 단계로 분류한다.

또한 멀티모달 분석 결과를 단일 지표로 통합하기 위해 식 (6)과 같이 가중합 기반 신뢰도 점수를 정의하였다. 건강기능식품 광고는 「식품 등의 표시·광고에 관한 법률」 제8조에 따라 거짓·과장 광고를 엄격히 금지하고 있으므로, 본 연구에서는 과장 여부의 중요도를 반영하여 과장 광고 판별 결과의 가중치를 0.7로 설정하였다[2].

$$Confidence\ Score = 1 - (\alpha \cdot Ad + (1-\alpha) \cdot Deepfake) \quad (6)$$

식 (6)에서 Ad는 광고 문구의 과장 확률, Deepfake는 영상의 딥페이크 조작 확률을 의미하며, 각각 0과 1 사이로 정규화된다. α는 가중치 계수로 본 연구에서는 α=0.7로 설정하였다. 최종 신뢰도 점수는 두 모달리티 확률의 가중합을 1에서 차감하여 산출되며, 값이 낮을수록 광고의 종합 위험도가 높음을 의미한다.

실험 데이터는 유튜브 플랫폼을 중심으로 공개 접근이 가능한 소셜미디어에 게시된 콘텐츠를 대상으로 하였으며, 검증되지 않은 전문가로 보이는 사람이 제품을 홍보하는 내용을 다루는 광고이다. 모든 광고는 숏폼을 대상으로 연구 목적에 한해 저장 분석하였다.

그림 8은 실제 구현된 서비스의 출력 예시이다. 해당 사례에서 딥페이크 확률은 0.9396, 과장 광고 확률은 0.9632로 산출되어 두 임계값을 모두 초과하였으므로 High Risk 단계로 분류되었으며, 이에 따른 광고 신뢰도 점수는 약 5% 수준으로 매우 낮게 산출되었다. 이는 본 시스템이 복합적인 조작이 가해진 고위험 광고를 효과적으로 식별하고, 이를 사용자에게 명확한 경고 형태로 전달할 수 있음을 보여준다.

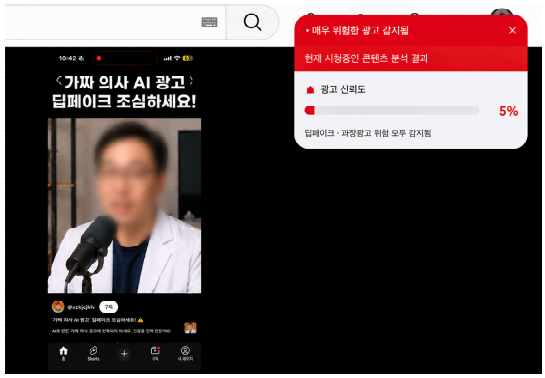

그림 8. 딥페이크 & 과장광고 탐지 결과
Fig. 8. Results of deepfake and exaggerated advertising detection

V. 결 론

본 논문은 온라인 광고 환경에서 증가하고 있는 딥페이크 악용 영상과 과장 광고의 위험에 효과적

으로 대응하기 위해, 텍스트와 영상 분석을 결합한 멀티모달 기반 통합 탐지 프레임워크를 제안하였다. 제안된 시스템은 KcELECTRA를 활용하여 광고 문구의 언어적 과장성을 판별하고, MobileNetV3 - LSTM 하이브리드 모델을 통해 영상 내 얼굴의 시공간적 조작 여부를 분석한 후, 두 분석 결과를 통합하여 콘텐츠의 종합적인 위험 수준을 산출한다.

실험 결과, 제안된 프레임워크는 텍스트와 영상 정보를 통합적으로 분석함으로써 자동화된 온라인 광고 모니터링 시스템으로서의 적용 가능성을 제시하였다. 특히 영상 정보만을 활용하는 경우에는 딥페이크 여부는 판단할 수 있으나 광고 내용의 과장성이나 규제 위반 여부를 판단하기 어렵고, 반대로 텍스트 정보만을 활용하는 경우에는 광고 문구의 과장성은 판별할 수 있으나 발화자의 신뢰성 또는 딥페이크 여부를 확인하기 어렵다는 한계가 있다. 이에 본 연구에서는 영상 기반 딥페이크 탐지와 텍스트 기반 과장 광고 판별 결과를 결합한 멀티모달 분석을 통해 온라인 광고의 위험도를 보다 종합적으로 판단할 수 있음을 확인하였다.

본 연구는 건강기능식품 광고를 중심으로 검증을 수행하였으나, 효능 과장 및 전문가 권위 악용이 빈번한 의약품, 화장품, 의료기기 등 유관 건강 산업 분야로의 확장 적용이 가능하다. 특히 이들 산업군은 광고 표현 양식과 소비자 기만 패턴이 유사하므로, 전이 학습(Transfer learning)을 통해 본 프레임워크를 효율적으로 최적화할 수 있을 것으로 기대된다.

향후 연구에서는 디퓨전 모델 기반 고품질 합성 영상에 대응하기 위해, 해당 유형의 위조 데이터를 포함한 확장 데이터셋을 구축할 계획이다. 특히 디퓨전 기반 생성 영상에서 나타나는 주파수 영역의 스펙트럼 왜곡, 노이즈 분포의 비자연적 패턴, 생성 단계에서 발생하는 미세한 잔여 아티팩트(Residual artifacts) 특성을 반영한 탐지 전략을 설계하고자 한다. 이를 위해 공간적 특징 기반 분석에 더하여 주파수 영역 특징 추출 및 시간적 일관성(Temporal coherence) 검증을 통합하고, rPPG 기반 생체 신호 분석을 결합한 다층적 탐지 구조를 구축할 예정이다. 또한 적대적 학습(Adversarial training)과 오픈셋(Open-set) 탐지 기법을 도입하여 학습되지 않은 신규 생성 모델에 대해서도 일반화 성능을 유지할 수 있는 강건한 프레임워크로 확장하고, 빠르게 진화하는 생성 모델 환경에서도 지속적인 대응이 가능하도록 진행될 예정이다.

종합적으로, 본 연구는 경량화된 영상 분석 모델과 한국어 특화 언어 모델을 결합하여 실시간성이 요구되는 온라인 환경에서 딥페이크와 과장 광고를 동시에 탐지할 수 있는 기술적 기반을 제시하였다. 이는 사후 처벌 중심의 기존 규제 체계를 보완하고, 소비자 피해를 사전에 예방하는 능동적 디지털 광고 생태계 구축에 기여할 수 있을 것으로 기대된다.

Acknowledgements

모든 저자들은 이 연구에 동등하게 기여하였음을 밝히는 바입니다.

References

[1] J. Lee and H. Oh, "A comparative study on deepfake-based disinformation response policies toward recommendations for South Korea", Journal of the Korean Society for Information Management, Vol. 42, No. 1, pp. 155-182, Mar. 2025. https://doi.org/10.3743/KOSIM.2025.42.1.155.

[2] S. Jung, "A study on consumer's deception effect of expression style on dishonest advertisement and brand attachment", The Korean Journal of Advertising, Vol. 22, No. 1, pp. 303-333, Mar. 2011.

[3] Kyunghyang Shinmun, https://www.khan.co.kr/article/202510211404001. [accessed: Jan. 28, 2026]

[4] N. Um, "Millennial consumers' attitude toward SNS false and exaggerative advertising through in-depth interview", Journal of Digital Convergence, Vol. 18, No. 10, pp. 459-467, Oct. 2020. https://doi.org/10.14400/JDC.2020.18.10.459.

[5] S. Park, "System development to analyze risk alert of deceptive and exaggerated advertisements based on intelligent cloud", The Journal of the Institute

of Internet, Broadcasting and Communication, Vol. 25, No. 1, pp. 215-225, Jan. 2025. https://doi.org/10.7236/JIIBC.2025.25.1.215.

[6] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators", arXiv preprint arXiv:2003.10555, pp. 1-18, Mar. 2020. https://doi.org/10.48550/arXiv.2003.10555.

[7] J. Kim and J. Bang, "Development of a sentiment classification and keyword extraction system based on chatbot and sentiment analysis model", Journal of the Korea Institute of Information and Communication Engineering, Vol. 29, No. 10, pp. 1302-1311, Oct. 2025. https://doi.org/10.6109/JKIICE.2025.29.10.1302.

[8] S. Jun and M. Noh, "Attribute-based sentiment analysis using online fashion review data", Korean Business Education Review, Vol. 40, No. 1, pp. 183-196, Feb. 2025. https://doi.org/10.23839/KABE.2025.40.1.183.

[9] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3", Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, pp. 1314-1324, Oct. 2019. https://doi.org/10.1109/ICCV.2019.00140.

[10] Y. Han, E. Kim, J. Bang, Y. Bae, J. Ok, J. Jeong, G. Hwang, J. Park, H. Park, W. Choi, and B. Lee, "Development of a Lightweight Deep Learning-Based Mobile Application for Pet Health Management", Proc. of KIIT Conference, Jeju, Korea, pp. 1152-1156, Jun. 2025.

[11] S. Jang, M. Lee, and T. Son, "Performance optimization of a lightweight MobileNet-based deepfake detection model", Journal of Defense and Security, Vol. 7, No. 1, pp. 388-403, Jul. 2025.

[12] S. Choi and S. Lee, "A Study on Comparison of Performance by Deepfake Detection Models Based on Dataset", Proc. of KIIT Conference, Jeju, Korea, pp. 41-43, Jun. 2025.

[13] T. Hwang, H.-J. Yu, and J.-R. Kim, "Emergency situation detection and speech recognition enhancement utilizing Whisper", The Journal of the Acoustical Society of Korea, Vol. 44, No. 2, pp. 132-143, Mar. 2025. https://doi.org/10.7776/ASK.2025.44.2.132.

[14] Ministry of Food and Drug Safety, https://www.mfds.go.kr/brd/m_218/view.do?seq=1154. [accessed: Jan. 28, 2026]

[15] P. Kwon, J. You, G. Nam, S. Park, and G. Chae, "KoDF: A large-scale Korean deepfake detection dataset", 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 10724-10733, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.01057.

[16] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images", Proceedings of the IEEE/CVF international conference on computer vision (ICCV), Seoul, Korea, pp. 1-11, Oct. 2019. https://doi.org/10.1109/ICCV.2019.00009.

[17] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks", IEEE Signal Processing Letters, Vol. 23, No. 10, pp. 1499-1503, Aug. 2016. https://doi.org/10.1109/LSP.2016.2603342.

[18] D. Nguyen, M. Astrid, A. Kacem, E. Ghorbel, and D. Aouada, "Vulnerability-Aware Spatio-Temporal Learning for Generalizable and Interpretable Deepfake Video Detection", Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10786-10796, Jan. 2025.

[19] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks", Proceedings of the 15th IEEE International Conference on Advanced Video and Signal Based

Surveillance (AVSS), Auckland, New Zealand, pp. 1-6, Nov. 2018. https://doi.org/10.1109/AVSS.2018.8639163.

[20] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", Neural Computation, Vol. 9, No. 8, pp. 1735-1780, Nov. 1997. https://doi.org/10.1162/neco.1997.9.8.1735.

[21] K. Yang, "Transformer-based Korean pretrained language models: A survey on three years of progress", arXiv preprint arXiv: 2112.03014, pp. 1-7, Nov. 2021. https://doi.org/10.48550/arXiv.2112.03014.

[22] M. Pathan, S. Khairnar, S. Joshi, H. Malshikare, A. Kharkar, and D. Suryawanshi, "Deepfake detection using deep learning: ResNext and LSTM", Proc. of the 2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, pp. 1-5, Aug. 2024. https://doi.org/10.1109/ICCUBEA61740.2024.10775025.

[23] S. P. Koritala, M. Chimata, S. N. Polavarapu, B. S. Vangapandu, T. K. Gogineni, and V. M. Manikandan, "A deepfake detection technique using recurrent neural network and EfficientNet", Proc. of the 2024 15th International Conference on Computing Communication and Networking Technologies, Kamand, India, pp. 1-6, Jun. 2024. https://doi.org/10.1109/ICCCNT61001.2024.10723875.

[24] S. Ambika, Y. Harini, and B. R. Sumanth, "Real-time deepfake detection using a hybrid MobileNet-LSTM model for image and video analysis", 2025 8th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, pp. 981-986, Jul. 2025. https://doi.org/10.1109/ICCMC65190.2025.11140856.

저자소개

등 덕 호 (Tehhao Teng)

2023년 8월 : 고려대학교 기계공학과(학사)
2025년 8월 : 성균관대학교 지능형로봇학과(석사)
2025년 12월 : 고려대학교 지능정보 SW아카데미 7기 수료(640H)
관심분야 : 컴퓨터 비전, 자연어 처리, 멀티모달 딥러닝

한 주 은 (Jueun Han)

2025년 12월 : 고려대학교 지능정보 SW아카데미 7기 수료(640H)
2022년 3월 ~ 현재 : 고려대학교 보건정책관리학부/ 뇌인지과학융합전공 학사과정
관심분야 : 딥러닝, 컴퓨터비전

임 채 림 (Chaerim Lim)

2024년 8월 : 전북대학교 바이오메디컬공학부(학사)
2025년 12월 : 고려대학교 지능정보 SW아카데미 7기 수료(640H)
관심분야 : 머신러닝, 딥러닝

박 현 호 (Hyunho Park)

2023년 7월 : The Hong Kong Polytechnic University, School of Accounting and Finance
2023년 10월 ~ 2025년 5월 : (주)니트머스 개발사업팀 소속
2025년 12월 : 고려대학교 지능정보 SW아카데미 7기 수료(640H)
관심분야 : 데이터 분석,데이터 엔지니어링

안 수 빈 (Subin Ahn)

2023년 2월 : 숙명여자대학교
글로벌서비스학부(학사)
2025년 12월 : 고려대학교 지능정보
SW아카데미 7기 수료(640H)
관심분야 : 머신러닝, 딥러닝

이 다 은 (Daeun Lee)

2020년 3월 ~ 2024 2월 :
중앙대학교 컴퓨터예술학부(학사)
2025년 12월 : 고려대학교 지능정보
SW아카데미 7기 수료(640H)
2025년 3월 ~ 현재 : 고려대학교
교육대학원 컴퓨터교육학과
관심분야 : 컴퓨터교육, 컴퓨터과학

유 길 상 (Gilsang Yoo)

2010년 2월 : 중앙대학교
영상공학과(박사)
2010년 3월 ~ 현재 :
(사)한국컴퓨터게임학회 부회장
2011년 3월 ~ 현재 : 고려대학교
정보대학 정보창의교육연구소/
지능정보 SW아카데미 교수
2023년 3월 ~ 현재 : (사)한국미디어 아트산업협회
수석부회장
관심분야 : 데이터사이언스, 데이터 시각화, 금융 데이터
분석, 3D영상 콘텐츠, 머신/딥러닝