한국데이터사이언스학회
Korean Association of Data Science

한국데이터사이언스학회
2026년도
하계 학술대회

PROCEEDINGS

2026. 6. 18.(목) – 6. 19.(금)

수원컨벤션센터

주최/주관 한신대학교 RISE사업단, 서울AI재단 Seoul AI Foundation

후원 NAVER, LIG, 대신정보통신(주), JIN INFRA (주)진인프라, SK broadband, 아이티센 엔텍

목 차

0041 AudioLDM2 Text-LOA 생성과정에서의 text condition 참조 경향 분석 7
김진형 (서강대학교), 박규태 (서강대학교), 김지환 (서강대학교)

0042 Whisper와 경량 보조 헤드를 활용한 음향 특성 융합 기반 마비말장애 중증도 분류 모델 9
박규태 (서강대학교), 김진형 (서강대학교), 김지환 (서강대학교)

0046 액티브 시니어의 기술효능감과 기술불안이 웨어러블 기기 행동의도에 미치는 영향: STAM모형을 중심으로 15
김지혜 (순천향대학교), 김미량 (순천향대학교)

0047 OCR+VLM 결합 기반 코딩 강의 영상의 파이썬 코드 스니펫 자동 추출 시스템 20
진세현 (성균관대학교), 김재현 (성균관대학교)

0064 LLM의 기후 사실 검증: 영-한 언어 격차와 조건별 성능 분석 22
이지훈 (한국전자기술연구원), 문재원 (한국전자기술연구원)

0059 공공 지역데이터의 교육 자원 적합성 평가 프레임워크 개발 26
김하영 (제주대학교), 조정원 (제주대학교)

0105 생성형 AI를 활용한 국가 R&D 성과물 이관 준비도 자동 평가 방법론: 개념적 프레임워크 30
정문권 (한국과학기술정책플랫폼협동조합), 이주영 (한국과학기술정책플랫폼협동조합)

0116 에이전트 기반 조직 행동 시뮬레이션을 통한 다원적 HR 정책 사전 검사 아키텍처 36
김성현 (성균관대학교), 김장현 (성균관대학교)

0010 반도체 CMP 공정의 MRR 예측 및 계측 절감을 위한 불확실성 보정 기반 선택적 가상계측 프레임워크 40
허석진 (한신대학교), 백수진 (한신대학교)

0013 긴급재난문자 및 지역 취약지표를 결합한 공공 사회위험 조기탐지 서비스 설계 46
김시현 (한신대학교), 안현 (한신대학교)

0030 공공데이터 기반 머신러닝을 활용한 전력 수요 예측 모델 비교 연구 55
지민석 (한신대학교), 백수진 (한신대학교)

0090 AI Agent 환경 내 능동형 접근제어 모델 61
홍현준 (고려대학교 세종캠퍼스), 최부영 (한신대학교), 이세영 (서강대학교), 고병수 (한신대학교)

0079 지역사회 시니어 체육 프로그램 참여와 주관적 건강 및 생활 만족도의 인과관계 분석 63
노유경 (한신대학교), 김유신 (한신대학교)

0080 프로야구 선수 평가 지표의 통계적 착시 현상 분석: 상대전적과 클러치 스탯을 중심으로 65
유준호 (한신대학교), 김유신 (한신대학교)

0081 시계열 데이터 분석을 활용한 미세먼지 농도 예측 연구 67
이연서 (한신대학교), 김유신 (한신대학교)

0020 규제 기반 ESG 보고 자동화를 위한 Compliance-Aware Prompting 프레임워크 제안 69
이해인 (동국대학교)

0025 가축분뇨 퇴비화 공정의 강화학습 기반 지능형 제어 시스템 개발 73
백유준 ((주) 지능로봇 스튜디오), 김상철 (전북대학교)

0028 Visual-Temporal Co-Attention for Gold ETF Trend Forecasting 78
SIV NASETH (전남대학교), 전재기 (전남대학교)

0011 사족보행 로봇의 청각 기반 상황 인지를 위한 자가소음 데이터베이스 구축 및 시간-주파수 특성 분석 79
신현재 (한신대학교), 김선만 (한신대학교)

0015 학사 규정 질의응답 모델의 오류 검토·교정 학습을 위한 합성 데이터셋 구축 84
박소영 (한신대학교), 김선만 (한신대학교)

0029 공중화장실 안전위생 통합관제용 AI 환경센서 개발 89
홍아름 (에어딥), 정우진 (에어딥), 김유신 (한신대학교)

0071 인공지능 역량은 기대와 우려를 동시에 높이는가? : 일반국민과 정보취약계층 6개 집단 비교 93
이명하 (순천향대학교)

0050 군집분석을 활용한 공공 흡연시설 설치 우선지역 선정 : 서울시 사례를 중심으로 96
하시윤 (한신대학교), 백수진 (한신대학교)

0052 기상 데이터 기반 돼지 구제역 발생 위험 분류 모델 101
조원희 (한신대학교), 백수진 (한신대학교)

0056 공개 EDA 데이터 기반 머신러닝과 U-Net을 활용한 반도체 라우팅 혼잡 Hotspot 예측 106
김영범 (한신대학교), 백수진 (한신대학교)

0072 청년 체크카드 소비 패턴과 금융상품 선호도 및 고객 이탈 분석 113
진성민 (한신대학교), 김유신 (한신대학교)

0075 미국 프로야구 MLB 투수 부상 위험 예측 모델 연구 115
김연승 (한신대학교), 김유신 (한신대학교)

0073 e스포츠 아시안 게임 국가대표 선출 지표 분석 118
김민준 (한신대학교), 김유신 (한신대학교)

0077 제조 설비 시계열 데이터를 활용한 이상 탐지 분석 120
최민영 (한신대학교), 김유신 (한신대학교)

0026 가축 분뇨 퇴비의 부숙도 지수 예측을 위한 머신러닝 기반 회귀모델 개발 및 평가 122
강희준 ((주) 로봇팜), 김승완 ((주) 그린맥스), 강희용 ((주) 로봇팜)

0037 중환자실 환자의 진정·진통 약물 처방 패턴을 활용한 장기 재원 예측 모델 개발 126
배지원 (건양대학교병원), 이승희 (건양대학교병원), 이지훈 (건양대학교병원)

0040 터치스크린과 가상 교실: 감염병 대유행이 견인한 영유아 발달 치료의 디지털 패러다임 쉬프트 130
유영의 (순천대학교), 장윤호 (청운대학교), 조보경 (순천향대학교), 김미량 (순천향대학교)

0060 모바일 브라우저 YOLO26 객체 탐지 파이프라인의 WebGPU–WASM 지연 분해 및 병목 분석 133
박찬빈 (호서대학교), 하동호 (호서대학교), 이호웅 (호서대학교)

0062 개인정보보호 및 상호운용성 향상을 위한 데이터 메시 기반 의료데이터 공유 아키텍처 구현 139
김현겸 (호서대학교), 조현태 (호서대학교), 이호웅 (호서대학교)

0065 이중 시점 융합 기반의 저조도 객체 분류 146
Muhammad Uzair (제주대학교), 김수균 (제주대학교)

0067 발달지연 유아 운동중재 연구의 지식 진화 분석: 키워드 네트워크 기반 연구동향 탐색 148
김아름 (순천향대학교), 김미량 (순천향대학교)

0069 공공데이터를 활용한 경기도 지역별 카드 소비 특성 분석 및 시각화 152
이도윤 (차의과학대학교), 정미현 (차의과학대학교)

0048 생활 습관 및 건강 지표 기반 당뇨 위험도 예측 모델 비교 분석 155
이성철 (한신대학교), 백수진 (한신대학교)

0111 AI Agent를 활용한 대학생 맞춤 정보 시스템 설계 및 구현에 대한 연구 161
최부영 (한신대학교), 배현욱 (한신대학교), 남재욱 (한신대학교), 전예은 (한신대학교), 홍승필 (한신대학교)

0112 모바일 환경에서 LLM모델의 양자화 스케일과 추론 지연·메모리 영향에 따른 인지 기반 최적화 167
손혁진 (한신대학교), 판반뚱럼 (한신대학교), 강민구 (CSU)

0110 사회 취약계층 건강 및 복지 서비스 효과성 평가 175
배유진 (한신대학교), 김유신 (한신대학교)

0117 히트곡 멜로디의 전개 및 화성 진행 패턴 데이터베이스 구축과 AI 작곡 보조 기법 178
허진명 (한신대학교), 김유신 (한신대학교)

0014 시계열 인과발견 결과의 실무적 해석을 위한 LPCMCI 기반 동적 요인 기여 프레임워크 181
김현우 (한동대학교), 최서현 (한동대학교), 김성은 (한동대학교)

0003 멀티모달 융합 매칭 기반 지능형 분실물 검색 시스템 191
김수연 (세종대학교), 김동호 (마크애니), 민병석 (세종대학교)

0012 119 신고 발화자의 마음위기 분류(C-CASA) 프레임워크 제안 연구 197
강미주 (가톨릭대학교), 이지영 (한신대학교), 정유진 (한신대학교), 홍승필 (한신대학교)

0018 Sentinel AI: 규칙 기반 엔진과 멀티모달 Local LLM을 결합한 시니어 실신·배회 탐지 지능형 관제 시스템 204
류찬 (한신대학교), 성낙준 (한신대학교)

0095 공공데이터 융합 기반 고령층 의료 이동 취약지역 발굴을 위한 K-MMVI 프레임워크(An Optimization Framework for Korea Medical Mobility Vulnerability Index (K-MMVI) Using Public Data Convergence) 209
신현진 (한신대학교), 김병재 (한신대학교)

0006 생성형 AI 인터페이스의 출처 검증 가능성과 신뢰도 표시 방식이 사용자 신뢰 교정에 미치는 영향: 인지 부하의 매개 효과를 중심으로 215
담서단 (성균관대학교), 김장현 (성균관대학교)

0021 한국어 단문 환경에서의 생성형 AI 작성 리뷰 탐지: KoBERT 기반 하이브리드 모델 218
이상철 (한국공학대학교), 이다예 (한국공학대학교), 이하영 (한국공학대학교), 남우민 (한국공학대학교), 박건철 (한국공학대학교)

0022 시민참여형 도시계획을 위한 생성형 AI 기반 퍼실리테이션 지원 시스템 제안 221
김정민 (한국공학대학교), Uprety Sudan Prasad (한국공학대학교), 남우민 (한국공학대학교), 박건철 (한국공학대학교)

0023 멀티 에이전트 토론과 가드레일 신뢰도 평가에 따른 다중 인공지능 답변 시스템 224
문태웅 (전남대학교), 이진형 (전남대학교), 나인섭 (전남대학교)

0114 생성형 AI 기반 비정형 정책데이터 분석을 위한 PRAF 프레임워크 개발 연구 226
이주영 (한국과학기술정책플랫폼협동조합)

0039 생성형 AI NPC의 VR 상호작용 제어를 위한 하네스 프레임워크 233
오석희 (한신대학교)

0061 바이브 코딩을 활용한 대학에서의 비IT전공자 대상 SW-AI 기초교육 방안 238
노해선 (성균관대학교), 김재현 (성균관대학교)

0078 학습자 상태 기반 협업형 LLM 에이전트 학습 지원 구조 설계(Design of a Learner State-Based Collaborative LLM Agent Architecture for Adaptive Learning Support) 243
이지원 (고려대학교) , 이가연 (한신대학교) , 진수빈 (한신대학교) , 강민구 (CSU)

0084 개발 과정 회고 보조를 위한 작업 중 어려움 탐지 및 로컬 LLM 기반 시스템 250
정영웅 (순천향대학교), 이하민 (순천향대학교), 홍민 (순천향대학교)

0099 Spring Boot 기반 온라인 학습 플랫폼의 설계 및 구현 256
왕홍우 (성균관대학교), 김장현 (성균관대학교)

0058 KLUE-BERT 기반 대학 행정조교 업무보조 시스템 259
전유빈 (순천향대학교), 김수현 (순천향대학교)

0051 올인원 IT 커리어 관리 플랫폼의 설계 및 구현 265
신미리 (한신대학교)

0104 DTW 알고리즘을 이용한 사용자 맞춤형 일본어 억양 교정 시스템 구현 268
전이현 (한신대학교), 김예나 (한신대학교), 성낙준 (한신대학교)

0009 연합학습 환경의 LLM 파인튜닝을 위한 누적 토큰 학습량 정합화 기반 학습 방법 연구 273
정진용 (가천대학교), 송인서 (가천대학교), 이강윤 (가천대학교)

0106 Evidential Deep Learning을 이용한 ABR 이미지 기반 난청 분류 및 불확실성 추정 277
권진오 (순천향대학교), 민지홍 (순천향대학교), 마준 (국립군산대학교), 홍민 (순천향대학교)

0032 온프레미스 프레젠테이션 기록, 요약 및 분석 시스템 283
조진만 (순천향대학교), 민지홍 (순천향대학교), 홍민 (순천향대학교)

0108 유방촬영술 기반 이진 분류에서 딥러닝 Backbone 아키텍처 성능 비교 분석 289
배준상 (순천향대학교), 김태헌 (순천향대학교), 민지홍 (순천향대학교), 홍민 (순천향대학교)

0034 Unity 엔진 기반 니파 바이러스 민·관 협력 대응 교육 시뮬레이션 시스템 개발 295
김상호 (순천향대학교), 강병훈 (순천향대학교), 권진오 (순천향대학교), 박민성 (순천향대학교), 전수영 (순천향대학교), 조현우 (순천향대학교), 정영웅 (순천향대학교), 배준상 (순천향대학교), 정현우 (순천향대학교), 민지홍 (순천향대학교), 이하민 (순천향대학교), 김태헌 (순천향대학교), 홍민 (순천향대학교)

0070 흉부 X-ray 기반 다중 병변 분류 모델의 Grad-CAM 기반 약지도 병변 위치 추정 연구 301
박현진 (한신대학교), 진수빈 (한신대학교), 전예은 (한신대학교), 류찬 (한신대학교), 마준 (국립군산대학교), 성낙준 (한신대학교)

0098 검증가능성의 역설: 소프트웨어 개발에서 쉬운 검증은 왜 AI 보조적 위임을 촉진하고 AI 주도적 위임에는 저항하는가 305
사박문 (성균관대학교), 김장현 (성균관대학교)

0031 Lawsuit: RAG 기반 민법 조문 환각 방지 시스템 310
전예은 (한신대학교), 진수빈 (한신대학교), 박현진 (한신대학교), 성낙준 (한신대학교)

0113 KNHANES 기반 근로 청년층의 의료 미충족 예측 모델 개발 및 고위험군 특성 분석 314
김명하 (차의과학대학교), 박한결 (차의과학대학교), 강정엽 (차의과학대학교), 김서윤 (차의과학대학교), 김은수 (차의과학대학교), 이상민 (차의과학대학교)

0101 성인학습이론에 기반한 비전공자 시니어 대상 파이썬 교육의 문제점 분석 및 생성형 AI 활용 개선 방안 연구 321
이준석 (성균관대학교), 김장현 (성균관대학교)

0103 전통시장 활성화를 위한 가상 방문객 기반 유동인구 최적화 시뮬레이션 연구 327
전이현 (한신대학교), 성낙준 (한신대학교)

0001 실내 포인트 클라우드 노이즈에 강건한 클라우드 서버 기반 3D LiDAR 공간 인식 시스템 332
박성재 (경기대학교), 김태훈 (경기대학교), 이세훈 (경기대학교), 김남기 (경기대학교), 안준호 (경기대학교)

0008 다기관 연합학습 클라이언트 환경 배포를 위한 가상화 방법 337
조성진 (가천대학교), 정민혁 (가천대학교), 이강윤 (가천대학교)

0043 기사 친화적 배차를 위한 AI 시뮬레이터 연구 : 인간 중심 알고리즘의 진화(A study on an AI simulator for Driver-Friendly DIspatching : Evolution of Human-Centered Algorithms) 342
김형주 (한신대학교), 정재영 (한신대학교), 이주성 ((주) 성화투어), 오준희 ((주) 성화투어), 신태환 ((주) 성화투어), 강민구 (CSU), 류승택 (한신대학교)

0076 직무 역량 벡터와 강의평 감성 키워드 분석을 결합한 설명 가능한 교내 커리큘럼 추천 시스템 353
김나영 (한신대학교), 진수빈 (한신대학교), 황희원 (한신대학교), 박현진 (한신대학교), 이용걸 (한신대학교)

0053 태그 참조 LLM 기반 AI 문서 보안 시스템 358
최은우 (세종대학교), 박종원 (세종대학교), 김동호 (마크애니), 민병석 (세종대학교)

0038 합성이미지 기반 Downstream 분류 모델의 원본(Real) 샘플 멤버십 추론 위험 분석 364
김효주 (서울여자대학교), 이혁재 (명지대학교), 김고은 (중앙대학교), 박서정 (숙명여자대학교), 류봄 (이화여자대학교), 우서현 (중앙대학교)

0045 합성 행동 데이터와 이상 탐지 모델 기반 블록체인 스마트 컨트랙트를 활용한 청소년 게임 과몰입 통합 관리 시스템 연구 372
김지환, 설순욱 (한국기술교육대학교), 홍지수 (숙명여자고등학교), 임경현 (Kent School)

0063 설문 기반 개인정보 수집·암호화·자동 파기 보조 시스템 설계 및 구현 381
유승우 (순천향대학교), 김수현 (순천향대학교)

0017 주파수 융합과 가우시안 적응 주의를 활용한 위장 객체 분할 연구 387
한인철 (제주대학교), 김수균 (제주대학교)

0054 NER 기반 용어 제약과 소형 언어 모델 미세 조정을 통한 한→영 도메인 번역 389
정성우 (성균관대학교), 이창준 (성균관대학교)

0057 Distributed Labor Market Intelligence: NLP and Semantic Job Recommendation at Scale 393
Neelam Patidar (CSU), 우종욱 (CSU)

0066 간소화된 설명인자 기반 혼합용매 용해도 예측 기계학습 모델 개발 399
김정임 (단국대학교), 여승연 (동국대학교), 탁경재 (단국대학교)

AudioLDM2 Text-LOA 생성 과정에서의 Text Condition 참조 경향 분석

김진형[O1], 박규태[2], 김지환[2]
[O1]서강대학교 수학과
[2]서강대학교 컴퓨터공학과
e-mail: jadon0207@sogang.ac.kr [O1], kt8199@sogang.ac.kr[2], kimjihwan@sogang.ac.kr[2]

Analyzing Text Condition Usage in AudioLDM2 Text-LOA Generation

JinHyeong Kim[O1], GyuTae Park[2], JiHwan Kim[2]
[O1]Dept. of Mathematics, Sogang University
[2]Dept. of Computer Science, Sogang University

요 약

본 연구는 AudioLDM2의 Text-LOA 생성 과정에서 GPT-2 기반 LOA generator가 text condition을 어떻게 참조하는지 분석하였다. 이를 위해 generated LOA vector 위치에서 condition prefix 위치로 향하는 self-attention을 추출하고, CLAP condition과 FLAN-T5 condition의 참조 비중을 비교하였다. AudioCaps test description을 대상으로 분석 결과, Text-LOA 생성 과정에서 CLAP condition이 FLAN-T5 condition보다 상대적으로 더 높은 비중으로 참조되었다. 또한 FLAN-T5 condition 내부에서는 동사가 가장 높은 attention weight를 보였다.

키워드: AudioLDM2, Text-LOA, Text-to-Audio Generation, CLAP, FLAN-T5

I. 서론

Text-to-Audio 생성 모델은 입력 텍스트 조건을 바탕으로 오디오를 생성하므로, 텍스트 조건이 중간 생성 과정에서 어떻게 참조되는지 이해하는 것이 중요하다. 실제 생성 결과에서 입력 텍스트에 포함된 세부 조건이 충분히 반영되지 않는 경우가 있으며, 이러한 한계가 어느 단계에서 발생하는지 분석할 필요가 있다. AudioLDM2 는 CLAP 및 FLAN-T5 기반 text condition 을 활용하여 Text-LOA(Language of Audio)를 생성하고, 이를 latent diffusion model 의 조건으로 사용한다[1]. 본

※이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구결과임 (No. RS-2025-25441313, 멀티모달 AI 에이전트 시대에 적합한 실무형 AI인재 육성 프로그램)

연구는 AudioLDM2 의 Text-LOA 생성 과정에서 CLAP condition 과 FLAN-T5 token condition 이 어떻게 참조되는지 분석한다.

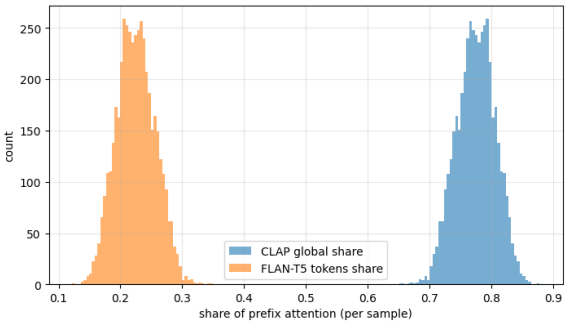


그림 1. Text condition별 attention share 분포

표1. Text description의 품사별 평균 attention weight

품사	Attention weight
Verb	0.084
Noun	0.073
Adj	0.071
Adv	0.051

II. 제안방법

본 연구는 AudioCaps test set 에 포함된 4,581 개의 text description 을 대상으로 AudioLDM2 의 Text-LOA 생성 단계에서 GPT-2 의 self-attention 을 이용해 condition-to-LOA attention 을 계산하였다[2]. Generated LOA vector 위치를 target position, condition prefix 위치를 source position 으로 설정하고, 각 LOA vector 가 CLAP global condition, FLAN-T5 token condition 을 어느 정도 참조하는지 계산하였다. 분석 지표로는 condition prefix 내부에서 CLAP condition 이 차지하는 비율을 CLAP condition share 로, FLAN-T5 token condition 이 차지하는 비율을 token condition share 로 정의하였다. 또한 FLAN-T5 token condition 은 품사별로 분류하여 평균 attention 을 비교하였다.

III. 결론

분석 결과, Text-LOA 생성 과정에서는 FLAN-T5 condition 보다 CLAP condition 이 상대적으로 더 높은 비중으로 참조되었으며, token condition 내부에서는 동사가 가장 높은 attention weight 를 보였다. 이는 LOA Generator 가 개별 token 수준의 세부 정보보다 입력 텍스트의 전반적인 의미를 요약한 global representation 에 더 크게 의존하는 경향이 있음을 시사한다. 또한 동사에 대한 높은 attention 은 Text-LOA 생성 과정에서 동작 정보가 상대적으로 중요하게 참조될 가능성을 보여준다.
향후에는 이러한 condition 참조 경향이 latent diffusion model 을 거쳐 생성된 최종 오디오에 어떻게 반영되는지 분석하고, 입력 텍스트의 세부 문맥을 더 효과적으로 반영할 수 있는 LOA 생성 방식을 연구할 필요가 있다.

참고문헌

[1] H. Liu et al., “AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871-2883, 2024.

[2] C. Kim et al., “AudioCaps: Generating Captions for Audios in the Wild,” Proc. NAACL-HLT, pp. 119-132, 2019.

※이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구결과임 (No. RS-2025-25441313, 멀티모달 AI 에이전트 시대에 적합한 실무형 AI인재 육성 프로그램)

Whisper와 경량 보조 헤드를 활용한 음향 특성 융합 기반 마비말장애 중증도 분류 모델

박규태[O(발표자)], 김진형[*], 김지환[**]
[O,**]서강대학교 컴퓨터공학과
[*]서강대학교 수학과
e-mail: {kt8199 [O], jadon0207[*], kimjihwan[**]}@sogang.ac.kr

Acoustic Feature Fusion-Based Dysarthria Severity Classification Model Using Whisper and Lightweight Auxiliary Heads

Gyutae Park[O], Jinhyeong Kim[*], Jihwan Kim[**]
[O, **]Dept. of Computer Science and Engineering, Sogang University
[*]Dept. of Mathematics, Sogang University

요 약

본 논문은 Whisper Encoder의 계층별 음성 표현과 명료도 및 발화 시간 기반 음향 특성을 융합하여 마비말장애 중증도를 분류하는 방법을 제안한다. 제안 모델은 Whisper의 계층별 음성 표현들을 활용하는 동시에, 각 헤드 및 음향 분석 등을 통해 얻은 13개 특성들을 추출하여 하나의 벡터로 결합한다. 이후 Temporal Transformer와 Layer Transformer를 순차적으로 적용하여 시간적 발화 패턴과 계층별 표현 정보를 통합한다. 이를 통해 사전학습 음성 모델의 일반적인 표현 능력과 마비말장애 음성의 특수한 음향 단서를 함께 활용하는 중증도 분류 구조를 구성한다.

키워드: 마비말장애, 심각도 분류, XAI, Whisper, 음향 특성 융합

I. 서론

마비말장애는 말운동장애 중 하나로, 의사소통과 삶의 질에 큰 영향을 미친다. 중증도 평가는 재활치료 계획 수립에 중요하지만 주로 임상의의 청지각적 판단에 의존해 시간과 비용이 많이 소요된다[1]. 이를 보완하기 위해 자동 중증도 분류 연구가 활발히 진행되고 있다. 본 연구는 Whisper[2]의 단일 사전학습 음성 표현을 활용한 음향 특성 융합 모델을 제안한다.

※ 이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획 평가원의 지원을 받아 수행된 연구임(No. 2022-0-00621, RS-2022- II220621,대화 기반 설명가능성을 멀티모달로 제공하는 인공지능 기술 개발)

II. 관련 연구

기존 마비말장애 중증도 분류 연구는 단일 사전 학습 모델 기반의 인코더에서 추출한 표현을 활용하여 분류를 하는 연구가 진행되었다[3][4]. 그러나 이러한 구조는 모델 예측 결과에 대한 설명성이 부족하였기에 단순 사전 학습 모델 인코더의 표현만을 활용하지 않고 발화 속도, 무음 구간, 명료도 등의 음향 특성과 멜 스펙트로그램, 원본 오디오를 결합한 다중 입력 융합 구조가 연구되었다[5]. 한편, 최근에는 Whisper 등의 사전 학습 모델 인코더에 헤드들을 결합하여 다양한 음성 태스크를 경량화된 구조로 수행하는 연구가 진행되고 있다[6].

III. 제안방법

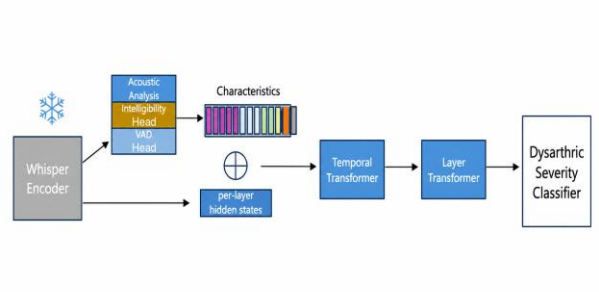


그림 1. Whisper 기반 음향 융합 모델 구조

기존 다중 입력 융합 구조는 복수의 특성 추출기와 인코더를 요구해 복잡한 구조와 함께 계산 비용이 증가한다. 이를 개선하기 위해 본 연구는 Whisper-Small Encoder를 단일 backbone으로 사용하고 그 위에 헤드를 결합한 구조를 제안한다. 구체적으로는 별도의 외부 분류 모델들을 따로 두는 대신 하나의 backbone 인코더 위에 VAD 헤드와 명료도 분류 헤드를 결합한다. 입력 음성에서 Whisper 계층별 표현벡터, 각 헤드들을 통해서 얻은 발화 시간 정보, 명료도 특성, 그리고 Praat 기반 음향 분석 결과를 추출하여 각 표현 값들을 결합하여 통합 표현 벡터들로 구성한 뒤, Temporal Transformer와 Layer Transformer로 통합하여 최종 마비말장애 중증도를 예측한다.

IV. 구현 및 결과

4.1 실험 데이터 및 평가 단위

실험은 마비말장애 환자가 수행한 단어 읽기 과제를 기반으로 하였으며, 모듈별 목적에 따라 VAD, 명료도, 최종 중증도 분류에 사용한 split과 평가 단위를 구분하였다. VAD 모듈은 Train 236개, Validation 37개, Test 42개로 구성되었고, 명료도 분류 모듈은 Train 247개, Validation 43개, Test 64개로 구성된 데이터를 사용하였다.

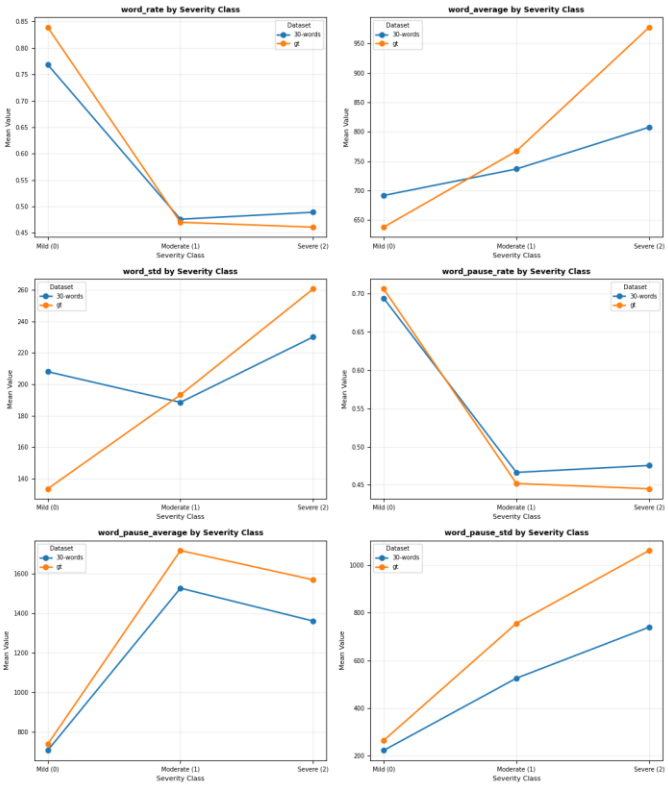


그림 2. GT와 VAD 결과 기반 결과 비교

※ 이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획 평가원의 지원을 받아 수행된 연구임(No. 2022-0-00621, RS-2022- II220621,대화 기반 설명가능성을 멀티모달로 제공하는 인공지능 기술 개발)

최종 중증도 분류기는 화자 기준 Train 231명, Validation 37명, Test 42명으로 구성된 split을 사용하였으며, 이는 발화 파일 기준으로 각각 282개, 45개, 55개에 해당한다.

최종 중증도 분류기 학습 시에는 학습이 완료된 각 헤드들의 체크포인트를 통합하여 사용하였다. 모든 음성 입력은 16 kHz로 변환하여 사용하였으며, backbone 모델로는 Whisper-Small Encoder를 사용하였다. VAD 헤드와 명료도 분류 헤드 학습 시에는 Whisper Encoder를 freeze하고, 각 task-specific head만 학습하였다.

4.2 모듈별 학습 설정 및 성능 평가

VAD 헤드는 Conv1D layer로 구성되며, Whisper Encoder의 마지막 특성 벡터를 입력으로 받아 각 frame의 음성유무를 이진 분류한다.

학습은 AdamW optimizer를 사용하여 learning rate 1e-3, batch size 4로 진행하였고, 총 60 epoch 동안 학습하였다.

이후 VAD의 frame-level 예측 결과를 음성 구간 단위로 변환하기 위해 후처리 파라미터를 조정하였다. 단어 읽기 데이터는 개별 발화가 짧고 단어 사이의 침묵 구간이 매우 짧게 나타나는 경우가 많다. 이에 따라 Validation set을 기준으로 onset threshold는 0.55, offset threshold는 0.50으로 설정하였고, smoothing kernel size는 1로 두어 과도한 평활화를 방지하였다.

또한 짧은 단어와 단어 사이의 gap을 보존하기 위해 최소 침묵 길이와 최소 음성 길이는 모두 0.0초로 설정하였으며, 짧은 음성 peak를 유지하기 위한 threshold는 0.65로 설정하였다. segment 전후 padding 역시 0.0초로 두어 예측 구간이 실제 단어 경계를 불필요하게 확장하지 않도록 하였다. 이러한 설정을 적용한 결과 VAD 헤드는 Test set에서 frame-level Accuracy 93.86%, segment IoU 87.84%를 보였다.

해당 VAD 결과로 발화 구간 수, 평균 발화 길이, 발화 길이의 표준편차, 쉼 구간 수, 평균 쉼 길이, 쉼 길이의 표준편차를 계산하였다. 그림 2는 실제 음성 구간 GT와 후처리된 VAD 예측 결과를 비교한 그림이다. 이를 통해 VAD 예측 구간이 GT 음성 구간과 유사하게 정렬되었으며, 이후 단어 단위 특성 추출 및 중증도 예측에 사용되는 음성 구간을 보다 안정적으로 정의할 수 있었다.

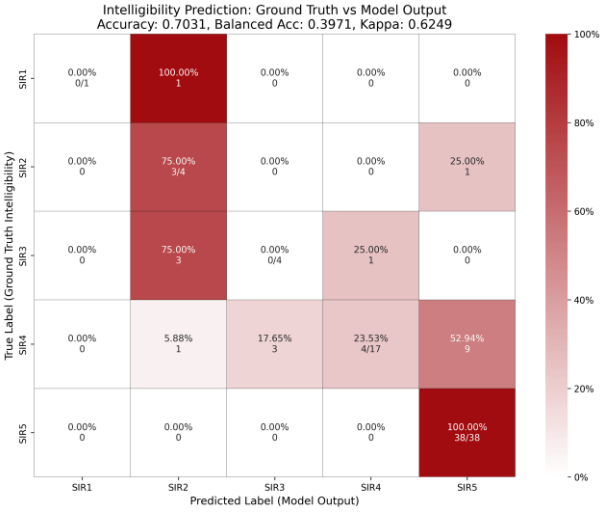


그림 3. 명료도 예측 성능

명료도 분류 헤드는 동일하게 Whisper Encoder의 마지막 특성 벡터를 입력으로 사용하며, Attention Pooling을 거쳐 SIR1부터 SIR5까지의 5개 명료도 등급을 분류한다. 명료도 분류 헤드는 AdamW optimizer를 사용하여 learning rate 1e-3, Train batch size 32, Validation/Test batch size 16으로 학습하였다. 또한 NoamAnnealing scheduler를 적용하였으며, warmup step은 800으로 설정하였다. 최종적으로 Test set 64개 샘플 기준 Accuracy 70.31%의 성능을 보였다.

그림 3을 통해 전체 Accuracy뿐만 아니라 각 명료도 등급별 예측 경향을 확인하였다. SIR5와 같이

※ 이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획 평가원의 지원을 받아 수행된 연구임(No. 2022-0-00621, RS-2022- II220621,대화 기반 설명가능성을 멀티모달로 제공하는 인공지능 기술 개발)

명료도가 높은 발화는 비교적 안정적으로 분류되었으나, 일부 낮은 명료도 등급에서는 인접 등급으로 예측되는 오류가 관찰되었다.

Praat 기반 음성 특성으로는 F0 변동, F0 표준편차, 평균 intensity, intensity 변동, 최대 intensity를 사용하였다. Praat 기반 특성은 마비말장애 발화에서 나타나는 pitch 및 energy 변동성을 반영하고, VAD 기반 특성은 발화 지속 시간, 발화 구간의 변동성, 평균 쉼의 길이 등을 통해 발화 속도 저하나 비정상적인 쉼 패턴을 반영한다. 이러한 특성들은 min-max scaling 후 Whisper Encoder의 표현 벡터와 결합하여 최종 중증도 분류기의 입력으로 사용하였다.

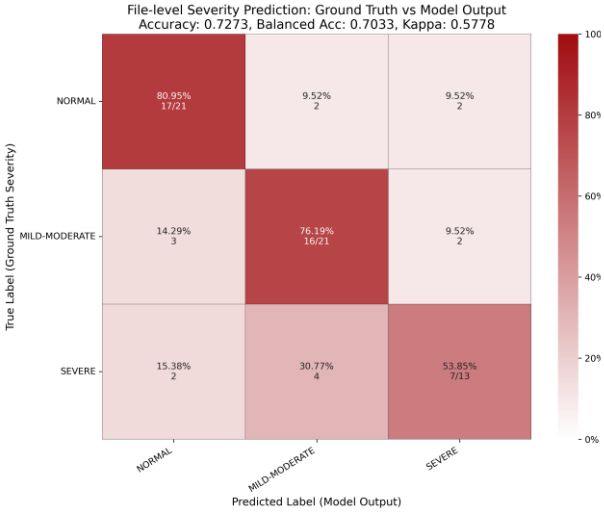


그림 4. File-level 중증도 예측 성능

최종 중증도 분류기에서는 사전 학습된 VAD 헤드와 명료도 분류 헤드의 체크포인트를 불러온 뒤, 해당 모듈들을 freeze하여 특성 추출기로 사용하였다. 최종 중증도 분류기의 입력 특성은 명료도 예측값, VAD 기반 발화 및 쉼 특성, Praat 기반 F0 및 intensity 특성, 그리고 발화자의 성별 정보로 구성하였다. 각 특성은 스칼라 값으로 정규화하여 사용하였으며, 전체 구성은 표 1과 같이 총 13개 특성으로 이루어진다.

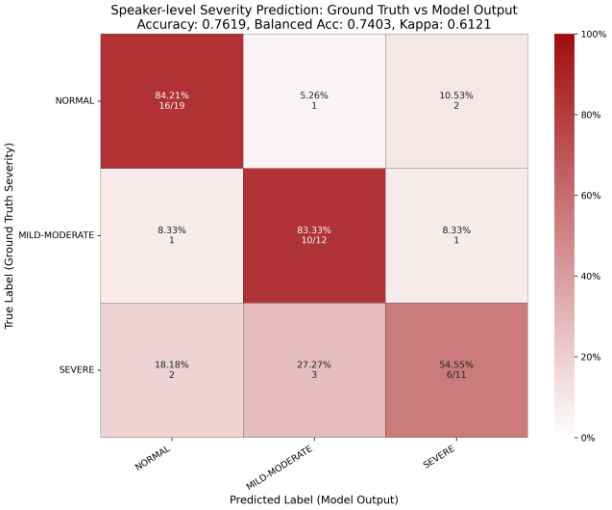


그림 5. Speaker-level 중증도 예측 성능

표 1. 최종 중증도 분류기에 사용된 입력 특성

특성 이름	특성 설명
성별	발화자의 성별 정보를 스칼라 값으로 반영한 특성
명료도 예측값	명료도 분류 헤드가 예측한 SIR 등급 정보를 반영한 특성
발화 구간 수	초 당 발화 개수
평균 발화 길이	검출된 각 발화 구간 길이의 평균
발화 길이 표준편차	발화 구간 길이의 변동성을 나타내는 값
쉼 구간 수	초 당 쉼 개수
평균 쉼 길이	검출된 쉼 구간 길이의 평균
쉼 길이 표준편차	쉼 구간 길이의 변동성을 나타내는 값
F0 변동	발화 구간 내 기본주파

※ 이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획 평가원의 지원을 받아 수행된 연구임(No. 2022-0-00621, RS-2022- II220621,대화 기반 설명가능성을 멀티모달로 제공하는 인공지능 기술 개발)

	수의 변동 정도를 나타내는 특성
F0 표준편차	기본주파수 분포의 변동성을 나타내는 값
평균 intensity	발화의 평균 음성 에너지 크기를 나타내는 특성
Intensity 변동	발화 내 음성 에너지의 변화 정도를 나타내는 특성
최대 intensity	발화 중 관찰된 최대 음성 에너지 크기

이 중 명료도 예측값은 발화 내용의 전반적인 이해 가능성을 반영하며, VAD 기반 특성은 발화 지속 시간과 쉼 패턴을 통해 말운동 조절의 변화를 나타낸다. Praat 기반 F0 및 intensity 특성은 pitch와 energy의 변동성을 반영하여 음성 산출의 안정성을 보완적으로 설명한다. 따라서 13개 특성은 Whisper Encoder의 계층별 표현만으로 포착하기 어려운 발화 시간, 명료도, 음향적 변동 정보를 함께 제공하며, 최종 중증도 분류 성능을 보완하는 역할을 한다.

이후 Whisper Encoder의 각 계층에서 추출한 hidden representation에 13개 입력 특성을 결합하여 계층별 통합 표현을 구성하였다. Temporal Transformer는 각 계층 표현 내의 시간축 정보를 통합하고, Layer Transformer는 서로 다른 Whisper 계층에서 얻어진 표현 간의 정보를 결합하여 최종 중증도 분류에 사용하였다.

중증도 분류기의 최종 분류 헤드는 linear layer로 구성되었으며, AdamW optimizer를 사용하여 learning rate 1e-4, batch size 1, weight decay 0.01, warmup step 1200으로 70 epoch 학습하였다.

그림 4의 평가는 발화 파일 단위로 수행되었으며, Test set의 총 55개 발화 파일에 대해 Accuracy 72.73%, Balanced Accuracy 70.33%를 보였다.

그림 5의 평가는 42명 화자 기준으로 수행되었으며, 동일 화자에 속한 여러 발화 파일 단위 예측 결과를 voting하여 최종 중증도를 결정하였다. 그 결과 화자 단위 Accuracy는 76.19%, Balanced Accuracy는 74.03%로 향상되었다. 이는 단일 발화에서 발생하는 예외적인 오분류가 화자 단위 voting을 통해 일부 보정될 수 있음을 보여준다.

V. 결론

본 연구는 Whisper-Small Encoder와 헤드들을 활용한 경량화된 마비말장애 분류 모델을 제안하였다. 최종 중증도 분류에서는 발화 파일 단위 평가에서 Accuracy 약 72%, 화자 단위 평가에서는 약 76%를 보였다. 단일 Whisper 기반의 경량 구조에서의 분류 가능성을 확인하였다.

후속 연구로는 해당 구조에 SHAP[7] 등의 설명 가능성 기법을 적용하여, 제안 모델이 중증도 판단에 활용되는 음향 특성과 계층별 표현의 기여도를 분석할 예정이다.

참고문헌

[1] Pamela M. Enderby, “Frenchay Dysarthria Assessment,” San Diego: College Hill Press, 1983.

[2] Alec Radford. et al., “Robust Speech Recognition via Large-Scale Weak Supervision,” in Proc. ICML, PMLR, vol. 202, pp. 28492-28518, 2023.

[3] Sanket Rathod. et al., “Whisper Features for

※ 이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획 평가원의 지원을 받아 수행된 연구임(No. 2022-0-00621, RS-2022- II220621,대화 기반 설명가능성을 멀티모달로 제공하는 인공지능 기술 개발)

Dysarthric Severity-Level Classification," in Proc. Interspeech 2023, Dublin, Ireland, pp. 1523-1527, Aug. 2023.

[4] Gyutae Park. et al., "CNN-based Severity Classification of Korean Dysarthric Speech using Whisper Encoder," in Proc. 한국정보과학회 2025 한국소프트웨어종합학술대회 논문집, pp. 1464-1466, Dec. 2025.

[5] Minseo Kim. et al., "Attribution-Based XAI for Dysarthria Assessment: Validation of Acoustic Feature Attributions," in Proc. the CIKM 2025 Workshop on Human-Centric AI: From Explainability and Trustworthiness to Actionable Ethics, COEX, Seoul, Korea, Nov. 2025.

[6] Yuan Gong. et al., "Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers," in Proc. Interspeech 2023, Dublin, Ireland, pp. 2798-2802, Aug. 2023.

[7] Scott M. Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. NeurIPS, vol. 30, 2017.

※ 이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획 평가원의 지원을 받아 수행된 연구임(No. 2022-0-00621, RS-2022- II220621,대화 기반 설명가능성을 멀티모달로 제공하는 인공지능 기술 개발)

액티브 시니어의 기술효능감과 기술불안이 웨어러블 기기 행동의도에 미치는 영향: STAM 모형을 중심으로

김지혜[O], 김미량*
*순천향대학교 휴먼헬스케어융합연구소
[O]순천향대학교 스포츠응용산업학과
e-mail:{김미량, 김지혜} mlkim@sch.ac.kr*, jiji1024@sch.ac.kr[O]

The Effects of Technology Self-Efficacy and Technology Anxiety on Behavioral Intention to Use Wearable Devices among Active Seniors: Focusing on the STAM

Jihye Kim[O] , Miryang Kim*
[O]Human Healthcare Convergence Institute, Soonchunhyang University, Korea
*Dept. of Sport Industry Studies, Soonchunhyang University, Korea

요 약

본 연구의 목적은 고령층 기술수용모델(STAM)을 기반으로 심리적 선행 요인인 기술효능감과 기술불안이 지각된 용이성, 지각된 유용성과 행동의도에 미치는 구조적 관계를 규명하고자 한다. 이를 위해 국내 50~60대 장·노년층을 대상으로 설문조사를 실시하여 최종 356부의 유효 표본을 확보하였으며, SPSS와 AMOS 프로그램을 활용한 확인적 요인분석 및 구조방정식 모델링(SEM)을 통해 연구가설을 검증하였다. 분석 결과는 다음과 같다. 첫째, 액티브 시니어의 기술효능감은 지각된 용이성에 유의한 정(+)적 영향을 미치는 것으로 나타났다. 둘째, 액티브 시니어의 기술불안은 지각된 용이성에 통계적으로 유의한 영향을 미치지 않은 것으로 나타났다. 셋째, 액티브 시니어의 지각된 용이성은 지각된 유용성에 유의한 정(+)적 영향을 미치는 것으로 나타났다. 마지막으로, 액티브 시니어의 지각된 유용성은 행동의도에 유의한 정(+)적 영향을 미치는 것으로 나타났다. 본 연구는 고령층의 기술 수용 초기 단계에서 핵심적인 심리적 요인을 중심으로 검증했다는 점에서 학술적 의의가 있으며, 향후 시니어 친화적 디지털 헬스케어 서비스 보급 및 맞춤형 교육 프로그램 설계를 위한 실무적 기초자료를 제공하고자 한다.

키워드: 고령층 기술수용모델(STAM), 웨어러블 기기, 기술효능감, 기술불안, 행동의도

※ 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2025S1A5C2A02022632)

I. 서론

고령화 사회의 가속화로 시니어의 건강관리 및 삶의 질 향상이 중요해진 가운데, 웨어러블 기기는 실시간 건강 정보 제공을 통한 개인 맞춤형 관리 도구로서 질병 예방과 의료비 절감에 기여하고 있다(김지혜, 윤지인, 2025; 고대선, 2021). 그러나 고령층의 디지털 기술수용 수준은 타 세대 대비 낮으며, 심리적 부담과 기술불안으로 인한 디지털 격차가 심화되고 있다(한국지능정보사회진흥원, 2021, 2024).

기존 연구는 주로 TAM을 통해 지각된 유용성과 용이성을 중심으로 설명하고 있으나(Davis, 1989). 고령층의 기술효능감, 기술불안과 같은 심리적 변인을 설명하는 데 한계가 있다(Peek et al., 2014). 이에 본 연구는 고령층 기술수용모델(STAM; Chen & Chan, 2014)의 심리적 요인에 초점을 맞춰 액티브 시니어의 기술효능감, 기술불안, 지각된 용이성, 유용성, 행동의도에 미치는 구조적 관계를 실증적으로 검증하고자 한다. 이는 액티브 시니어의 친화적 디지털 헬스케어 서비스 개발 및 보급을 위한 기초자료를 제공하고자 한다.

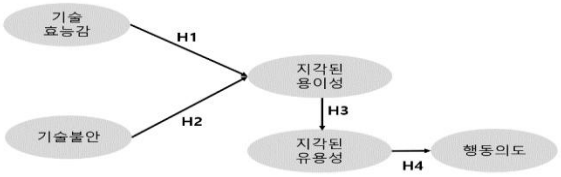


그림 1. 연구모형

II. 연구방법

2.1 연구대상자 및 조사도구

스마트 워치 및 스마트 밴드 사용 경험이 있는 50~60대 성인 남녀를 대상으로 유목적 표집법을 적용하였다. 전문 리서치 기관과 연구자 네트워크를 통해 온라인 설문을 실시하였으며 최종 356부의 데이터를 분석에 활용하였다.

5점 Likert 척도 설문지를 사용하였다. 기술효능감(3문항)은 Chen과 Chan(2014)이 재개념화하고 이진명 등(2025)이 활용한 척도를 수정 적용하였으며, 기술불안(2문항,역코딩)은 Rose와 Fogarty (2006) 및 Chen과 Chan(2014)의 척도를 박명아 등(2024)의 연구를 참고해 수정 활용하였다. 지각된 용이성(3문항)과 유용성(4문항) 은 Davis(1989)의 TAM 기반 척도를 참고해 수정하였고, 행동의도(4문항)는 Venkatesh et al.(2003)의 척도를 본 연구 목적에 맞게 보완하여 측정하였다.

2.2 타당도와 신뢰도 및 자료 처리

조사도구의 타당성 및 신뢰성 검증을 위해 확인적 요인분석을 실시한 결과, 적합도 지수 (χ^2=471.183, df=109, RMSEA=.087,TLI=.895,CFI=.905)가 권장 기준에 근접하거나 충족하여 측정모형의 타당성을 확보하였다(Hu & Bentler, 1999). 또한, 집중타당도 분석 결과 AVE 값(.519~.752)과 CR(.723~.924) 값이 모두 기준치를 상회하여 타당성이 입증되었다(김계수, 2007). Cronbach's α값 은 .723~.924로 나타나 모든 변수에서 내적 일관성과 신뢰성이 확보되었다(Nunnally & Bernstein,1994).

본 연구에서 수집된 설문 자료를 바탕으로 연구가설을 검증하기 위해 SPSS, AMOS을 활용하여 분석을 진행하였다. 연구대상자의 인구통계학적 특성을 파악하기 위해 빈도분석을 실시하였으며 조사도구의 타당성과 신뢰성을 파악하기 위해 확인적 요인분석(CFA)과 신뢰도 분석을 실시하였다. 각 변인간의 관련성을 살펴보기 Pearson의 상관관계 분석을 실시하였으며 본 연구의 가설 검증을 위해 구조방정식 모형(SEM)을 실시하였다.

III. 결과 및 논의

※ 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2025S1A5C2A02022632)

2)

상관관계 분석 결과, 기술효능감, 지각된 용이성, 지각된 유용성, 행동의도 간에는 유의한 정적 상관관계가 나타났으며, 기술불안은 이들 모든 변수와 유의한 부적 상관관계를 보였다. 특히 지각된 유용성과 행동의도 간의 상관관계(r=.676,p<.001)가 가장 높게 나타났다. 모든 잠재 변수 간 상관계수 값은 −.220~.676범위로,기준인 .80을 초과하지 않아 구성타당도에 문제가 없음이 확인되었다(Kline, 2013).

표1. 연구모형 적합도

	$\chi^2(df)$	RMSEA	TLI	CFI
측정모형	471.183(109)	.087	.895	.905
구조모형	486.527(113)	.086	.892	.901

액티브 시니어의 웨어러블 기기 기술수용요인 간 구조적 관계를 검증하기 위해 SEM 분석 결과, 적합도 지수(χ^2=486.527,CFI=.901,TLI=.892,RMSEA=.086)가 권장 기준을 충족하거나 근접하여 전반적으로 양호한 적합도를 확보한 것으로 판단되었다(Browne & Cudeck, 1993; Hair et al, 2019; Marsh, Balla, & McDonald, 1988).

표2. 가설 검증 결과

경로	B	β	S.E.	가설
기술효능감 → 지각된 용이성	.660	.728***	.072	채택
기술불안 → 지각된 용이성	.048	.067	.043	기각
지각된 용이성 → 지각된 유용성	.446	.532***	.061	채택
지각된 유용성 → 행동의도	1.066	.807***	.083	채택

***p<.001

첫째, 액티브 시니어의 기술효능감은 지각된 용이성에 유의한 정적 영향을 미치는 것으로 나타났다(β=.728, p<.001). 이는 시니어 사용자가 웨어러블 기술에 대해 '스스로 잘 다룰 수 있다'는 주관적 자신감을 가질수록, 기기 사용법을 더 쉽게 받아들이며 선행연구(이진명 외, 2025; 최봄이 외, 2021)를 통해 본 연구 결과를 지지하고 있다.

둘째, 액티브 시니어의 기술불안은 지각된 용이성에 통계적으로 유의한 영향을 미치지 않는 것으로 나타났다(β=.067, p=.265). 이는 기술불안을 주요 심리적 방해요인으로 간주했던 과거 연구들(박명아 외, 2024; Chen & Chan, 2014)과 대비되는 결과이다. 이러한 차이는 최근 웨어러블 기기의 UI/UX가 고령 친화적(Age-friendly)이고 직관적으로 단순화됨에 따라, 기술에 대한 불안감이 낮으며 더 이상 기술 수용에 방해요인이 되지 않음을 시사한다.

셋째, 액티브 시니어의 지각된 용이성은 지각된 유용성에 유의한 정적 영향을 미치는 것으로 나타났다(β=.532, p<.001). 이는 시니어 사용자가 웨어러블 기기의 조작이 쉽다고 인지할수록 해당 기술이 자신의 건강 및 체력 관리에 더 유용하다고 인식하는 것으로 나타났다. 이러한 결과는 TAM의 전통적인 메커니즘이 시니어 집단에도 견고하게 작동하고 있음을 재확인시켜 준다(Chen & Chan, 2014; Davis, 1989).

넷째, 액티브 시니어의 지각된 유용성은 행동의도에 정적 영향을 미치는 것으로 나타났으며 전체 연구 가설 검증에서 가장 영향력이 있는 핵심 요인으로 분석되었다(β=.807, p<.001). Chu, Chen과 Wang(2019)은 STAM에서 제시된 요인들이 시니어의 웨어러블 사용에 어떻게 작용하는지 탐색한 결과, 건강관리 기능의 유용성 때문에 웨어러블 기기를 사용하는 것으로 나타나 본 연구결과를 지지하고 있다.

본 연구는 액티브 시니어의 웨어러블 기기 수용에 있어 기술효능감의 촉진 효과와 지각된 유용성의 주된 영향력을 규명하였다. 특히 불안과 같은 부정적 정서를 완화하기보다 기술 활용에 대한 자신감, 긍정적 효능감의 강화가 이용만족과 행동의도를 높이는 데 더 유효함이 입증되었다(김지혜, 윤지인, 2025). 또한 신체 기능 향상 및 건강 증진과 같은 실질적 유용성의 제공이 액티브 시니어의 웨어러블 기기의 지속적인 사용에 결정적인 요인임을 재확인하였다.

※ 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2025S1A5C2A02022632)

IV. 결론 및 제언

본 연구의 결과를 바탕으로 액티브 시니어 친화적 웨어러블 기술 활성화를 위해, 제조사 및 개발사는 단순한 기능 확장을 넘어 만성질환 모니터링 및 응급 상황 대응과 같이 시니어가 실질적 효용을 체감할 수 있는 가치 중심의 서비스 설계에 집중하고, 직관적인 UI/UX 유지와 함께 지역사회 내 웨어러블 기기 체험 및 교육 프로그램을 통해 사용자의 기술효능감을 고취하는 실무적 전략을 수립해야 한다.

나아가 본 연구가 개인의 심리적 요인에 초점을 맞춰 고령자의 노화 관련 변인. 사회적 지지, 기술을 돕는 환경적 요인을 충분히 다루지 못한 점을 보완하기 위해, 향후 연구에서는 STAM의 핵심 변인인 사회적 영향과 촉진 조건을 통합하고 대상자의 건강 상태 및 질환 여부에 따른 세분화된 집단 비교 분석을 수행함으로써 액티브 시니어 기술 수용 모델의 설명력을 더욱 입체적으로 확장할 것을 제언한다.

참고문헌

[1] 고대선(2021). 고령자를 위한 스포츠 웨어러블 디바이스 이용의도에 미치는 요인 탐색. 한국웰니스학회지, 16(3), 149-155.

[2] 김지혜, 윤지인(2025). 액티브 시니어의 웨어러블 디바이스에 관한 기술주비도, 이용만족, 행동의도의 구조적 관계 분석. 한국여성체육학회지, 39, 1-11.

[3] 김계수(2007). New Amos 16.0 구조방정 식모형 분석. 서울: 한나래.

[4] 박명아, 이현주, 신혜리, 김영선(2024). 고령자의 디지털 금융서비스 기술수용 연구: 고령자 기술수용모형(STAM)을 중심으로, Journal of Information Technology Applications & Management, 31(3), 119-147.

[5] 이진명, 박성지, 전민선(2025). 고령자 기술수용모델(STAM)을 적용한 돌봄로봇 수용의도 결정요인 분석. 소비자학연구, 36(4), 145-167.

[6] 최봄이, 주수산나, 전혜정(2021). 노인의 웨어러블 로봇에 대한 지각된 유용성과 용이성 관련 요인. 한국노년학, 41(5), 825-841.

[7] 한국지능정보사회진흥원(2021). 2020 디지털장보격차 실태조사. 한국지능정보사회진흥원

[8] 한국지능정보사회진흥원(2024). 2023 디지털장보격차 실태조사. 한국지능정보사회진흥원

[9] Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing Structural Equation Models, 136-162. Sage.

[10] Chen, K., & Chan, A. (2014). Gerontechnology acceptance by elderly Hong Kong Chinese: a senior technology acceptance model(STAM). *Ergonomics, 57*, 635-652.

[11] Chu, Y., Chen, S., & Wang, M. (2019). Why do older adults use wearable devices: A case study adopting the Senior Technology Acceptance Model (STAM). *2019 Portland International Conference on Management of Engineering and Technology (PICMET)*, Portland, OR, USA.

[12] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-340.

[13] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis* (8th ed.). Cengage.

[14] Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives.

※ 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2025S1A5C2A02022632)

Structural equation modeling:a multidisciplinary journal, 6(1), 1-55.

[15] Kline, P. (2013). *Handbook of psychological testing.* Routledge.

[16] Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*(3), 391-410.

[17] Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric Theory (3rd ed.). McGraw-Hill.

[18] Peek, S. T., Wouters, E. J., Van Hoof, J., Luijkx, K. G., Boeije, H. R., & Vrijhoef, H. J. (2014). Factors influencing acceptance of technology for aging in place: a systematic review. *International Journal of Medical Informatics, 83*(4), 235-248.

[19] Rose, J., & Fogarty, G. J. (2006). Determinants of perceived usefulness and perceived ease of use in the technology acceptance model: senior consumers' adoption of self-service banking technologies. *In Proceedings of the 2nd biennial conference of the academy of world business, marketing and management development: Business across borders in the 21st century, 2*(10), 122-129.

[20] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly, 27*(3), 425-278.

※ 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2025S1A5C2A02022632)

OCR+VLM 결합 기반 코딩 강의 영상의 파이썬 코드 스니펫 자동 추출 시스템

진세현[O], 김재현[*]

[O]성균관대학교 컴퓨터교육학과

[*]성균관대학교 컴퓨터교육학과

e-mail:{chinsehyun[O], jaekim[*]}@skku.edu

OCR+VLM Integrated System for Automatic Python Code Snippet Extraction from Programming Lecture Videos

Sehyun Chin[O], Jaehyoun Kim[*]

[O]Dept. of Computer Education, SUNGKYUNKWAN University

[*]Dept. of Computer Education, SUNGKYUNKWAN University

요 약

본 연구는 코딩 강의 영상에서 학습자가 화면 속 코드를 직접 입력하기 위해 특정 구간을 반복 재생하거나 일시정지해야 하는 불편을 줄이기 위해 영상 내 파이썬 코드를 자동으로 추출하여 시간대별 복사 가능한 스니펫으로 제공하는 OCR+VLM 결합 시스템을 제안한다. 온라인 강의 영상은 코드 편집기, 강의 슬라이드, 자막, 터미널 출력 등이 한 프레임 안에 함께 나타나는 경우가 많다. 단순 OCR과 VLM을 단독으로 사용하는 방식은 설명 문구나 문제 텍스트 등을 코드로 오인하는 한계가 있다. 본 연구에서는 OpenCV 기반으로 코드 편집기 영역을 탐지하고 해당 영역에 OCR을 적용한 뒤 VLM을 이용해 코드 여부와 문맥을 최종 검토하는 파이프라인을 구성하였다. 실험 결과 OCR only와 VLM only는 코드가 아닌 텍스트까지 스니펫으로 출력하는 반면 OCR+VLM 방식은 실제 코드만 설별하는데 더 안정적인 경향을 보였다.

키워드: OCR, VLM, 코딩 강의, 코드 스니펫, 영상 분석

※ 본 논문은 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 첨단분야 혁신융합대학사업의 연구결과입니다.

I. 서론

온라인 코딩 강의에서는 강사가 직접 코드를 입력하며 설명을 하는 경우가 많아 학습자는 영상을 반복 재생하거나 일시정지하며 코드를 입력해야 한다. 특히 코드가 여러 단계에 걸쳐 수정, 완성되는 경우 코드 입력의 어려움이 커진다. 이에 본 연구는 강의 영상에서 파이썬 코드를 자동으로 추출하고 시간대별로 복사 가능한 스니펫을 제공하는 OCR+VLM 결합 시스템을 제안한다.

II. 관련 기술

OCR은 이미지나 영상 화면의 문자를 텍스트로 변환하는 기술이며 본 연구에서는 강의 영상에서 코드 후보를 추출하는 데 사용한다. VLM은 이미지와 텍스트 지시문을 함께 처리하여 화면 내용을 해석하는 모델이며 OCR 결과가 실제 코드인지 검토하는 데 활용한다. 본 연구는 OCR의 문자 인식 결과와 VLM의 화면 해석 능력을 결합하여 코드 외 텍스트를 줄이고자 한다.

III. 파이썬 코드 스니펫 자동 추출 시스템

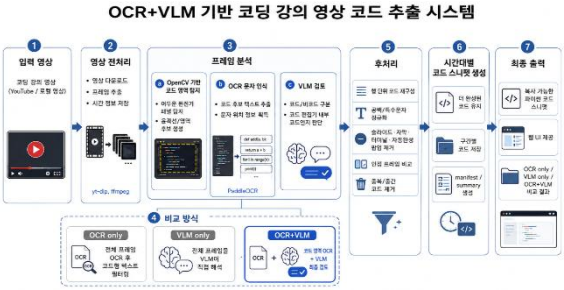


그림 1은 OCR과 VLM을 활용한 파이썬 코드 스니펫 자동 추출 시스템의 다이어그램이다.

IV. 실험 및 결과

OCR only, VLM only, OCR+VLM 세 방식을 비교한 결과, OCR only는 코드 외 텍스트가 함께 추출되었고 VLM only는 일부 화면 요소를 코드로 오인하였다. 반면 OCR+VLM은 실제 코드를 보다 안정적으로 추출하는 경향을 보였다.

V. 결론

본 논문에서는 OCR과 VLM을 활용하여 코딩 교육용 영상에서 코드 영역을 추출하는 시스템을 구현하였다. 실험 결과 실제 코드 추출이 가능함을 확인하였으며 향후에는 인식 정확도 향상을 위한 연구를 진행할 예정이다.

참고문헌

[1] 정영웅, 강병훈, 배준상, 손서영, 박정아, 김성엽, 홍민, "대학 행정 문서 데이터화를 위한 PaddleOCR 기반 전/후처리 파이프라인 설계 및 구현, " 한국데이터사이언스학회 동계 종합학술대회 논문집, pp. 88-91, 2025.

[2] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," Proceedings of the 39th International Conference on Machine Learning, Vol. 162, pp. 12888-12900, July 2022.

※ 본 논문은 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 첨단분야 혁신융합대학사업의 연구결과입니다.

LLM의 기후 사실 검증: 영-한 언어 격차와 조건별 성능 분석

이지훈*O, 문재원*
*O한국전자기술연구원 AI기반모델연구센터
e-mail: {jhlee31, jwmoon}@keti.re.kr*O

LLM Climate Fact Verification: English-Korean Gap and Condition-wise Performance Analysis

Jihoon Lee*O, Jaewon Moon*
*OAI Foundation Model Research Center, Korea Electronics Technology Institute

요 약

LLM은 한국어로 기후 사실을 검증할 때 영어보다 정확도가 낮다. 이 격차를 메우는 방법들의 효과를 한국어 특화 모델과 환경 도메인을 대상으로 함께 비교한 연구는 찾기 어렵다. 한국어 특화 모델 2종과 범용 다국어 모델 3종을 Poly-FEVER 한국어 주장과 Climate-FEVER 영어 근거에 적용해 모델·근거·교차언어·응답 유보 네 가지를 12개 조건 격자에서 비교했다. 동일 규모·동일 양자화에서 한국어 특화 모델이 한국어 입력에 한정해 유의한 우위를 보였고 영어 근거가 가장 안정적인 회복 수단이었던 반면 교차언어 추론은 효과가 없었다. 한국어 특화 효과가 모델 크기로 환원되지 않음을 확인했고, 향후 NEI를 포함한 3-class 확장과 환경 도메인 전반의 일반화가 후속 과제로 남는다.

키워드: 다국어 LLM, 기후 사실 검증, 한국어 평가, 응답 유보, 신뢰가능 AI

I. 서론

LLM이 정보 탐색에 일상적으로 쓰이면서 모델이 내놓는 사실의 정확성이 사회적 문제로 떠올랐다. 환경·기후 정보는 잘못된 답변의 파급이 정책 신뢰에 직접 닿는 영역이라 위험이 더 크다. 한국어 입력에서 LLM의 사실 검증 정확도는 영어 대비 낮다[1]. 이 격차가 단순 번역 품질이 아니라 모델의 언어별 이해 차이에서 비롯된다는 분석도 있다[2]. 그

※이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2024-00460011, 인류세의 기후테크 역량 강화를 위한 기후·환경 데이터 구축·처리 플랫폼)

러나 한국어 LLM의 신뢰성·환각 평가는 일반 지식 도메인에 머물러 환경·기후를 다루지 않았다. 다국어 응답 유보와 교차언어 추론도 각각 제안됐을 뿐, 사실 검증·환경·한국어 특화 모델을 함께 놓고 본 평가는 검색 가능한 범위에서 비어 있다.

알려진 네 가지 평가 조건(한국어 특화 모델 선택, 영어 근거 제공, 교차언어 추론, 응답 유보)을 같은 격자에서 비교해 이 빈자리를 메우고자 한다. EXAONE을 동급 범용 모델과 양자화 정밀도까지 정렬해 비교함으로써 모델 크기 효과로부터 한국어 특화 효과를 분리한다.

II. 방법론

한국어 LLM의 기후 사실 검증에서 영-한 격차를 측정하고, 이를 다루는 네 가지 조건의 효과를 같은 격자에서 비교한다. 네 조건은 모두 다국어 사실 검증 맥락에서 선행 연구가 논의한 것들이며, 평가는 추가 학습 없이 zero-shot으로 수행해 배포된 모델을 그대로 쓰는 실무 상황을 반영한다.

2.1 데이터 및 모델

Poly-FEVER의 한국어·영어 주장과 Climate-FEVER의 영어 근거(evidence)를 결합했다[3]. 영어 주장 텍스트의 정규화 매칭으로 두 데이터를 연결해 SUPPORTS 583건과 REFUTES 214건, 총 797건을 확보했으며 라벨 불균형으로 주 지표는 macro-F1로 둔다. 한국어 주장에 영어 근거를 결합한 구성은 환경 자료 다수가 영어 위키피디아에 있는 실무 RAG 환경을 반영한다. 평가 대상은 한국어 특화 EXAONE-4.0-32B와 1.2B[4], 범용 다국어 Qwen3-32B와 14B, Llama-3.1-8B-Instruct 다섯 가지다. 이 가운데 EXAONE-32B와 Qwen3-32B는 크기와 정밀도가 같고 EXAONE 1.2B와 32B는 같은 계열에서 크기만 다르다. 이 비교 설계로 한국어 특화와 크기 효과를 분리하고자 했다.

2.2 실험 조건 및 프롬프트

비교 조건은 모델 선택(한국어 특화·범용), 영어 근거 제공 여부, 교차언어 추론(한국어 주장에 영어 reasoning trace 유도), 응답 유보(UNCERTAIN 허용) 네 가지다. 근거와 교차언어를 조합해 6개 기본 조건을 만들고(표 1), 각각을 응답 유보 미허용·허용 두 가지로 분기해 총 12개가 된다. 모든 프롬프트는 출력 형식을 한 줄로 강제했다. 교차언어 추론은 모델 내부 사고 언어가 아니라 영어 reasoning trace 출력을 유도하는 프롬프트 조건으로 한정해 정의한다.

표 1. 6개 기본 실험 조건 구성

조건	언어	영어 근거	교차언어 추론
A	영어	제공	미적용
B	영어	미제공	미적용
C	한국어	제공	미적용
D	한국어	미제공	미적용
E	한국어	제공	적용
F	한국어	미제공	적용

2.3 평가 지표

모델 출력의 첫 줄에서 SUPPORTS, REFUTES, UNCERTAIN 세 라벨을 규칙 기반으로 추출했다. 주 지표는 macro-F1이며 정확도(accuracy)와 클래스별 F1을 보조로 본다. 조건별 효과는 영-한 차이(Δ_lang), 근거 효과(Δ_evid), 교차언어 효과(Δ_cl)로 측정한다. 응답 유보 조건에서는 응답률(Coverage), 유보율, 응답한 사례 중 환각률(hallucination rate)을 함께 보고한다. 각 격자의 macro-F1에 대해 797개 쌍을 1,000회 복원추출해

※이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2024-00460011, 인류세의 기후테크 역량 강화를 위한 기후·환경 데이터 구축·처리 플랫폼)

부트스트랩 95% 신뢰구간을 산출했고 동일 규모 비교는 paired 부트스트랩으로 신뢰구간이 0을 포함하는지 유의성을 판단했다.

III. 실험 결과

5개 모델로 12개 조건 전체를 zero-shot 추론한 결과를 사실성 평가, 한국어 특화 효과 검증, 불확실성과 환각 분석으로 정리한다.

3.1 조건별 사실성 비교

5개 모델의 6개 기본 조건 macro-F1을 표 2에 정리했다. 영-한 격차(Δ_lang)는 5개 모델 중 4개에서 양수로, 성능이 높은 모델에서 한국어 입력은 영어보다 정확도가 낮다. 예외는 EXAONE-1.2B로, 영어 무근거 조건이 0.409까지 떨어져 격차가 역전됐다. 근거 효과(Δ_evid)는 모든 모델에서 양수이며, 작거나 한국어가 약한 모델일수록 회복 폭이 크다. EXAONE-32B의 회복 폭이 작은 것은 원래 정확도가 높아 여지가 좁기 때문이다.

표 2. 5개 모델의 6개 기본 조건별 macro-F1

모델	A	B	C	D	E	F
EXAONE-1.2B	0.602	0.409	0.707	0.635	0.643	0.499
EXAONE-32B	0.783	0.759	0.761	0.727	0.767	0.725
Qwen3-14B	0.734	0.725	0.710	0.656	0.668	0.683
Qwen3-32B	0.797	0.755	0.732	0.726	0.716	0.697
Llama-8B	0.771	0.719	0.742	0.615	0.709	0.641

교차언어 효과(Δ_cl)는 5개 모델 모두에서 미미하거나 나타나지 않았다. 다국어 수학 추론에서 보고된 교차언어 효과가 사실 검증 도메인에서는 재현되지 않았고, 소형 모델에는 오히려 해로웠다. 표 2에서 보듯 한국어 직접 조건(D)에 근거를 더하면(C) 5개 모델 모두 정확도가 오르지만, 교차언어를 더해도(E) 더 오르지 않는다.

3.2 동일 규모 모델 간 비교

동일 규모 비교 세 가지를 보면 같은 32B·4bit인 EXAONE-32B와 Qwen3-32B의 한국어 조건 평균 macro-F1 차이는 +0.027이고 신뢰구간이 0을 포함하지 않는다. 같은 비교의 영어 조건 차이는 사실상 동률(−0.013~+0.004)이었다. 한국어 특화 효과가 한국어 입력에 한정해 나타난다.

EXAONE 계열(1.2B→32B)의 한국어 조건 차이는 +0.124, Qwen3 계열(14B→32B)은 +0.038로 둘 다 신뢰구간이 0을 포함하지 않았다. 한국어 특화와 모델 크기는 독립적으로 정확도에 기여하나 특화가 크기를 완전히 대체하지는 않는다. EXAONE-1.2B의 한국어 macro-F1은 0.621로 범용 8B·14B 모델보다 낮다. 그림 1에서 32B 영역의 EXAONE이 Qwen3보다 한국어 마커가 위에 있어 이 차이가 시각적으로 드러난다.

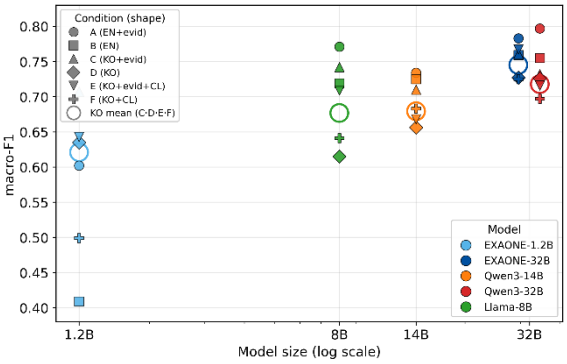


그림 1. 모델 크기에 따른 조건별 macro-F1 분포

3.3 불확실성과 환각 분석

응답 유보를 허용하면 한국어 특화 모델은 모르는

※이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2024-00460011, 인류세의 기후테크 역량 강화를 위한 기후·환경 데이터 구축·처리 플랫폼)

문항에 '모르겠다'를 더 자주 택해 자신의 한계를 드러낸다. 표 3은 이를 대표적으로 보여주는 세 사례다. 동급 32B에서 EXAONE-32B는 유보율 0.462로 어려운 문항의 절반 가까이를 보류했고, 그 결과 환각률이 0.082로 가장 낮았다. 반면 Qwen3-32B는 거의 모든 문항에 답해(유보율 0.009) 환각률이 0.225로 높았다. 즉 정확도 우위에 더해, 한국어 특화 모델은 모르는 것을 모른다고 인식하는 신뢰도 보정(calibration)에서도 앞선다. 다만 이는 신중한 판단으로도, 단순 회피로도 해석될 수 있다. 소형 모델인 EXAONE-1.2B는 교차언어 조건에서 유보율이 0.699까지 치솟아, 응답 유보가 오히려 응답률을 무너뜨리기도 한다. 환각이 집중되는 지점을 보면, 영어로는 정답이나 한국어로는 오답이 된 사례 50건에서 전문 용어와 숫자 인용에 오류가 몰려 한국어 입력의 약점이 드러난다.

표 3. 모델별 응답 유보와 환각 양상 (대표 3사례)

모델	조건	응답률	유보율	환각률
EXAONE-32B	한국어·근거 없음	0.538	0.462	0.082
Qwen3-32B	한국어·근거 없음	0.991	0.009	0.225
EXAONE-1.2B	한국어+교차언어	0.289	0.699	0.117

IV. 결론

본 연구는 한국어 기후 사실 검증에서 LLM의 영-한 정확도 격차를 측정하고, 알려진 네 가지 방법(한국어 특화 모델 선택, 영어 근거 제공, 교차언어 추론, 응답 유보)의 효과를 5개 오픈 모델과 12개 조건 격자에서 비교했다. 같은 32B·4bit 조건에서 EXAONE-32B는 Qwen3-32B 대비 한국어 macro-F1에서 +0.027의 유의한 우위를 보였고 같은 비교의 영어 조건에서는 동률이었다. 한국어 특화 효과는 모델 크기로 환원되지 않고 한국어 입력에 한정해 발현된다.

영어 근거 제공은 모든 모델에서 정확도를 끌어올렸고 약한 모델일수록 회복 폭이 컸다. 반면 교차언어 추론은 어떤 크기에서도 효과가 없어, 다국어 수학 추론에서 보고된 교차언어 효과가 사실 검증 과제로는 이어지지 않았다. 응답 유보에서는 한국어 특화 모델이 '모를 때 모른다'를 더 자주 드러냈으나 응답률은 떨어졌다.

다만 평가는 SUPPORTS/REFUTES 2-class에 한정되며, EXAONE 계열 내부 비교는 양자화 정밀도가 섞여 순수 크기 효과를 분리하지 못한다. 후속 연구에서는 NEI를 포함한 3-class 확장, 대기·해양·생태 등 환경 도메인으로의 일반화, 학습 기반 방법, 한국어 특화 모델의 보정(calibration) 기제 분석 등을 다루고자 한다.

참고문헌

[1] H. Zhang, S. Anjum, H. Fan, W. Zheng, Y. Huang, and Y. Feng, "Poly-FEVER: A multilingual fact verification benchmark for hallucination detection in large language models," arXiv preprint arXiv:2503.16541, 2025.

[2] D. Kang, S. Hwang, D. Kim, H. Kim, and G. G. Lee, "Why do multilingual reasoning gaps emerge in reasoning language models?" arXiv preprint arXiv:2510.27269, 2025.

[3] T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, and M. Leippold, "Climate-FEVER: A dataset for verification of real-world climate claims," in NeurIPS Workshop on Tackling Climate Change with Machine Learning, 2020.

[4] LG AI Research, "EXAONE 4.0: Unified large language models integrating non-reasoning and reasoning modes," arXiv preprint arXiv:2507.11407, 2025.

※이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2024-00460011, 인류세의 기후테크 역량 강화를 위한 기후·환경 데이터 구축·처리 플랫폼)

공공 지역데이터의 교육 자원 적합성 평가 프레임워크 개발

김하영[O], 조정원[*]
[O]제주대학교 지능소프트웨어교육연구소 연구원
[*]제주대학교 컴퓨터교육과 교수, 지능소프트웨어교육연구소 소장
e-mail: hayoung@stu.jejunu.ac.kr[O], jwcho@jejunu.ac.kr[*]

Development of a Public Local Educational Data Framework for Suitability Evaluation

HaYoung Kim[O], Jungwon Cho[*]
[O]Researcher, Intelligent Software Education Research Institute, Jeju National University
[*] Professor, Dept. of Computer Education & Director, Intelligent Software Education Research Institute, Jeju National University

요 약

본 연구는 AI 융합 교육 프로그램 개발의 선행 단계로서, 공공 지역데이터가 교육 자원으로 기능하기 위한 적합성 조건을 이론적으로 규명하고 평가 프레임워크를 개발하는 데 목적이 있다. 기존 연구들이 데이터 기반 학습의 효과를 검증하는 데 집중한 반면, 어떤 데이터가 어떠한 조건을 갖출 때 교육 자원으로 적합한가에 대한 사전 진단 논의는 부족하였다. 본 연구는 이를 데이터 맥락성(Data Contextuality)으로 개념화하고, 접근 가능성, 교육 맥락 정합성, 인지적 수준 적합성으로 구성된 평가 프레임워크(PLED-F)를 선행연구 분석을 통해 도출하였다. 본 프레임워크는 향후 지역 공공데이터를 활용한 중등 AI 융합 교육 프로그램 개발 시 데이터 선정 단계에서 체계적 근거로 활용될 수 있으며, 타 교과의 지역 공공데이터 교육 자원화 연구로 확장 가능하다.

키워드: 데이터 맥락성, 공공 지역데이터, AI 융합 교육, 교육 자원 적합성, 장소기반교육

I. 서론

2022 개정 교육과정에서 데이터 소양이 핵심 역량으로 강조되며 AI 융합 교육 내 데이터의 질과 맥락에 대한 교육공학적 접근이 중요 해졌다[1,2]. 그러나 현행 교육은 글로벌 범용 데이터셋이나 교과서 예시에 치우쳐 있어[3], 학습자가 실생활과 연계해 주도적으로 문제를 발견하는 데 한계가 있다[4]. 중등 학습자의 동기는 주제의 개인적 관련성에 큰 영향을 받으므로 생활권 중심 데이터의 활용이 요

※ 이 논문 또는 저서는 2021년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2021S1A5C2A04088646)

구되지만[5,6], 관련 장소기반교육 연구는 사후 효과 측정에 편중되어 있어 교사가 객관적 기준 없이 직관에 의존해 수업 데이터를 선정하는 실천적 공백이 존재한다[7-9]. 이에 본 연구는 공공 지역데이터의 교육 자원화 조건을 탐색하고, 학습자 경험 및 교육 목표와의 상호 연관성을 의미하는 데이터 맥락성(Data Contextuality)으로 개념화하여 접근 가능성, 교육 맥락 정합성, 인지적 수준 적합성으로 구성된 공공 지역데이터 교육 자원 적합성 평가 프레임워크를 개발하는 데 목적이 있다.

II. 관련 연구

데이터 기반 학습에서 데이터의 맥락성은 학습의 질을 결정하는 핵심 변인이다[10]. 선행 교육공학 연구에 따르면 학습자가 데이터의 출처와 생성된 실제 배경 맥락을 연계하여 파악할 때 고차원적 인지 활동이 적극적으로 활성화되며 데이터 비판적 사고력이 유의미하게 향상된다[11,12]. 본 연구는 이처럼 학습자 개인의 삶과 교과 목표가 융합되는 영역을 '데이터 맥락성(Data Contextuality)'으로 개념화한다. 데이터 맥락성의 구체화를 위해 다음의 선행연구 이론을 바탕으로 구성요소를 도출하였다. 첫째, 장소기반교육 관점에서 생활권 생태·환경 데이터 활용은 장소 귀속감과 실천적 행동 의지를 강화하는 매개체가 된다[7,13,15]. 둘째, 자기결정이론에 따라 학습자가 문제를 스스로 정의·분석하고 지역에 기여하는 과정은 내재적 학습 동기를 자극하는 요인이다[16,17]. 셋째, 컴퓨팅 사고력 관점에서 교육 데이터는 학습자 발달 단계에 알맞은 인지적 수준 적합성을 유지해야 한다[18-20]. 즉, 데이터 맥락성은 고정된 속성이 아닌 학습자-데이터-목표 간의 상호작용적 관계이다.

III. 제안방법

본 연구는 선행연구 분석을 중심으로 한 이론적 탐색 연구이다. 탐색을 통해 도출한 데이터 맥락성을 구체화하여 공공 지역데이터 교육 자원 적합성 평가 프레임워크(PLED-F)를 제안한다. 도출된 내용은 다음 표1, 표2 와 같다.

표1. 이론적 계열별 선행연구 검토 및 데이터 맥락성 도출

이론적 계열	공통 강조 조건 및 평가 기준 도출 근거	도출된 구성요소
맥락기반학습 데이터 소양	데이터 출처·생성 맥락 이해 시 인지 처리 및 비판적 사고력 향상	교육 맥락 정합성 [10],[12]
장소기반교육 (PBE)	생활권 생태 맥락 학습 자원화 자기효능감·귀속감 향상 실증	교육 맥락 정합성 [7],[13],[14],[15]
자기결정이론 (SDT)	자율성·유능감·관계성 충족 시 내재적 동기 강화	교육 맥락 정합성 [16],[17]
공공데이터 인프라·시각화	개방형 API/CSV 구조, 갱신 주기, 결측치 20% 이하 기준 설정 근거	접근 가능성 [9],[11]
컴퓨팅사고력 (CT)·	핵심 변수 10개 이하, 중등 통계 수준,	인지적 수준 적합성 [18],[19],[20]

표2. 데이터 맥락성의 구성 요소 및 내용

구성	개념적 정의 및 도출 근거	공공 지역데이터의 적용 기준
접근 가능성	비용이나 고도의 기술적 장벽 없이 교사와 학습자가 데이터에 접근할 수 있는 정도 [9,11]	- 개방형 API 및 CSV 형식 제공 여부 - 교육 현장 활용에 적합한 데이터 최신성 및 안정적 인프라 확보
교육 맥락 정합성	데이터의 속성(주제·지역·시간 범위)이 교과 성취기준 및 학습자의 실생활권과 일치하는 정도 [1,7,15,16]	- 학습자 실제 거주지 인근 관측 데이터 확보 - 교육과정 성취기준 연계성 유무 - 학습자가 직접 체감한 지역 사회 현상 포함
인지적 수준 적합성	데이터의 구조적 복잡성과 정형성이 학습자의 발달 단계 및 AI 분석 도구 수준에 부합하는 정도 [18-20]	- 학습자가 통제 가능한 변수 규모 - 중등 수학 · 과학 수준의 통계 정합성

※ 이 논문 또는 저서는 2021년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2021S1A5C2A04088646)

PLED-F는 순차 점검 방식으로 진행된다. 접근 가능성에서는 기준 미충족 시 대체 데이터를 탐색하며, 이를 통과하면 교육 맥락 정합성 검토를 통해 교육과정 성취기준과의 연계를 분석한다. 인지적 수준 적합성 검증에서는 CT 4요소 루브릭을 평가 척도로 적용하여 학습자의 데이터 처리 가능 수준을 종합적으로 평가한다.

IV. 결론

본 연구는 현장 교육에서 직관에 의존하던 데이터 선정 한계를 개선하고자 선행연구 분석을 바탕으로 데이터 맥락성을 개념화하고, 공공 지역데이터 교육 자원 적합성 진단 프레임워크(PLED-F)를 제안하였다. 이는 객관적인 자원 선정 기준을 제공하여 지역 공공데이터를 타 교과 AI 융합 교육 학습 자원화하기 위한 방법론적 범용성을 제공한다는 점에서 의의가 있다. 하지만 초기 탐색적 문헌 검토 수준으로서 체계적 문헌 분석 절차와 전문가 델파이 검증을 거치지 못한 한계가 있다. 후속 연구에서는 델파이 분석을 통해 프레임워크의 구조적 타당성을 보완하고, PLED-F를 거쳐 선정된 데이터를 투입한 중등 AI 융합 교육 프로그램을 개발·적용하여 학습자의 내재적 동기, AI 자기효능감 및 역량에 미치는 효과를 실증적으로 검증할 계획이다.

참고문헌

[1] 교육부. (2022). 2022 개정 교육과정 (교육부 고시 제2022-33호). 교육부.

[2] Carmi, N., Arnon, S., & Orion, N. (2015). Transforming environmental knowledge into behavior: The mediating role of environmental emotions. The Journal of Environmental Education, 46(3), 183-201. https://doi.org/10.1080/00958964.2015.1028393

[3] Kong, S. C., Cheung, W. M. Y., & Zhang, G. (2021). Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. Computers and Education: Artificial Intelligence, 2, 100026. https://doi.org/10.1016/j.caeai.2021.100026

[4] Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. Journal of Learning Analytics, 2(2), 5-13. https://doi.org/10.18608/jla.2015.22.2

[5] Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. Contemporary Educational Psychology, 61, 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

[6] Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. Educational Psychologist, 41(2), 111-127. https://doi.org/10.1207/s15326985ep4102_4

[7] Sobel, D. (2004). Place-based education: Connecting classrooms and communities. Education for Meaning and Social Justice, 17(3), 63-64.

[8] Smith, G. A. (2002). Place-based education: Learning to be where we are. Phi Delta Kappan, 83(8), 584-594. https://doi.org/10.1177/003172170208300806

[9] 기상청. (2024). 기상청 공공데이터포털 활용 가이드. 기상청. https://data.kma.go.kr

[10] Schank, R. C. (1999). Dynamic memory revisited. Cambridge University Press. https://doi.org/10.1017/CBO9780511527920

[11] Börner, K., Bueckle, A., & Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. Proceedings of the National Academy of Sciences, 116(6), 1857-1864.

※ 이 논문 또는 저서는 2021년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2021S1A5C2A04088646)

https://doi.org/10.1073/pnas.1807180116
[12] Wolff, A., Gooch, D., Cavero Montaner, J. J., Rashid, U., & Kortuem, G. (2016). Creating an understanding of data literacy for a data-driven society. The Journal of Community Informatics, 12(3). https://doi.org/10.15353/joci.v12i3.3275
[13] Sobel, D. (2004). Place-based education: Connecting classrooms and communities. Orion Society.
[14] Hamilton, E., & Marckini-Polk, L. (2023). The impact of place-based education on middle school students' environmental literacy and stewardship. Cogent Education, 10(1), 2163789. https://doi.org/10.1080/2331186X.2022.2163789
[15] Soucy, A. R., Rahimzadeh-Bajgiran, P., De Urioste-Stone, S., Eitel, K., Jansujwicz, J., & Brownlee, M. (2025). Impact of a place-based educational approach on student and community members' experiences and learning within a post-secondary GIS course. Journal of Geoscience Education, 73(4), 386-401. https://doi.org/10.1080/10899995.2024.2379194
[16] Deci, E. L., & Ryan, R. M. (2000). The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior. Psychological Inquiry, 11(4), 227-268. https://doi.org/10.1207/S15327965PLI1104_01
[17] Wang, Y., Wang, H., Wang, S., Wind, S. A., & Gill, C. (2024). A systematic review and meta-analysis of self-determination-theory-based interventions in the education context. Learning and Motivation, 87, 102015. https://doi.org/10.1016/j.lmot.2024.102015
[18] Wing, J. M. (2006). Computational thinking. Communications of the ACM, 49(3), 33-35. https://doi.org/10.1145/1118178.1118215
[19] Hsu, H. P. (2025). Programming to Prompting: Developing Computational Thinking through Large Language Model-Based Generative Artificial Intelligence. TechTrends, 69, 485-506. https://doi.org/10.1007/s11528-025-01052-6
[20] Xu, J., Chen, Z., Chen, M., Li, Y., & Xu, X. (2026). Unpacking the Effects of GenAI on Cultivating Students' Computational Thinking: A Meta-Analysis. Journal of Educational Computing Research, 64(4), 1024-1067. https://doi.org/10.1177/07356331261419586

※ 이 논문 또는 저서는 2021년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2021S1A5C2A04088646)

생성형 AI를 활용한 국가 R&D 성과물 이관 준비도 자동 평가 방법론: 개념적 프레임워크

정문권[O], 이주영[*]
[O*]한국과학기술정책플랫폼협동조합 스마트미래융합본부
e-mail: mkjeong@kspp.re.kr[O], jylee@kspp.re.kr[*]

A Generative-AI Approach to Automated Assessment of Transfer Readiness for National R&D Deliverables: A Conceptual Framework

Jung Moonkwon[O], Lee Jooyoung[*]
[O*]Smart Future Convergence Division, Korea Science Policy Platform co-op

요 약

국가 R&D를 통해 구축된 데이터 플랫폼·소프트웨어·시제품 등의 성과물은 사업 종료 이후 운영 주체의 부재와 제도적 연계의 미흡으로 인해 충분히 활용되지 못하는 경우가 빈번히 발생한다. 이러한 문제를 완화하기 위해서는 성과물의 이관 이전 단계에서 이관 준비도를 평가할 필요가 있다. 하지만 기존의 전문가 설문·델파이 방식은 비용·시간·표본 규모의 제약, 평가자 주관성, 재현 가능성의 한계를 지닌다. 본 연구는 이러한 문제를 대규모 언어모델을 활용한 자동 텍스트 평가 방법론을 통해 해결하는 개념적 프레임워크를 제안한다. 프레임워크는 다음과 같은 절차로 구성된다: (1) 이관 준비도를 다차원 평가 기준(운영 주체 명확성, 제도·법적 연계성, 기술 성숙·유지보수성, 사용자 효용, 확산·지속가능성)으로 구조화한다, (2) 언어모델이 R&D 성과 문서를 해당 기준에 따라 점수화하며 판단 근거를 함께 제시한다, (3) 동일 문서에 대한 전문가 평가와의 일치도를 일치율 및 카파 계수로 정량 검증한다, (4) 위치·장황성·자기선호 등 언어모델 평가자의 편향과 모델·프롬프트 변형에 따른 안정성을 점검한다. 본 연구는 생성형 AI를 활용하여 평가의 자동화·정량화·재현가능성을 제고하고, 그 신뢰성 검증 절차를 설계한다는 점에서 의의가 있다.

키워드: 생성형 AI, 대규모 언어모델, 자동 텍스트 평가, 국가 R&D 성과 이관, 책임있는 AI

This research was supported by the Korea Institute of Marine Science & Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries, Republic of Korea, in 2026 (RS-2022-KS221571, Development and Demonstration of an AI-based Fishing Vessel Safety Design Data Platform).

I. 서론

국가 R&D를 통해 산출된 데이터·소프트웨어·알고리즘·시제품 등은 기술적 가치를 지니지만, 사업 종료와 함께 전담 인력·예산 지원이 종료되고 관련 제도와의 연계가 단절되면서 실제 활용으로 이어지지 못하는 문제가 반복된다. 정부 역시 제5차 연구성과 관리·활용 기본계획에서 연구성과 활용·확산의 획기적 제고를 핵심 목표로 제시하여, 기존 성과 활용이 미흡하다는 문제의식을 공식화한 바 있다[9].

성과물을 안정적인 공공 인프라로 이관·정착시키기 위해서는, 이관에 앞서 해당 성과물의 이관·운영 준비 수준(이관 준비도)을 사전 진단하는 절차가 필요하다. 그러나 현행 평가는 소수 전문가의 설문·델파이·정성 심사에 의존하여 비용과 시간이 많이 소요되고, 동원 가능한 전문가 표본이 제한되어 다수 성과물에 일관되게 적용하기 어려우며, 평가자 주관에 따라 결과가 상이해지고 동일 조건에서 재현하기 어렵다는 한계를 지닌다.

본 연구는 이러한 평가 문제를 R&D 성과 문서의 텍스트 데이터를 정량 점수로 변환하여, 자연어처리 분야에서 정착되고 있는 생성형 AI(대규모 언어모델) 기반 자동 텍스트 평가 기법을 이관 준비도 평가에 적용하는 개념적 프레임워크와 그 신뢰성 검증 설계를 제안한다.

II. 관련 연구

2.1 R&D 성과의 활용과 사장 문제

R&D 성과의 저조한 활용·확산은 국가연구개발사업 성과분석 등을 통해 지속적으로 점검되어 온 정책 의제이며[10], 성과물의 운영주체 이관·제도화·사업화 연계가 핵심 과제로 다루어져 왔다. 다만 이러한 논의는 대체로 정성·사례 중심이어서, 다수 성과물을 일관된 기준으로 정량 비교·진단하는 도구는 충분히 정립되지 않았다. 실제로 공공연구기관의 기술이전율은 2022년 31.7%에 그쳐 신규 확보 기술의 상당수가 당해연도 이전으로 연결되지 못하였으며, 기술이전 건수 또한 전년 대비 감소하였다[11]. 나아가 기술활용률은 2019년 3.8%, 보유기술 운용률은 9.3% 수준에 머물러 확보된 기술의 대부분이 실제 활용·운용으로 이어지지 못하고 있으며, 건당 기술료 징수액도 2012년 5,170만 원에서 2019년 2,910만 원으로 하락하였다[10]. 이러한 지표는 성과물의 산출 자체보다 이관·활용 단계에서 구조적 병목이 발생하고 있음을 보여준다.

2.2 준비도·성숙도 평가 프레임워크

성과물의 준비 상태를 구조화된 항목으로 평가하려는 시도로는, 기술 성숙도를 아홉 단계로 규정한 기술성숙도(TRL)[6], 소프트웨어 제품 품질을 기능적 합성·신뢰성·유지관리성·보안성 등 다차원으로 규정한 ISO/IEC 25010[7], 그리고 최신성·정확성·상호연계성을 중심으로 한 공공데이터 품질관리 기준[8] 등이 있다. 본 연구의 평가 기준은 이러한 다차원 평가 체계의 항목 설계 방식을 차용하되, 평가 수행 주체를 사람에서 언어모델 기반 자동 평가로 전환한다는 점에서 차별화된다.

2.3 언어모델 기반 자동 텍스트 평가

언어모델을 평가자로 활용하는 자동 텍스트 평가는 평가 자동화 분야의 주요 방법으로 정착되고 있다. Zheng 등은 언어모델 평가자가 사람 평가자와 80% 이상(동점 제외 시 약 85%) 일치한다는 결과를 보고하는 한편, 위치 편향·장황성 편향·자기선호 편향을 규명하였다[1]. 또한 Gilardi 등은 언어모델의

This research was supported by the Korea Institute of Marine Science & Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries, Republic of Korea, in 2026 (RS-2022-KS221571, Development and Demonstration of an AI-based Fishing Vessel Safety Design Data Platform).

분류 정확도가 크라우드 작업자보다 평균 약 25% 포인트 높고 평가자 간 일치도 역시 높다고 보고하였다[2]. 평가 기준을 자연어로 제시하고 단계적 추론을 통해 점수화하는 방법(G-Eval)은 요약 평가에서 사람 판단과의 스피어만 상관계수 0.514를 달성하였다[3]. 이뿐만 아니라, 언어모델 평가자의 자기 선호 편향을 정량적으로 측정하려는 후속 연구도 제시되었다[4]. 사람과 언어모델의 일치도는 일반적으로 일치율과, 우연 일치를 보정한 코헨의 카파로 정량화한다.

코헨의 카파(κ)는 두 평가자가 동일 대상을 동일 범주로 분류할 때, 우연에 의한 일치를 제거하고 남은 실질적 일치 정도를 −1에서 1 사이로 나타내는 지표이다. κ는 실제 일치율(Po)에서 우연 기대 일치율(Pe)을 보정하여 κ=(Po−Pe)/(1−Pe)로 계산되며, 단순 일치율이 우연만으로 부풀려지는 문제를 완화한다. 본 연구는 자동 평가와 전문가 평가의 일치도를 카파로 정량화하고, 통상적 해석 기준(표 1)에 따라 신뢰 수준을 판정한다. 한편 정부 정책문서·연구계획서 등 행정 텍스트를 언어모델로 자동 평가하려는 시도는 연구계획서 심사나 정책 준수 점검 등 일부 영역에서 초기 단계로 보고되고 있으나, 표준 일치도 지표 보고가 미비하거나 전문가 심사를 보조하는 수준에 머물러 있다. 특히 R&D 성과물의 이관 준비도를 언어모델로 점수화하고 전문가 평가와의 일치도를 정량 검증한 연구는 아직 확인되지 않는다. 본 연구는 이 공백에 대응하는 개념적 프레임워크와 신뢰성 검증 설계를 제안한다는 점에서 차별성을 가진다.

표1. 카파 해석 기준

κ 범위	일치 강도
< 0	일치 없음(Poor)
0.00–0.20	미미(Slight)
0.21–0.40	약함(Fair)
0.41–0.60	보통(Moderate)
0.61–0.80	상당(Substantial)
0.81–1.00	거의 완전(Almost perfect)

III. 제안방법

본 방법의 출발점은, 이관 준비도를 판단하는 근거가 결국 성과계획서·최종보고서·매뉴얼 등 성과 문서의 텍스트에 내재한다는 점이다. 인간 전문가 역시 이 문서를 읽고 운영 주체·제도 연계·유지보수 체계의 서술 여부를 근거로 점수를 부여한다. 따라서 평가는 본질적으로 "문서를 읽고 기준별 점수로 변환하는" 텍스트 분류·점수화 과제이며, 이는 대규모 언어모델이 강점을 보이는 작업 유형과 정확히 일치한다. 본 연구가 평가 문제를 데이터사이언스 문제로 재정의하는 근거가 여기에 있다.

3.1 이관 준비도 평가 기준 정의

이관 준비도를 상호 독립적인 다차원 기준으로 구조화하고, 각 기준에 1~5점 척도와 점수별 판정 기준 문장을 명시한다. 각 점수에 판정 기준 문장(앵커)을 명시하는 이유는 평가의 재현가능성과 해석가능성을 확보하기 위함이다. 기준이 모호하면 같은 문서라도 평가 시점·평가자에 따라 점수가 달라진다. TRL[6]·ISO/IEC 25010[7]이 단계별 정의를 명문화하여 신뢰도를 확보한 것과 동일한 설계 원리이며, 자연어 기준을 제시해 점수화하는 G-Eval[3]의 접근과도 일치한다.

표2. 이관 준비도 평가 차원과 핵심 질문

This research was supported by the Korea Institute of Marine Science & Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries, Republic of Korea, in 2026 (RS-2022-KS221571, Development and Demonstration of an AI-based Fishing Vessel Safety Design Data Platform).

평가 차원	핵심 질문
운영주체 명확성	이관 이후 운영·책임 주체가 구체적으로 명시되어 있는가
제도·법적 연계성	관련 법령·지침·표준과의 연계 근거가 제시되는가
기술 성숙·유지보수성	표준 준수 및 지속적 유지보수 체계가 마련되어 있는가
사용자 효용	실제 수요자의 활용 가치·수요 근거가 제시되는가
확산·지속가능성	확산 경로와 지속 운영(예산·거버넌스) 계획이 있는가

3.2 언어모델 자동 채점

R&D 성과 문서를 입력하면 언어모델이 각 기준에 따라 채점하고 판단 근거 문장을 함께 제시한다. 동일 문서를 여러 번 반복 채점하여 점수의 평균과 분산을 산출한다.

언어모델의 출력은 확률적이어서 동일 입력에도 점수가 변동할 수 있다. 본 방법이 동일 문서를 반복 채점하여 평균과 분산을 산출하는 것은, 점수의 중심값뿐 아니라 불확실성까지 측정하기 위함이다. 분산이 큰 항목은 모델이 확신하지 못하는 영역으로, 기준 문장의 모호성 또는 문서의 정보 부족을 시사하는 진단 신호로 활용된다. 이는 단일 점수만 제시하는 기존 정성 평가가 제공하지 못하는 정보다.

3.3 사람·언어모델 일치도 검증

전문가가 동일한 기준으로 독립 채점한 결과를 기준으로 일치율, 점수 상관(스피어만·피어슨), 코헨의 카파(평가자가 3인 이상인 경우 크리펜도르프 알파)를 산출한다. 전문가가 2인 이상인 경우 전문가 간 일치도도 함께 보고하여 사람 평가 기준선과 비교한다.「」

자동 평가의 타당성은 "사람 기준선에 얼마나 근접하는가"로 입증된다. 핵심은 전문가 간 일치도(인간-인간)를 상한 기준선으로 설정하고, 여기에 자동 평가-전문가 일치도(LLM-인간)를 비교하는 것이다. 만약 LLM-인간 일치도가 인간-인간 일치도에 근접한다면, 자동 평가는 "사람만큼 일관적"이라고 주장할 수 있다. 실제로 Zheng 등[1]은 LLM-인간 일치도(약 85%)가 인간-인간 일치도와 대등함을 보여 이 논증 구조의 타당성을 입증하였다. 본 연구는 이 비교 설계를 이관 준비도 평가에 반영한다.

3.4 편향·신뢰성 점검

위치·장황성·자기선호 편향[1][4]을 점검하고, 입력 순서·프롬프트 표현·사용 모델 변형에 대한 점수 안정성과 근거 문장의 해석 가능성을 평가한다. 편향 점검을 절차에 포함하는 이유는, 언어모델 평가자가 내용이 아니라 형식에 휘둘릴 위험 때문이다. 위치 편향(제시 순서), 장황성 편향(긴 문서 선호), 자기선호 편향(자사 계열 모델·유사 시스템 선호)[1][4]은 평가의 공정성을 훼손한다. 또한 프롬프트 표현이나 사용 모델을 바꿨을 때 점수가 크게 흔들린다면 그 평가는 신뢰할 수 없다. 본 절차는 이러한 강건성·공정성 점검을 명시적으로 포함함으로써 "책임 있는 AI" 원칙을 평가 설계에 반영한다.

IV. 개념 적용 예시

가상의 R&D 성과 문서 1건을 위 기준으로 평가하는 상황을 가정한다. 언어모델은 운영 주체 명확성에서 낮은 점수와 함께 운영 이관 주체에 대한 서술 부재를, 기술 성숙·유지보수성에서 높은 점수와 함께 표준 준수·유지보수 계획이 구체적으로 기술되

This research was supported by the Korea Institute of Marine Science & Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries, Republic of Korea, in 2026 (RS-2022-KS221571, Development and Demonstration of an AI-based Fishing Vessel Safety Design Data Platform).

어 있음을 근거로 제시한다. 차원별 점수는 방사형 차트로 시각화되어 성과물의 강점과 보완 영역을 직관적으로 제시한다. 본 예시는 실제 데이터 수집 이전의 개념 실증 수준에 해당한다.

V. 한계 및 향후 계획

언어모델 자동 평가의 신뢰도는 평가 대상의 도메인 특성과 루브릭 설계 방식에 따라 크게 달라진다. 한 산업 분야 문서 관련성 주석 실증에서 언어모델의 코헨의 카파는 0.136~0.401에 그쳐 지도학습 분류기(0.669)에 미치지 못하였고[5], 영양·임상심리 등 고도의 전문지식이 요구되는 과제에서도 전문가–언어모델 일치율이 64~68% 수준으로 전문가 간 일치율(약 72~75%)을 밑돌았다[12]. 의료 안전·독성 판단처럼 모델의 안전장치가 평가 자체를 방해하는 과제에서는 카파가 0에 가깝거나 음수로 나타나기도 하였다[13]. 반면 평가 루브릭이 명확하고 점수별 앵커와 참조 예시가 제공된 경우에는, 한 전문 영역 평가에서 전문가–언어모델 가중 카파 중앙값이 0.60으로 전문가 간 값(0.58)과 대등한 수준에 이르렀다[14]. 이는 본 연구가 3.1절에서 점수별 앵커 문장과 명시적 기준을 강조한 설계가 신뢰도 확보의 핵심 조건임을 뒷받침한다. 따라서 본 방법은 전문가 평가를 대체하기보다 사람 평가 기준선 대비 검증을 전제로 한 사전 진단 도구로 활용되어야 한다.

표3. 도메인별 언어모델-전문가 일치도 사례

도메인(사례)	일치도 및 시사점
챗봇·범용 평가	일치율 약 85%로 사람 수준 도달[1]
구조화된 전문 평가(앵커·예시 제공)ㅅ	가중 카파 약 0.60, 전문가 간(0.58)과 대등[14]
산업 문서 관련성 주석	카파 0.136~0.401, 분류기(0.669) 미달[5]
영양·임상심리 등 고난도 전문 판단	일치율 64~68%로 전문가 간 수준 미달[12]
의료 안전·독성 판단	카파 0에 근접 또는 음수[13]

VI. 결론

본 연구는 R&D 성과물의 사장 방지를 위한 의사결정 지원에, 생성형 AI 기반 자동 텍스트 평가 방식을 도입하는 개념적 프레임워크와 신뢰성 검증 설계를 제안하였다. 본 연구의 제안 방법은 기존 전문가 평가의 비용·주관성·재현 곤란 문제를 보완하고, 다수 성과물에 일관된 정량 기준을 적용할 수 있다는 점에서 의의가 있다.

참고문헌

[1] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, I. Stoica, J. E. Gonzalez, and H. Zhang, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in Proc. NeurIPS, Datasets and Benchmarks Track, 2023. doi: 10.48550/arXiv.2306.05685.

[2] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," Proc. Natl. Acad. Sci. U.S.A., vol. 120, no. 30, p. e2305016120, 2023. doi: 10.1073/pnas.2305016120.

[3] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu,

This research was supported by the Korea Institute of Marine Science & Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries, Republic of Korea, in 2026 (RS-2022-KS221571, Development and Demonstration of an AI-based Fishing Vessel Safety Design Data Platform).

"G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment," in Proc. EMNLP 2023, Singapore, pp. 2511-2522. doi: 10.18653/v1/2023.emnlp-main.153.

[4] K. Wataoka, T. Takahashi, et al., "Self-Preference Bias in LLM-as-a-Judge," 2024.

[5] J. V. M. Correia, M. M. Bell, J. V. R. Amorim, J. Queiroz, D. Pedronette, I. R. Guilherme, and F. L. de Oliveira, "Analysis of Automated Document Relevance Annotation for Information Retrieval in Oil and Gas Industry," in Proc. 2025 Conf. Empirical Methods in Natural Language Processing (EMNLP): Industry Track, Suzhou, China, 2025, pp. 1878-1889. doi: 10.18653/v1/2025.emnlp-industry.132.

[6] J. C. Mankins, "Technology Readiness Levels: A White Paper," NASA, 1995.

[7] ISO/IEC 25010:2023, Systems and software engineering — SQuaRE — Product quality model, 2023.

[8] 행정안전부·한국지능정보사회진흥원, 공공데이터 품질관리 매뉴얼.

[9] 과학기술정보통신부, 제5차 연구성과 관리·활용 기본계획(2026-2030), 2025.

[10] 한국과학기술기획평가원, 국가연구개발사업 성과분석.

[11] 과학기술정보통신부, 「2024년 공공 기술이전·사업화 현황조사」(정부승인통계 제115022호), 2024. [Online]. Available: https://www.index.go.kr/unity/potal/main/EachDtlPageDetail.do?idx_cd=1148

[12] M. Szymanski et al., "Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks," in Proc. ACM Conf. Intelligent User Interfaces (IUI), 2025. doi: 10.1145/3708359.3712091.

[13] A. Bavaresco et al., "LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks," in Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL): Short Papers, 2025. doi: 10.18653/v1/2025.acl-short.20.

[14] S. Kumar, M. Groh et al., "When Large Language Models are Reliable for Judging Empathic Communication," Nature Machine Intelligence, 2025. doi: 10.48550/arXiv.2506.10150.

This research was supported by the Korea Institute of Marine Science & Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries, Republic of Korea, in 2026 (RS-2022-KS221571, Development and Demonstration of an AI-based Fishing Vessel Safety Design Data Platform).

에이전트 기반 조직 행동 시뮬레이션을 통한 다원적 HR 정책 사전 검사 아키텍처

김성현[O(발표자)], 김장현[*(교신저자)]
[O]성균관대학교 인터랙션사이언스학과
[*]성균관대학교 인터랙션사이언스학과/인간AI인터랙션융합전공
e-mail: seonghyune@g.skku.edu[O], alohakim@skku.edu[*]

An Architectural Framework for Pluralistic HR Policy Screening via Agent-Based Organizational Simulations

Seong Hyun Kim[O], Jang Hyun Kim[*]
[O]Dept. of Interaction Science, Sungkyunkwan University
[*]Department of Interaction Science/Human-Artificial Intelligence Interaction, Sungkyunkwan University

요 약

HR 정책은 단일 성과 지표로 환원되지 않는다. 하나의 정책이 생산성이나 협업을 개선하면서도 동시에 직원의 웰빙, 유지, 공정성 인식에는 비용을 발생시킬 수 있기 때문이다. 본 연구는 정책 배포 이전 단계의 스크리닝을 위한 게임마스터(GM) 중재 멀티에이전트 시뮬레이터를 제안한다. 시뮬레이터는 자연어 정책(X)과 기업 맥락(Z)을 입력받아 143개 합성 페르소나로 구성된 가상 조직을 구성하고, 4주 단위의 다단계 워크플로(인식, 대화, 관리자 중재, 감정 합성, 행동 계획, GM 결과 해석)를 통해 국지적 반응을 조직 수준의 네 가지 채널 (생산성, 유지, 웰빙, 공정성)로 변환한다. 각 채널은 5를 무변화 기준으로 하는 0-10 척도로 보고된다. 독립적 HR 전문가의 의견이 가장 수렴하는 최저 분산 부분집합에서, 시뮬레이터는 전문가 prior와 97.5%의 방향 일치도와 0.919의 평균절대오차를 달성한다. 본 도구는 인과적 예측이나 자동화된 HR 의사결정이 아니라, 정책의 다채널 트레이드오프를 사전에 가시화하는 스크리닝 보조 수단으로 한정된다.

키워드: HCI, HR Simulation, AI Agent

I. 서론

HR 정책은 단일 목표에만 독립적으로 영향을 주는 경우가 드물다. 전면 출근 정책은 협업과 관리 가시성을 높이면서도 원격근무에 적응한 직원의 자율성·웰빙·유지를 떨어뜨릴 수 있고, 실시간 성과 대시보드는 실행 가시성을 높이지만 직원이 이를 감시로 받아들이면 신뢰와 공정성 인식을 훼손할 수 있다. 두 경우 모두 정책은 단순히 '좋다' 혹은 '나쁘다'로 규정되지 않으며, 그 기대 가치는 어떤 조직 결과를 기준으로 보느냐, 그리고 누가 그 비용을 부담하느냐에 따라 달라진다.

따라서 핵심 질문은 정책이 평균 점수를 올리는지가 아니라, 여러 조직 가치 채널을 동시에 어느 방향으로 움직이는지다. 본 연구는 생산성, 유지, 웰빙, 공정성의 네 채널에 주목한다. 이들은 상호 의존적이지만 상호 대체되지 않으며, 단일 스칼라로 합치면 정작 검토가 필요한 트레이드오프가 가려진다. 유용한 AI 보조 정책 도구는 정책을 하나의 점수 권고로 압축하기보다, 의사결정자가 채널 간 트레이드오프를 검토하도록 도와야 한다.

단발성 LLM 판단은 기준선으로 유용하지만, 작성된 정책이 조직 반응으로 전환되는 사회적 과정에 대한 가시성은 거의 제공하지 못한다. 우리는 이 문제를 GM 중재 멀티에이전트 시뮬레이터로 다룬다. 의도된·실제·인식된 HR 실천을 구분하는 전략적 HRM 관점을 따라[1], 시뮬레이터는 정책 언어를 직접적인 수치 충격이 아니라 직원이 해석하고 동료가 논의하며 관리자가 중재한 뒤 조직 수준 결과로 누적되는 입력으로 다룬다. 목표는 실제 결과를 예측하는 것이 아니라, 이 단계들을 분리해 보여주는 검사 가능한 트레이스를 구성하고 전문가 기대와 정렬되는 다채널 판단을 산출하는 것이다.

II. 관련 연구

전략적 HRM은 HR 개입이 작성된 정책 내용만으로 조직에 영향을 주지 않으며, 의도된 정책이 실행을 거쳐 실제 실천으로, 직원 해석을 거쳐 인식된 실천으로 전환된다고 본다[1]. 선행 연구는 HRM 실천 체계를 이직·생산성·기업 성과와 연결해 왔으며[2], 이후 리뷰는 HRM–성과 연결이 실행 과정과 직원 반응에 더 주목해야 함을 지적한다[3]. 또한 직무요구–자원(JD-R) 이론은 직무 요구와 자원의 변화를 긴장·몰입·성과와 연결하고[4], 능력–동기–기회(AMO) 관점은 성과를 직원의 능력·동기·기여 기회와 연결하며[5], 사회적 정보처리·이직 전염·심리적 계약·변화 준비도 관점은 직원 반응이 동료 해석과 관리 소통을 통해 확산·강화·분기하는 이유를 설명한다[6]–[9]. 본 연구는 이러한 이론을 완결된 인과 모델이 아니라 잠재 상태와 상호작용 단계를 설계하기 위한 메커니즘 앵커로 활용한다.

에이전트 기반 모델은 이질적 에이전트, 국지적 상호작용, 창발적 거시 행동을 다루는 데 적합하며[10]–[12], 최근 생성형 에이전트 연구는 언어모델이 기억·계획·사회적으로 일관된 행동을 지원할 수 있음을 보였다[13]. 특히 Concordia의 게임마스터 패턴은 조정 모듈이 에이전트의 의도된 행동을 세계 상태 결과로 해석하는 구조를 정식화했으며[14], 본 연구의 GM 계층은 이 패턴을 차용한다. 대규모 생성 에이전트 시스템은 시뮬레이션 집단에 대한 설문·개입을 지원하고[15], 페르소나 집계 접근은 다양한 합성 에이전트가 단일 LLM 응답보다 행동 예측을 개선할 수 있음을 시사한다[16]. 다만 LLM을 심리 시뮬레이터로 보는 연구들은 시뮬레이션 주장에 명시적 과제 설정과 인간 응답 대비 검증이 필요함을 강조하므로[17], 우리는 인간 등가 시뮬레이션이 아니라 전문가 prior 정렬이라는 더 좁은 목표를 설정한다. 마지막으로 다원적 정렬 연구는 AI 시스템이 단일 목표가 아닌 복수의 가치에 직면함을 강조하는데[18], [19], HR 정책 평가도 네 채널이 서로 다른 방향으로 움직일 수 있다는 점에서 동일한 구조를 가진다. 우리는 이를 문제 설정과 평가 규율로 활용한다.

III. 제안방법

각 시나리오는 정책(X)과 맥락(Z)의 쌍으로 정의된다. 정책은 근무시간·보상·평가·모니터링·복지 등의 핵심 개입을 기술하고, 맥락은 조직 성숙도, 노동시장 압력, 사업 성과, 인력 우려, 리더십 프레이밍 등 배경 조건을 명시한다. 시뮬레이터는 이 입력으로 직원 120명, 라인 관리자 15명, HR 비즈니스 파트너 5명, 경영진 3명의 총 143개 합성 페르소나로 구성된 고정 가상 조직을 만든다. 페르소나 스키마와 인구학적 grounding은 PersonaHub[20]와 Nemotron-Personas-Korea[21]를 참고했으나, 최종 로스터는 HR 정책 시뮬레이션을 위해 과제 특

화하여 구성했다. 각 정책은 3개 시드로 4주에 걸쳐 매주 직원 인식, 동료 대화, 관리자 1:1, 성찰 응답을 표본 추출하며, 이 표본은 조직 전체를 망라하려는 것이 아니라 정책이 해석·논의·중재되어 행동으로 전환되는 과정의 대표적 트레이스를 제공한다.

주차별 워크플로는 여섯 단계로 진행된다. 인식은 정책을 역할별 기억에 임베딩하고, 대화는 반응이 동료 교류를 통해 강화·반박·재구성되게 하며, 관리자 중재는 정책 방향을 보존하면서 우려를 인정하는 1:1을 모델링하고, 감정 합성은 팀 단위의 분위기·긴장·공정성 우려·변화 준비도를 요약하며, 행동 계획은 표본 직원이 실제로 무엇을 할지 (노력, 협업, 회의 참여, 순응, 저항, 구직 행동)를 산출하고, GM 결과 해석은 그 트레이스를 읽어 팀·조직 수준의 결과를 결정한다. 모든 LLM 기반 구성요소는 단일 고정 백엔드(Qwen3.5-397B-A17B)로 생성되어, 평가가 모델 백엔드 변이가 아니라 제안 워크플로에 집중되도록 한다[22].

GM 계층은 미시적 에이전트 반응과 거시적 조직 결과를 잇는 다리다. 개별 직원의 좌절·감사·이탈 의도는 곧바로 조직 수준의 생산성·유지·웰빙·공정성과 같지 않다. GM 모듈은 직원 반응이 팀 구조, 동료 영향, 관리 라우팅, 업무 재분배, 이직 동역학과 상호작용한 뒤 조직 수준에서 무엇이 사실이 되는지를 결정함으로써 집계 단계를 조작화한다. 이후 prior 기준 보정(calibration) 단계가 동일 맥락에서 정책이 없는 기준선 대비의 효과에 점수를 맞춰, 보고 점수를 정책의 한계 효과에 집중시킨다.

시뮬레이터는 5를 무변화 기준으로 하는 0–10 척도로 네 채널을 보고한다. 생산성은 지속가능한 산출 역량(실행 품질, 협업 처리량, 혁신, 지식 유지)을, 유지는 잔류 의향과 구직·이직 압력을[7], 웰빙은 스트레스·번아웃·자율성·심리적 안전·소속감을[4], 공정성은 절차적·분배적·상호작용적 정의 인식을 포착하며[23], 동일 척도가 전문가 평정과 시뮬레이터 출력 모두에 사용된다. 이들 채널은 생산성·이직·직원 경험·공정성에 관한 전략적 HRM 및 작업시스템 관점의 관심사와 부합한다[2], [5]. 예비 검증으로, 전문가 간 분산이 가장 낮은 10개 정책 부분집합(전문가가 시뮬레이터 출력을 보기 전에 평정)에서 시뮬레이터는 방향 일치도 97.5%, 평균절대오차 0.919를 달성하여, 전문가 기대가 수렴할 때 같은 정성적 방향을 약 1점 이내에서 회복함을 보인다.

IV. 결론

본 연구는 HR 정책 효과를 단일 효용 점수로 압축하지 않고 네 개의 조직 가치 채널에 걸쳐 표현하는 GM 중재 멀티에이전트 시뮬레이터를 제시했다. 시스템은 자연어 정책과 기업 맥락을 입력받아 이질적 가상 조직을 주차별 워크플로로 전진시키고, 국지적 직원·동료·관리자 반응을 GM 해석 계층을 통해 조직 수준의 네 결과 채널로 변환하며, 전문가 prior가 가장 안정적인 부분집합에서 강한 정렬을 보였다.

본 시스템은 배포 후 결과를 예측하는 인과 예측기가 아니라, 다채널 트레이드오프를 가시화하고 위험 채널과 검토가 필요한 메커니즘을 표면화하는 스크리닝·진단 보조물로 이해되어야 한다. HR 맥락의 특성상 점수가 객관적 사실이나 근로자 목소리의 대체물로 오용되어서는 안 되며, 개인에 대한 처벌적·고위험 결정이나 감시 정당화에 사용되어서는 안 된다. 후속 연구는 정렬 감사를 전체 정책 세트로 확장하고, 멀티에이전트 과정의 기여를 분리하는 ablation을 추가하며, 다양한 LLM 백엔드에 대한 강건성을 검증한다.

참고문헌

[1] Wright, P. M. and Nishii, L. H. Strategic HRM and organizational behavior: Integrating multiple levels of analysis. CAHRS Working Paper WP06-05, Cornell University, 2006.

[2] Huselid, M. A. The impact of human resource management practices on turnover, productivity, and corporate financial performance. Academy of Management Journal, 38(3):635–672, 1995.

[3] Guest, D. E. Human resource management and performance: Still searching for some answers. Human Resource Management Journal, 21(1):3–13, 2011.

[4] Bakker, A. B. and Demerouti, E. The job demands–resources model: State of the art. Journal of Managerial Psychology, 22(3):309–328, 2007.

[5] Appelbaum, E., Bailey, T., Berg, P., and Kalleberg,

A. L. Manufacturing Advantage: Why High-Performance Work Systems Pay Off. Cornell University Press, Ithaca, NY, 2000.

[6] Salancik, G. R. and Pfeffer, J. A social information processing approach to job attitudes and task design. Administrative Science Quarterly, 23(2):224–253, 1978.

[7] Felps, W., Mitchell, T. R., Hekman, D. R., Lee, T. W., Harman, W. S., and Holtom, B. C. Turnover contagion: How coworkers' job embeddedness and job search behaviors influence quitting. Academy of Management Journal, 52(3):545–561, 2009.

[8] Rousseau, D. M. Psychological and implied contracts in organizations. Employee Responsibilities and Rights Journal, 2(2):121–139, 1989.

[9] Armenakis, A. A. and Harris, S. G. Crafting a change message to create transformational readiness. Journal of Organizational Change Management, 15(2):169–183, 2002.

[10] Bonabeau, E. Agent-based modeling: Methods and techniques for simulating human systems. PNAS, 99(suppl. 3):7280–7287, 2002.

[11] Macy, M. W. and Willer, R. From factors to actors: Computational sociology and agent-based modeling. Annual Review of Sociology, 28:143–166, 2002.

[12] Epstein, J. M. Generative Social Science: Studies in Agent-Based Computational Modeling. Princeton University Press, Princeton, NJ, 2006.

[13] Park, J. S., O'Brien, J., Cai, C., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In UIST, pp. 1–22, 2023.

[14] Vezhnevets, A. S., Agapiou, J. P., Aharon, A., et al. Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia. arXiv:2312.03664, 2023.

[15] Piao, J., Yan, Y., Zhang, J., et al. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. arXiv:2502.08691, 2025.

[16] Bhattacharyya, A., Borah, A., Singla, Y. K., Shah, R. R., Ch, C., and Krishnamurthy, B. Social agents: Collective intelligence improves LLM predictions. In ICLR, 2026.

[17] Lin, Z. Large language models as psychological simulators: A methodological guide. Advances in Methods and Practices in Psychological Science, 9(1), 2026.

[18] Gabriel, I. Artificial intelligence, values, and alignment. Minds and Machines, 30(3):411–437, 2020.

[19] Sorensen, T., Moore, J., Fisher, J., et al. A roadmap to pluralistic alignment. In ICML, PMLR 235:46280–46302, 2024.

[20] Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., and Yu, D. Scaling synthetic data creation with 1,000,000,000 personas. arXiv:2406.20094, 2024.

[21] NVIDIA Corporation. Nemotron-Personas-Korea: Synthetic personas aligned to real-world distributions for Korea. Hugging Face dataset, 2026.

[22] Qwen Team. Qwen3.5-397B-A17B. Hugging Face model card, 2026.

[23] Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., and Ng, K. Y. Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. Journal of Applied Psychology, 86(3):425–445, 2001.

반도체 CMP 공정의 MRR 예측 및 계측 절감을 위한 불확실성 보정 기반 선택적 가상계측 프레임워크

허석진[O], 백수진[*]
[O]한신대학교 AI · SW대학 컴퓨터공학부/AI시스템반도체학
[*]한신대학교 AI · SW대학 AI시스템반도체학
e-mail:{mrtrade2[O], croso79[*]}@hs.ac.kr

An Uncertainty-Calibrated Selective Virtual Metrology Framework for MRR Prediction and Metrology Reduction in Semiconductor CMP Processes

SEOKJIN HUH[O], SUJIN BAEK[*]
[O]Dept. of Computer Engineering & AI System Semiconductor, College of AI · SW, Hanshin University
[*]Dept. of AI System Semiconductor, College of AI · SW, Hanshin University

요 약

반도체 CMP 공정에서 Material Removal Rate(MRR)는 wafer 표면 평탄화 품질을 관리하는 핵심 지표이나, 모든 wafer-stage를 실제 계측하는 것은 시간·비용 부담이 크다. 본 연구는 PHM 2016 CMP Data Challenge의 센서 시계열을 WAFER_ID와 STAGE 기준으로 구조화하고, 통계적·시간적 요약값, chamber별 feature 및 proxy feature를 추출하여 MRR 가상계측 모델을 구축하였다. 또한 단일 예측값만으로는 계측 생략 여부를 판단하기 어렵기 때문에, CQR 기반 예측구간으로 sample별 불확실성을 정량화하고 예측구간 폭에 따라 가상계측 허용 여부를 결정하였다. XGBoost는 validation set에서 MAE 2.66, RMSE 3.62, R^2 0.984의 성능을 보였으며, CQR 예측구간은 목표 coverage 90%에 대해 empirical coverage 0.900을 보였다. 선택적 계측 정책에서는 허용오차 δ=7 기준 81.4%의 계측 절감률과 허용오차 초과율(violation rate) 4.29%를 달성하였다. δ=15 기준에서는 전체 wafer-stage를 가상계측으로 대체해도 violation rate가 0.70%로 나타나, 공정 허용오차 수준에 따라 계측 부담과 품질 위험 간 균형을 조절할 수 있음을 보였다.

키워드 : 반도체 공정, CMP, 가상계측, MRR 예측, 불확실성 정량화, Conformalized Quantile Regression, 선택적 실제 계측

I. 서론

반도체 제조 공정에서는 미세한 조건 변화가 후속 공정 품질과 수율에 직접 영향을 미치므로, 단위 공정의 품질 지표를 모니터링하는 것이 중요하다. 특히 CMP(Chemical Mechanical Planarization) 공정은 wafer 표면을 기계적·화학적으로 연마하여 평탄화하는 핵심 공정이며, 노광 및 배선 공정의 정밀도와 안정성에 영향을 준다. 이때 Material Removal Rate(MRR)는 공정 중 제거된 물질의 평균 제거율로 CMP 공정 상태와 품질 변동을 나타내는 주요 관리 지표이다.

그러나 모든 wafer-stage에 대해 MRR을 직접 계측하는 것은 계측 시간, 생산 흐름 지연, 운영 비용 증가를 유발한다. 이를 완화하기 위해 센서 데이터를 활용한 가상계측이 사용되지만, 기존 접근은 주로 예측 성능 향상에 집중해 왔다.

하지만 실제 현장에서는 “MRR을 얼마나 잘 예측하는가”뿐만 아니라, “이 예측을 믿고 실제 계측을 생략해도 되는가”가 더 중요한 의사결정 문제가 된다. 단일 예측값만 제공하는 모델은 안정적으로 예측 가능한 sample과 큰 오차가 발생할 수 있는 sample을 구분하기 어렵기 때문이다.

이에 본 연구는 CMP 센서 시계열을 WAFER_ID와 STAGE 기준의 wafer-stage 단위로 구조화하고, 통계량, chamber별 feature 및 proxy feature를 추출하여 MRR 예측 모델을 구축한다. 또한 CQR 기반 예측구간을 통해 예측 불확실성을 정량화하고, 예측구간 폭을 기준으로 가상계측 허용 여부와 실제 계측 필요 여부를 판단하는 선택적 계측 프레임워크를 제안한다.

최종적으로 XGBoost 기반 MRR 예측과 불확실성 보정 정책을 결합함으로써 가상계측을 단순 예측 도구에서 품질 위험과 계측 부담의 균형을 조절하는 공정 의사결정 도구로 확장하였다. 이를 통해 제안한 방법은 선택적 계측 정책 적용 시 허용오차 δ=7 기준 81.4%의 계측 절감률과 4.29%의 violation rate를 달성하였으며, 결과적으로 공정 허용오차 수준에 따라 계측 부담과 품질 위험 간 균형 조절이 가능함을 확인하였다.

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

II. 선행 연구

2.1 CMP 공정과 MRR 예측

CMP 공정은 wafer 표면을 polishing pad에 접촉시킨 상태에서 압력, 회전, slurry 공급 등을 조절하여 물질을 제거하는 공정이다.

일반적인 CMP 공정에서는 wafer가 carrier에 의해 지지되고, 인가 압력 하에서 polishing pad와 접촉한다. 이때 회전하는 pad/table 및 wafer carrier의 상대 운동과 slurry 공급에 의해 slurry - pad - wafer 계면에서 화학적 반응과 기계적 마찰이 동시에 발생하며, 이를 통해 wafer 표면 물질이 제거되고 평탄화가 이루어진다[1].

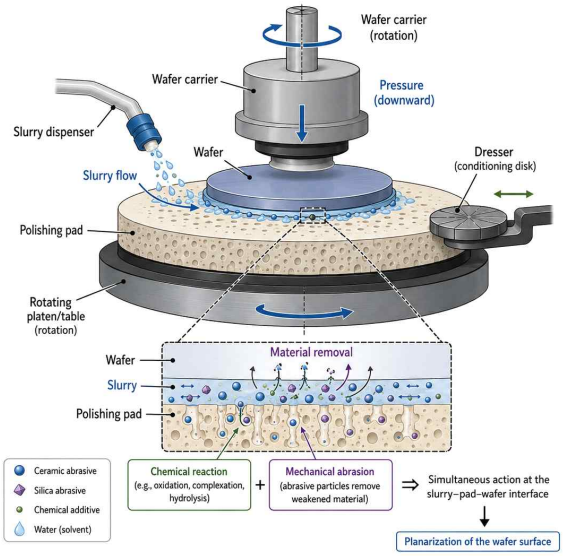


그림 1. CMP 공정 개요

MRR은 polishing pressure, wafer - pad 상대속도, slurry flow, pad surface morphology 등 다양한 공정 및 장비 조건에 영향을 받는 핵심 품질 변수로 알려져 있다[2]. 또한 polishing 과정에서 pad 표면의 glazing, porosity 및 roughness 변화로 인해 제거율과 평탄도가 저하될 수 있으며, dresser conditioning은 pad 표면을 재생하는 역할을 하지만 동시에 pad wear를 유발할 수 있다[3].

따라서 본 연구에서는 polishing pad, dresser, membrane 등 소모성 component의 사용 상태를 MRR 예측을 위한 주요 입력 특성으로 고려하였다.

MRR 예측은 공정 결과를 사전에 추정함으로써 실제 계측 부담을 줄이고 공정 이상을 조기에 감지할 수 있다는 점에서 가상계측 문제로 볼 수 있다.

특히 CMP 공정에서는 pressure, slurry flow, rotation, pad/dresser usage와 같은 공정 변수들이 제거율에 영향을 줄 수 있으므로, 센서 데이터 기반 예측 모델의 활용 가능성이 크다.

2.2 센서 시계열 기반 가상계측

가상계측에서는 실제 계측값이 제품 또는 공정 단위로 제공되는 경우가 많으므로, 센서 시계열을 해당 계측 단위에 맞게 요약하는 과정이 중요하다.

특히 CMP 데이터에서는 시간별 센서 row와 wafer-stage별 MRR label이 분리되어 있으므로, 예측 대상과 입력 단위를 일치시키기 위한 feature aggregation이 필요하다.

2.3 예측 불확실성과 선택적 예측

가상계측 모델을 실제 공정 의사결정에 적용하기 위해서는 단일 예측값만으로는 충분하지 않다. 예측값이 실제 제거율과 얼마나 차이날 수 있는지를 함께 제시해야 실제 계측 생략 여부를 판단할 수 있으며, 이를 위해 예측 불확실성을 정량화한 예측구간(prediction interval)이 필요하다.

본 연구에서는 Quantile Regression과 Conformal Prediction을 결합한 Conformalized Quantile Regression(CQR)을 사용하였다. CQR은 하한·상한 분위수 예측을 기반으로 sample별 예측구간을 생성하고, calibration set을 이용해 목표 coverage를 만족하도록 구간을 보정할 수 있기 때문이다[4].

따라서 제안 프레임워크에서는 CQR을 통해 wafer-stage별 MRR 예측구간을 산정하고, 예측구간 폭을 모델 불확실성의 지표로 사용하였다. 예측구간 폭이 작은 sample은 예측 신뢰도가 높은 것으로 간주하여 가상계측으로 대체하고, 예측구간 폭이 큰 sample은 실제 계측 대상으로 분류하였다.

2.4 사용 데이터셋

반도체 제조 공정 데이터는 공정 recipe, 장비 운전 조건, 센서 로그, 품질 지표 등 기업의 핵심 기술 정보와 밀접하게 연결되어 있어 외부 공개가 매우 제한적이다. 특히 CMP와 같은 단위 공정 데이터는 장비별 운용 전략, 소모성 부품의 사용 이력, 공정 조건 변화, 품질 관리 기준을 간접적으로 드러낼 수 있으므로 최신 제조라인에서 수집된 실제 데이터를 연구 목적으로 확보하기 어렵다. 이러한 제약은 반도체 공정 AI 연구에서 재현 가능한 실험과 방법론 비교를 어렵게 만드는 요인으로 작용한다.

이에 본 연구는 공개적으로 활용 가능하면서도 CMP 공정의 장비 열화와 제거율 예측 문제를 포함하는 PHM Society 2016 Data Challenge 데이터셋[5]을 사용하였다. 해당 데이터셋은 특정 기업의 실제 공정 정보를 직접 노출하지 않도록 익명화 및 스케일링된 형태로 제공되며, CMP 장비의 시간 경과에 따른 성능 저하와 평균 제거율 예측 문제를 다룰 수 있도록 구성되어 있다. 또한 물리 기반 해석과 통계적 예측 접근을 함께 고려하도록 설계되어, 공정 도메인 지식을 반영한 데이터사이언스 방법론의 검증에도 적합하다.

따라서 본 연구에서 해당 데이터셋을 활용한 것은 공개 데이터 사용의 편의성 때문만이 아니라, 보안상 접근이 어려운 반도체 제조 데이터를 대신하여 가상계측과 선택적 실제 계측 방법론을 검증할 수 있는 공신력 있는 벤치마크를 확보하기 위함이다. 이로써 실험 재현성과 방법론 비교 가능성 및 평가 객관성을 모두 함께 확보할 수 있다.

III. 제안 방법

3.1 전체 프레임워크

본 연구는 CMP 공정 센서 시계열을 활용하여 wafer-stage 단위 MRR을 예측하고, 예측 불확실성을 기반으로 실제 계측 여부를 결정하는 선택적 가상계측 프레임워크를 제안한다.

전체 과정은 그림 2와 같이 네 단계로 구성된다. 첫째, 원본 센서 시계열을 WAFER_ID와 STAGE 기준으로 그룹화하여 wafer-stage 단위 sample을 구성한다.

둘째, 각 그룹에서 공정 상태를 요약하는 통계적·시간적·도메인 기반 feature를 추출한다.

셋째, 추출된 feature를 입력으로 사용하여 MRR point prediction model을 학습한다.

넷째, CQR 기반 예측구간을 생성하고, 예측구간 폭을 기준으로 가상계측 허용 여부와 실제 계측 필요 여부를 판단한다.

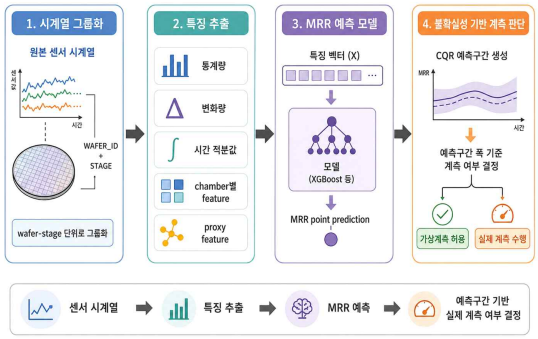


그림 2. 전체 프레임워크 개요

3.2 데이터 구조 및 전처리

본 연구에서 사용한 PHM 2016 CMP 데이터는 센서 시계열 파일과 removal rate label 파일로 구성된다. 센서 데이터는 CMP 공정 중 시간별로 수집된 다변량 측정값이며, row는 시점별 공정 상태를 나타낸다. 제거율 label은 WAFER_ID, STAGE, AVG_REMOVAL_RATE로 구성된다. STAGE는 A 또는 B로 구분되는 서로 다른 processing stage 유형을 의미하며, AVG_REMOVAL_RATE는 WAFER_ID와 STAGE 조합에 대응되는 예측 대상이다.

따라서 본 연구는 원본 row를 독립 sample로 사용하지 않고, WAFER_ID와 STAGE를 기준으로 동일 공정 구간의 센서 관측값을 하나의 시계열 그룹으로 묶었다. 이는 MRR label이 wafer-stage 단위로 제공되기 때문이며, 입력 단위와 예측 단위를 맞추기 위함이다. row 단위로 학습할 경우, 동일 wafer-stage의 일부 시점이 학습과 평가에 동시에 포함되어 data leakage가 발생할 수 있다. 결과적으로 wafer-stage의 시간별 row가 독립 sample로 분리되어 학습과 평가에 동시에 사용되는 row-level leakage를 방지하였다.

이에 본 연구는 wafer-stage 단위 feature vector를 생성한 뒤 removal rate label과 병합하였다. 최종적으로 train 2,006개, calibration set 431개, validation 430개의 wafer-stage sample을 구성하였다.

표 1. 데이터셋 구성

Split	Sample 수	Wafer 수	Stage A	Stage B	Feature 수
Train	2,006	1,699	1,177	829	1,723
Calibration	431	415	241	190	1,723
Validation	430	416	254	176	1,723

3.3 Feature Extraction

본 연구에서는 주요 feature group을 구성하였으며, 각 wafer-stage 그룹에서 센서 시계열의 공정 특성을 반영하기 위해 통계량, 변화량, 시간 적분값, chamber별 feature, 도메인 기반 proxy feature를 추출하였다. 먼저 각 센서 변수에 대해 평균, 표준편차, 최솟값, 최댓값, 중앙값, 사분위수, 범위, 시작값, 종료값, 변화량, slope 등을 계산하여 공정 중 센서값의 대표 수준과 변동 패턴을 요약하였다.

CMP 공정의 MRR은 특정 시점의 순간적인 센서값만으로 결정되기보다, 공정 시간 동안 누적된 압력, slurry 공급량, 회전 조건의 영향을 함께 받는다. 따라서 본 연구에서는 각 센서값을 시간축에 대해 적분하여 공정 중 누적 작용량을 근사하였다. 예를 들어 slurry flow의 시간 적분값은 총 slurry 공급량을, pressure 관련 변수의 적분값은 wafer에 가해진 누적 압력 조건을, rotation 변수의 적분값은 누적 기계적 연마 조건을 반영하는 feature로 해석할 수 있다.

추가적으로 하나의 wafer-stage는 여러 chamber 구간을 포함할 수 있으므로, chamber별 공정 특성을 반영하기 위한 feature를 별도로 생성하였다. 각 chamber에 대해 duration, sensor mean, standard deviation, maximum, minimum, integral 등을 계산하여 전체 공정 요약값만으로는 포착하기 어려운 구간별 차이를 반영하였다.

마지막으로 공개 데이터에는 실제 polishing pad 상태, dresser 마모도, wafer-pad 계면 상태와 같은 물리적 상태 변수가 직접 제공되지 않기 때문에, 센서 변수 조합을 활용한 도메인 기반 proxy feature를 구성하였다. 예를 들어 pressure와 duration의 곱은 압력의 누적 작용량을, slurry flow 적분값의 합은 slurry 총 투입량을, head rotation과 wafer rotation의 차이는 wafer-pad 상대 운동 조건을 근사한다. 이러한 proxy feature는 CMP 공정에서 제거율에 영향을 줄 수 있는 물리적 요인을 직접 측정값 없이 간접적으로 표현하기 위한 것이다.

표 2. 주요 feature group

Feature group	예시	의미
Basic statistics	mean, std, min	센서 시계열의 전체 공정 특성
Chamber-level features	ch1 pressure mean, ch1 slurry integral	chamber별 공정 구간 특성
Usage features	backing film usage, dresser usage	소모성 부품 상태 proxy
Physics-aware proxy	pressure × duration	CMP 물리적 조건 근사

3.4 MRR 예측 모델

추출된 wafer-stage feature vector를 입력으로 사용하고, AVG_REMOVAL_RATE를 예측 대상으로 설정하여 MRR point prediction model을 학습하였다.

CMP 공정 feature는 pressure, slurry flow, rotation, usage 변수 간 비선형 상호작용을 포함할 가능성이 높기 때문에, 본 연구에서는 정형 데이터에서 비선형 관계 학습에 강점을 갖는 트리 기반 회귀 모델을 중심으로 비교하였다.

비교 모델로는 Stage Mean Predictor, Random Forest, Extra Trees, Gradient Boosting, LightGBM, XGBoost, CatBoost를 사용하였다. 특히 XGBoost는 gradient boosting 기반의 트리 앙상블 모델로, 정형 데이터에서 비선형 관계와 변수 간 상호작용을 효과적으로 학습할 수 있어 본 연구의 MRR 예측 후보 모델에 포함하였다[6].

또한 구성한 train split은 모델 학습에 사용하고, validation split에서 모델별 예측 성능을 비교하였다. 이후 가장 우수한 성능을 보인 모델을 후속 불확실성 보정 및 선택적 계측 정책 평가의 기준 모델로 사용하였다. 평가지표는 MAE, RMSE, R^2를 사용하였으며, calibration set은 CQR 기반 예측구간 보정을 위해 사용하였다.

3.5 CQR 기반 예측구간 보정

본 연구에서는 sample별 예측 불확실성을 산정하기 위해 CQR을 사용하였다.

먼저 train set을 이용해 lower quantile model과 upper quantile model을 학습하였다. 이후 calibration set을 사용하여 conformal score를 계산하였다.

각 sample i에 대해 lower prediction을 L_i, upper prediction을 U_i 실제값을 y_i라고 할 때, conformal score는 다음과 같이 정의된다.

$$S_i = \max(L_i - y_i,\ y_i - U_i,\ 0)$$

Calibration set에서 계산된 conformal score의 보정 분위수를 q로 사용하여, validation sample의 예측구간을 다음과 같이 확장하였다.

$$\hat{C}(x) = [\hat{L}(x) - q,\ \hat{U}(x) + q]$$

목표 coverage를 0.90으로 설정하고, lower quantile model은 α/2 = 0.05 분위수, upper quantile model은 1 − α/2 = 0.95 분위수를 예측하도록 학습하였다.

3.6 선택적 실제 계측 정책

먼저 각 wafer-stage에 대해 MRR 예측값 $\hat{y}$, 예측구간 하한 L, 상한 U, 예측구간 폭 W를 계산한다.

$$W = U - L$$

선택적 실제 계측 정책은 다음과 같다.

$$W \le \tau_\delta \Rightarrow \text{가상계측허용}$$
$$W > \tau_\delta \Rightarrow \text{실제계측수행}$$

여기서 τ_δ는 허용오차 δ 조건에서 선택된 예측구간 폭 threshold로, 가상계측 허용 여부를 결정하는 불확실성 기준이다. 또한 τ_δ를 하나의 고정값으로 설정하지 않고, calibration set에서 예측구간 폭 W의 후보 threshold를 변화시키며 선택적 계측 정책의 운영점을 탐색하였다. 구체적으로 허용오차 δ별로 calibration set에서 violation rate가 목표 위험 수준(ρ=0.05) 이하로 유지되는 후보 중 계측 절감률이 가장 큰 threshold를 τ_δ로 선택하고, 이를 validation set에 적용하여 최종 정책 성능을 평가하였다. 이때, δ가 커질수록 violation 판정 기준이 완화되어 더 큰 τ_δ가 선택될 수 있지만, 목표 위험 수준 ρ 자체는 고정된다. 또한 ρ는 calibration set에서 허용하고자 하는 최대 violation rate이다.

공개 데이터에는 실제 제조라인의 spec limit이 포함되어 있지 않으므로, 본 연구에서는 선택적 계측 정책의 민감도 분석을 위해 여러 허용오차 δ를 운영 기준으로 설정하였다.

$$Violation = \begin{cases} 1, & |y - \hat{y}| > \delta \\ 0, & |y - \hat{y}| \le \delta \end{cases}$$

최종적으로 선택적 계측 정책의 성능은 계측 절감률, accepted MAE, violation rate를 기준으로 평가하였다.

IV. 실험 및 결과

4.1 MRR 예측 성능

Validation set 기준 MRR 예측 성능은 표 3과 같다.

표 3. MRR 예측 모델 성능 비교

Model	MAE ↓	RMSE ↓	R^2 ↑
Stage Mean	26.48	31.20	-0.195
Random Forest	3.91	16.52	0.665
Extra Trees	4.32	18.45	0.582
Gradient Boosting	5.09	30.15	-0.116
LightGBM	9.85	23.92	0.298
XGBoost	**2.66**	**3.62**	**0.984**
CatBoost	4.89	13.20	0.786

본 연구에서는 여러 회귀 모델의 validation 성능을 비교하고, 가장 우수한 성능을 보인 XGBoost를 후속 불확실성 보정 및 선택적 계측 정책 평가의 기준 모델로 사용하였다.

최종 모델로 선정한 XGBoost는 squared error 기반 회귀 목적함수를 사용하였으며, 주요 hyperparameter는 트리 수 n_estimators=600, 최대 깊이 max_depth=4, learning rate=0.03으로 설정하였다. 또한 과적합을 완화하기 위해 subsample=0.9, colsample_bytree=0.9를 적용하였다.

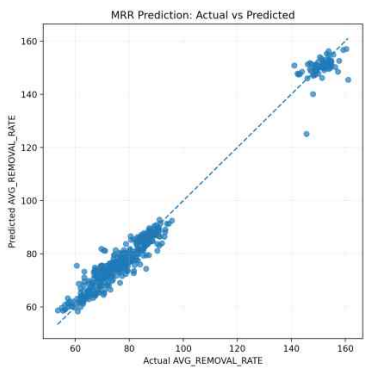


그림 3. XGBoost의 MRR 예측 산점도

추가적으로 최종 모델인 XGBoost의 stage별 성능을 확인한 결과, Stage A와 Stage B 모두에서 MAE 3 이하의 안정적인 예측 성능을 보였다. 이는 제안한 wafer-stage 단위 feature extraction과 MRR 예측 모델이 특정 stage에만 편향되지 않고, 서로 다른 processing stage에서도 일관된 예측 성능을 보였음을 의미한다.

다만 본 연구의 초점은 stage 간 차이 분석이 아니라 예측 불확실성 기반 선택적 실제 계측 정책이므로, stage별 결과는 모델 안정성 확인을 위한 보조 분석으로 활용하였다.

표 4. XGBoost 모델의 Stage별 validation 성능

Stage	Sample 수	MAE ↓	RMSE ↓	R^2 ↑
A	254	2.803	3.850	0.988
B	176	2.456	3.251	0.885

4.2 예측구간 보정 성능

CQR 기반 예측구간의 품질은 표 5와 같다. PICP는 실제값이 예측구간 안에 포함된 비율을 의미하며, MPIW는 예측구간 폭의 평균값을 의미한다.

Calibration set의 PICP는 0.9049로 나타났으며, 이는 보정 분위수 q가 목표 coverage 수준에 맞게 산출되었음을 확인하는 진단 지표로 해석할 수 있다.

독립적인 평가 기준인 validation set에서는 PICP가 0.9000으로 나타나 목표 coverage 0.90과 일치하였다.

이는 제안한 예측구간이 단순 point prediction을 넘어, 각 wafer-stage 예측값의 신뢰도를 정량화하는 기준으로 활용될 수 있음을 의미한다.

표 5. CQR 기반 예측구간 보정 성능

Split	Target Coverage	PICP	MPIW	Median Width	Coverage Error
Calibration	0.90	0.9049	25.83	17.16	0.0049
Validation	0.90	0.9000	25.09	17.74	0.0000

그림 4는 CQR 기반 예측구간을 시각화한 결과이다. 대부분의 실제 MRR이 예측구간 내부에 포함되어, 제안한 불확실성 보정 방법이 목표 coverage를 안정적으로 만족함을 확인할 수 있다. 특히 고MRR 영역 또는 regime transition 구간에서 예측구간이 넓어지는 현상은 모델이 해당 sample의 불확실성을 높게 평가했음을 의미하며, 이는 선택적 실제 계측 정책의 판단 근거로 활용된다.

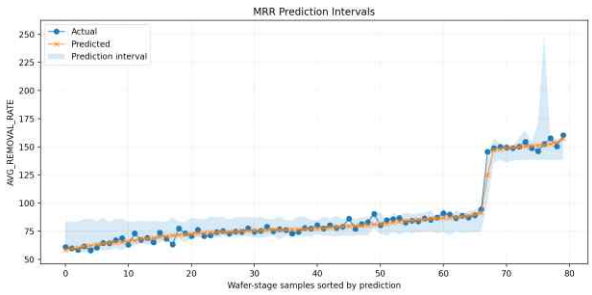


그림 4. CQR 기반 MRR 예측구간 시각화

4.3 선택적 실제 계측 정책 평가

선택적 계측 정책의 성능은 허용오차 δ에 따라 달라진다. 이때 δ는 모델의 하이퍼파라미터가 아니라, 공정에서 허용 가능한 MRR 예측 오차 기준을 의미한다.

본 연구에서는 공개 데이터에 실제 제조라인의 spec limit이 포함되어 있지 않으므로, 민감도 분석을 위해 허용오차 δ를 변화시키며 선택적 계측 정책의 민감도를 분석하였다. 따라서 실제 적용 시 δ는 공정별 품질 기준, 제품 요구사항 및 계측 운영 정책에 따라 달라질 수 있다. δ가 클수록 violation rate가 낮아지는 것은 자연스러운 결과이며, 서로 다른 δ는 서로 다른 공정 운영 기준으로 해석해야 한다.

또한 선택된 τ_δ는 calibration set에서 결정되므로, validation set에서의 violation rate는 데이터 분포 차이에 따라 목표 위험 수준 ρ=0.05와 일부 차이를 보일 수 있다.

표 6. 허용오차 δ에 따른 선택적 가상계측 정책의 민감도 분석

δ	τ_δ	가상계측 허용 수	실제 계측 필요 수	계측 절감률	Accepted MAE	Violation Rate
3	12.89	3	427	0.70%	1.52	0.00%
5	14.11	17	413	3.95%	1.93	5.88%
7	23.24	350	80	81.40%	2.55	4.29%
15	479.58	430	0	100%	2.66	0.70%

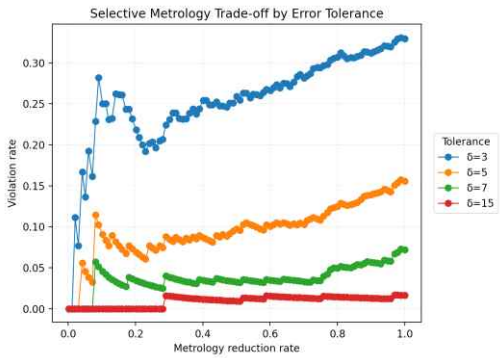


그림 5. 허용오차 δ별 예측구간 폭 threshold 변화에 따른 계측 절감률 - 위험률 trade-off

허용오차 δ=7 기준에서는 validation set의 430개 wafer-stage 중 350개를 가상계측으로 허용하여 81.4%의 계측 절감률을 달성하였다. 이때 violation rate는 4.29%로 나타나, 엄격한 품질 기준에서도 5% 이하의 위험률을 유지하면서 실제 계측 부담을 크게 줄일 수 있음을 확인하였다.

반면, δ=15 기준에서는 전체 430개 wafer-stage를 모두 가상계측으로 대체해도 violation rate가 0.70%에 그쳤다. 이는 공정에서 ±15 수준의 MRR 오차를 허용할 수 있는 운영 조건이라면, 가상계측 적용 범위를 크게 확대할 수 있음을 보여준다. 다만 δ=15의 우수한 결과는 모델이 더 우수하게 작동했기 때문이라기보다, 공정 허용오차 기준이 완화되었기 때문으로 해석해야 한다.

또한 δ=5의 경우 accepted sample 수가 17개로 작아 단일 violation에도 violation rate가 5.88%로 나타나므로, 낮은 acceptance 영역에서는 sample 수와 함께 해석해야 한다.

그림 5는 예측구간 폭 threshold 변화에 따른 계측 절감률과 violation rate의 trade-off를 나타낸다. 불확실성이 낮은 sample부터 가상계측으로 허용할수록 초기에는 낮은 violation rate를 유지할 수 있으나, 계측 절감률을 높일수록 허용오차를 초과하는 sample 비율이 증가한다. 이는 실제 공정에서 요구되는 허용오차 수준에 따라 적절한 운영점을 선택해야 함을 의미한다.

따라서 본 연구의 선택적 계측 정책은 공정 허용오차 수준에 따라 운영점을 조정할 수 있다. 엄격한 조건에서는 불확실성이 큰 sample만 실제 계측 대상으로 남기고, 완화된 조건에서는 전체 가상계측으로 확장할 수 있다.

4.4 Feature Importance 분석

XGBoost 모델의 gain 기준 feature importance 분석 결과, 상위 feature는 주로 usage 계열 변수로 나타났다.

표 7. 주요 Feature Importance (XGBoost gain 기준)

순위	Feature	Importance
1	Chamber 1 backing film usage integral	0.245
2	Chamber 1 membrane usage integral	0.188
3	Chamber 1 backing film usage mean	0.080
4	Chamber 1 pressurized sheet usage integral	0.076
5	Dresser table usage minimum	0.074
6	Chamber 3 backing film usage integral	0.066
7	Membrane usage slope	0.057
8	Chamber 1 stage rotation standard deviation	0.026
9	Initial machine ring-location ID	0.024
10	Chamber 1 head rotation standard deviation	0.023

세부적으로는 backing film, membrane, pressurized sheet, dresser table 등 소모성 부품 및 장비 상태와 관련된 usage 계열 feature가 MRR 예측에 큰 영향을 미치는 것으로 나타났다.

이는 CMP 공정에서 제거율이 단순한 단일 센서값이 아니라 slurry - pad - wafer 계면에서 발생하는 화학적 반응과 기계적 마찰의 복합 작용에 의해 결정된다는 기존 연구와 일치한다[1].

또한 MRR은 polishing pressure, wafer - pad 상대속도, slurry flow, pad surface morphology 등 다양한 공정 및 장비 조건에 영향을 받는 것으로 알려져 있으며[2], polishing 과정에서 pad 표면의 glazing, porosity, roughness 변화가 제거율과 평탄도 저하로 이어질 수 있다[3].

Dresser conditioning은 pad 표면을 재생하여 polishing 성능을 회복시키는 역할을 하지만, 동시에 pad wear를 유발할 수 있다는 점에서 dresser 및 pad 관련 usage 변수는 장비 component 상태를 반영하는 중요한 간접 지표로 해석될 수 있다[3].

이는 제안한 feature extraction이 단순 통계량에 머무르지 않고, CMP 공정에서 누적 사용량과 소모성 component 상태가 제거율 변화에 미치는 영향을 모델 입력으로 효과적으로 반영했음을 보여준다.

따라서 본 연구의 feature importance 결과는 제안 모델이 단순한 통계적 패턴뿐 아니라 CMP 공정의 소모성 component 상태와 장비 열화 특성이 MRR 예측에 미치는 영향을 학습했음을 시사한다.

Ⅴ. 결론

본 연구는 반도체 CMP 공정 센서 시계열 데이터를 활용하여 wafer-stage 단위 MRR 가상계측 모델을 구축하고, 예측 불확실성 기반 선택적 실제 계측 정책을 제안하였다. 원본 센서 row를 직접 학습하지 않고, WAFER_ID와 STAGE 기준으로 시계열을 그룹화한 뒤 전체 통계량, chamber별 feature, CMP 도메인 기반 proxy feature를 구성하였다.

실험 결과, XGBoost 기반 MRR 예측 모델은 validation set에서 MAE 2.66, RMSE 3.62, R^2 0.984의 높은 예측 성능을 보였다. 또한 CQR 기반 예측구간은 validation set에서 목표 coverage 90%를 정확히 달성하여, 가상계측 예측값의 신뢰도를 정량화할 수 있음을 확인하였다.

선택적 실제 계측 정책에서는 허용오차 δ=7 기준 81.4%의 계측 절감률과 4.29%의 violation rate를 달성하였으며, δ=15 기준에서는 전체 validation sample을 가상계측으로 대체해도 violation rate가 0.70%로 나타났다. 이는 공정에서 허용 가능한 MRR 오차 수준에 따라 실제 계측 부담을 크게 줄일 수 있음을 보여준다.

본 연구의 의의는 CMP 가상계측을 단순한 예측 문제에서 확장하여, 예측 불확실성을 기반으로 실제 계측 여부를 결정하는 risk-aware metrology decision 문제로 재정의했다는 점에 있다.

향후 연구에서는 실제 공정별로 정의되는 허용오차와 계측 주기, 운영상 제약을 반영해 threshold 선택 기준을 구체화하고, 장비·기간·stage 변화에 따른 예측구간 폭과 계측 절감 효과의 재현성을 검증할 계획이다.

참고문헌

[1] J. Seo, “A review on chemical and mechanical phenomena at the wafer interface during chemical mechanical planarization,” Journal of Materials Research, Vol. 36, pp. 235-257, 2021.
[2] T. Yu, Z. Li, and D. Wu, “Predictive modeling of material removal rate in chemical mechanical planarization with physics-informed machine learning,” Wear, Vol. 426-427, pp. 1430-1438, 2019.
[3] B. J. Hooper, G. Byrne, and S. Galligan, “Pad conditioning in chemical mechanical polishing,” Journal of Materials Processing Technology, Vol. 123, No. 1, pp. 107-113, 2002.
[4] Y. Romano, E. Patterson, and E. J. Candès, “Conformalized Quantile Regression,” Advances in Neural Information Processing Systems, Vol. 32, 2019.
[5] PHM Society, “PHM Data Challenge 2016: Wafer Chemical-Mechanical Planarization System,” Annual Conference of the Prognostics and Health Management Society, pp. 1-6, 2016.
[6] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

긴급재난문자 및 지역 취약지표를 결합한 공공 사회위험 조기탐지 서비스 설계

김시현*,O, 안현*
*한신대학교 AI·SW대학
e-mail: {ksh402, hyunahn}@hs.ac.kr*

Design of a Public Social Risk Early Detection Service Integrating Emergency Alert Messages and Regional Vulnerability Indicators

Sihyeon Kim*,O, Hyun Ahn*
*School of Computing and Artificial Intelligence, Hanshin University

요 약

본 연구는 긴급재난문자와 지역 취약지표를 결합하여 공공 사회위험을 조기에 파악할 수 있는 서비스 구조를 설계하는 것을 목적으로 한다. 긴급재난문자는 폭염, 호우/침수, 화재/산불, 한파 등 지역사회에서 발생하는 위험 정보를 신속하게 반영하는 공공 데이터이지만, 주로 개별 알림 정보로 활용되어 왔다. 이에 본 연구는 2024년부터 2025년까지의 긴급재난문자 데이터를 수집하고, 문자 내용 기반 키워드 분류를 통해 재난유형을 구분하였다. 이후 수신지역 정보를 활용하여 전국 시군구 단위로 데이터를 매핑하고, 재난유형별 문자 발생 빈도를 지역별 재난 신호 지표로 변환하였다. 또한 고령인구비율, 독거노인가구비율, 응급의료기관 수를 지역 취약지표로 활용하여 재난유형별 위험 신호와 지역 대응 취약성을 함께 반영하였다. 각 지표는 정규화 과정을 거친 뒤 결합되어 사회위험점수로 산출되며, 이를 기반으로 재난유형별 취약지역을 도출하고 지도 및 그래프 형태로 시각화하는 서비스 흐름을 제안하였다. 본 연구는 긴급재난문자를 단순한 알림 데이터가 아니라 지역별 위험 신호 데이터로 활용하고, 지자체가 폭염, 호우/침수, 화재/산불 등 재난유형별 대응 우선지역을 파악하는 데 활용 가능한 분석 기반을 제시한다.

키워드: 긴급재난문자, 사회위험점수, 조기탐지 서비스, 지역 취약지표, 데이터 파이프라인

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

I. 서론

최근 폭염, 집중호우, 산불, 한파, 실종 등 다양한 재난 및 안전 이슈가 지역사회에서 반복적으로 발생하고 있다. 이러한 상황에서 긴급재난문자는 시민에게 위험 정보를 신속하게 전달하는 중요한 공공 안전 수단으로 활용된다. 긴급재난문자는 특정 시점의 재난 상황, 주의사항, 대피 안내, 실종자 정보 등을 포함하고 있어 지역사회에서 나타나는 위험 신호를 실시간에 가깝게 반영하는 공공 데이터로 볼 수 있다[1].

그러나 긴급재난문자는 대부분 개별 알림 정보로 활용되는 경우가 많다. 해당 문자가 어느 지역에서 얼마나 자주 발생했는지, 특정 재난유형이 어느 지역에서 반복되는지, 취약계층이 많은 지역과 재난 신호가 어떻게 겹치는지는 체계적으로 분석되기 어렵다. 따라서 긴급재난문자는 단순 알림 데이터가 아니라 지역별 공공 사회위험을 파악하기 위한 데이터 자원으로 활용될 필요가 있다.

특히 재난위험은 단순히 재난 신호의 발생 빈도만으로 설명되기 어렵다. 같은 폭염 문자가 발송되더라도 고령인구와 독거노인가구 비율이 높은 지역에서는 피해 위험이 상대적으로 커질 수 있다[2]. 또한 응급의료기관이 부족한 지역은 재난 발생 이후 신속한 대응이 어려울 수 있다. 이에 본 연구는 긴급재난문자의 재난유형별 발생 빈도와 지역 취약지표를 결합하여 전국 시군구별 공공 사회위험점수를 산출하는 데이터 파이프라인을 제안한다.

본 연구에서 제안하는 서비스 구조는 긴급재난문자를 단순한 알림 데이터가 아니라 지역별 위험 신호 데이터로 활용할 수 있다는 점에서 의의가 있다. 또한 재난문자 발생 빈도와 고령인구비율, 독거노인가구비율, 응급의료기관 수 등 지역 취약지표를 결합함으로써 지자체가 재난유형별 대응 우선지역을 파악하는 데 활용될 수 있다. 이를 통해 폭염, 호우/침수, 화재/산불 등 재난유형별 취약지역을 사전에 점검하고, 취약계층 보호와 공공 안전 인프라 개선을 위한 기초 자료로 활용될 것으로 기대된다.

II. 관련 연구

긴급재난문자는 재난 발생 시 지역 주민에게 위험 정보를 신속하게 전달하는 공공 안전 커뮤니케이션 수단이다. 선행연구에서는 재난문자를 자연어처리 기반으로 분류하거나[3], 재난문자의 긴급성을 평가하는 연구가 수행되었다[4]. 또한 COVID-19 시기 긴급재난문자의 발송 현황을 월별·지역별로 분석하고 시각화한 연구도 있다[5]. 이러한 연구들은 긴급재난문자가 단순한 개별 메시지가 아니라 시간적·공간적으로 분석 가능한 공공 데이터임을 보여준다.

한편 재난위험은 재난 발생 신호뿐만 아니라 지역의 구조적 취약성과 함께 해석될 필요가 있다. 고령화, 인구감소, 재난 대응 인프라 접근성 등은 재난 취약지역 판단에서 중요한 요소로 논의되어 왔으며[2], 안전취약계층을 고려한 재난안전 대응 서비스의 필요성도 제기되어 왔다[6]. 본 연구는 기존 연구와 달리 긴급재난문자의 재난유형별 발생 빈도를 시군구 단위 재난 신호로 변환하고, 고령인구비율, 독거노인가구비율, 응급의료기관 수와 결합하여 사회위험점수를 산출한다는 점에서 차별성을 가진다.

III. 제안방법

3.1 분석 데이터 및 전처리

본 연구는 2024년 1월 1일부터 2025년 12월 31일

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

까지의 긴급재난문자 데이터를 분석 대상으로 설정하였다. 2023년 자료는 연중 자료가 불완전하고, 2026년 자료는 분석 시점 기준으로 일부 기간만 존재하므로 연도 간 비교 가능성과 계절적 완전성을 확보하기 위해 제외하였다.

본 연구에서 활용한 데이터는 표 1과 같다. 긴급재난문자 데이터는 행정안전부 긴급재난문자 API를 활용하였으며[1], 고령인구비율과 독거노인가구비율은 KOSIS 자료를 사용하였다[7-8]. 응급의료기관 현황은 공공데이터포털의 전국응급의료기관표준데이터를 활용하였다[9].

데이터명	출처	기준	활용 목적
긴급재난문자	행정안전부	2024~2025	재난 신호
고령인구비율	KOSIS	2024.12	고령층 취약성
독거노인가구비율	KOSIS	2024	사회적 취약성
응급의료기관 현황	공공데이터포털	수집 시점 기준	대응 인프라

표 1. 활용 데이터 목록

긴급재난문자 데이터에서는 재난문자 고유번호, 발송일시, 수신지역명, 문자내용을 활용하였다. 발송일시는 분석 기간 확인과 재난문자 집계 기준으로 활용하였고, 수신지역명은 시도·시군구 매핑에 사용하였다. 하나의 문자가 복수 지역에 발송된 경우에는 해당 지역별로 행을 분리하여 시군구 단위 집계가 가능하도록 처리하였다.

재난유형은 문자 내용 기반 키워드 매칭 방식으로 분류하였다. 분류 유형은 폭염, 호우/침수, 화재/산불, 한파, 실종/배회, 대설, 강풍/태풍, 지진, 감염병, 도로/교통안전, 생활안전, 기타의 총 12개로 구성하였다. 예를 들어 폭염은 '폭염', '무더위', '온열질환' 등의 키워드로, 호우/침수는 '호우', '침수', '하천', '산사태', '토사유출' 등의 키워드로 분류하였다. 전처리 결과, 원본 기준 고유 재난문자는 38,125건으로 나타났다. 복수 지역에 발송된 문자를 시군구별로 분리한 결과, 최종 분석에 사용한 지역 단위 재난문자 데이터는 49,900행으로 구성되었다. 또한 재난유형 분류 결과 기타 유형의 비율은 7.83%로 나타났다.

3.2 데이터 파이프라인

본 연구의 데이터 파이프라인은 긴급재난문자 원자료를 지역 단위의 재난 신호 지표로 변환하고, 이를 지역 취약지표와 결합하여 최종 사회위험점수를 산출하는 구조로 설계하였다. 전체 처리 흐름은 그림 1과 같다.

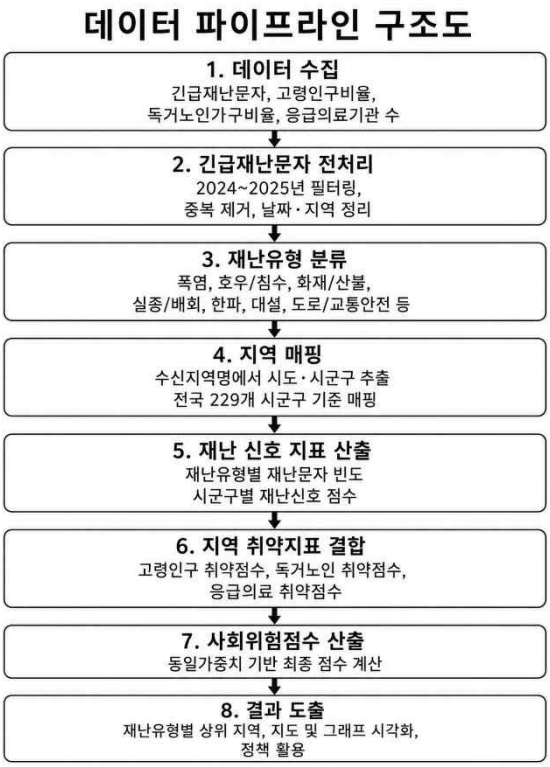


그림 1. 공공 사회위험 조기탐지 서비스 파이프라인

그림 1에서 제시한 바와 같이, 먼저 긴급재난문자

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

데이터와 고령인구비율, 독거노인가구비율, 응급의료기관 수 등 지역 취약지표 데이터를 수집하였다.
이후 긴급재난문자 데이터에서 2024년과 2025년 자료만 추출하고, 동일 레코드 중복 제거, 날짜 정리, 수신지역 정리, 문자내용 정제를 수행하였다.
다음으로 문자 내용 기반 키워드 매칭을 통해 재난유형을 분류하고, 수신지역명에서 시도와 시군구를 추출하여 전국 229개 시군구 기준으로 지역 매핑을 수행하였다.
이후 시군구별 재난문자 건수를 단순 전체 건수로 집계하지 않고, 폭염, 호우/침수, 화재/산불 등 재난유형별로 분리하여 지역별 재난신호 지표를 산출하였다. 마지막으로 고령인구 취약점수, 독거노인 취약점수, 응급의료 취약점수를 결합하여 재난유형별 사회위험점수를 계산하였다. 이러한 절차는 특정 지역의 전체 위험 수준뿐 아니라, 재난유형별 대응 우선지역을 도출할 수 있다는 점에서 의미가 있다.

3.3 사회위험점수 산출

본 연구에서는 공공 사회위험을 특정 지역에서 재난 신호가 반복적으로 나타나고, 동시에 지역의 구조적 취약성이 높은 상태로 정의하였다. 이에 따라 사회위험점수는 재난문자 기반 재난 신호 지표와 지역 취약지표를 결합하여 산출하였다. 재난 신호 지표는 재난유형별 긴급재난문자 발생 빈도를 활용하였으며, 지역 취약지표는 고령인구비율, 독거노인가구비율, 응급의료기관 수를 활용하였다. 사회위험점수 산출에 사용한 지표와 가중치는 표 2와 같다.

구분	지표	산출방식	해석	가중치
재난 신호	재난유형별 재난문자 빈도	재난유형별 문자 건수 정규화	해당 재난유형의 위험 신호	0.25
지역 취약성	고령인구 비율	최댓값 기준 정규화	고령층 취약성	0.25
	독거노인 가구비율	최댓값 기준 정규화	사회적 취약성	0.25
	응급의료 기관 수	1 - 기관 수 정규화	응급의료 취약성	0.25

표 2. 사회위험점수 산출 지표 및 가중치

지역 r, 재난유형 d에 대한 긴급재난문자 건수를 N_{r,d}라고 할 때, 재난신호 점수 A_{r,d}는 해당 재난유형에서 가장 많은 문자 건수를 기준으로 정규화하였다.

$$A_{r,d} = N_{r,d} / \max_r(N_{r,d}) \quad (1)$$

즉, 동일 재난유형에서 재난문자 건수가 가장 많은 지역은 1의 값을 가지며, 나머지 지역은 최댓값 대비 상대적 비율로 산출된다. 이를 통해 폭염, 호우/침수, 화재/산불 등 재난유형별로 지역 간 위험 신호 수준을 비교할 수 있다.

지역 취약지표는 고령인구비율, 독거노인가구비율, 응급의료기관 수를 활용하였다. 고령인구 취약점수 E_r와 독거노인 취약점수 L_r는 각각 해당 지표의 최댓값을 기준으로 정규화하였다.

$$E_r = Old_r / \max_r(Old_r) \quad (2)$$

$$L_r = Alone_r / \max_r(Alone_r) \quad (3)$$

여기서 Old_r는 지역 r의 고령인구비율, Alone_r는 지역 r의 독거노인가구비율을 의미한다. 두 지표는 값이 클수록 재난 상황에서 취약성이 높은 것으로 해석하였다.

응급의료 취약점수 M_r는 응급의료기관 수가 적을수록 취약성이 높아지는 구조를 반영하기 위해 다

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

음과 같이 산출하였다.

$$M_r = 1 - Med_r / max_r(Med_r) \quad (4)$$

여기서 Med_r는 지역 r의 응급의료기관 수를 의미한다. 응급의료기관 수가 많은 지역은 응급 대응 여건이 상대적으로 양호하다고 판단하여 낮은 취약점수를 부여하고, 기관 수가 적은 지역은 높은 취약점수를 갖도록 하였다.

최종적으로 지역 r의 재난유형 d에 대한 사회위험점수 R_{r,d}는 재난신호 점수와 세 가지 지역 취약점수를 결합하여 다음과 같이 산출하였다.

$$R_\{r,d\} = 100 \times (0.25A_\{r,d\} + 0.25E_r + 0.25L_r + 0.25M_r \quad (5)$$

본 연구에서 활용한 네 가지 지표 조합에 대해 공식적으로 제시된 표준 가중치는 확인되지 않았다. 따라서 본 연구는 공공 사회위험 조기탐지 서비스 설계 단계의 탐색적 연구라는 점을 고려하여 특정 지표의 영향력을 임의로 과대 반영하지 않고, 재난신호 점수, 고령인구 취약점수, 독거노인 취약점수, 응급의료 취약점수에 동일가중치를 적용하였다. 이에 따라 본 연구의 사회위험점수는 국가 공식 지수가 아니라, 재난유형별 위험 신호와 지역 취약성을 결합하여 대응 우선지역을 도출하기 위한 탐색적 지표로 해석할 필요가 있다.

3.4 분석 결과

먼저 최종 분석에 사용한 지역 단위 긴급재난문자 데이터를 재난유형별로 집계한 결과, 폭염 관련 문자가 12,429건으로 가장 많이 나타났으며, 호우/침수 11,849건, 화재/산불 6,096건 순으로 높은 빈도를 보였다. 그림 2는 전체 분석 데이터에서 재난유형별 긴급재난문자 빈도를 나타낸 것이다. 이를 통해 분석 기간 동안 폭염과 호우/침수와 같은 기후위험 관련 재난문자가 높은 비중을 차지했음을 확인할 수 있다.

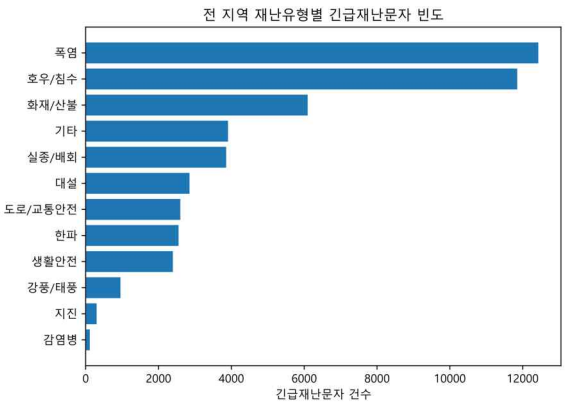


그림 2. 전 지역 재난유형별 긴급재난문자 빈도

이후 본 연구는 재난문자를 단순히 전체 건수 기준으로만 분석하지 않고, 재난유형별로 분리한 뒤 시군구 단위로 재집계하였다. 이를 통해 각 지역에서 어떤 재난유형의 신호가 상대적으로 강하게 나타나는지 파악하고, 재난유형별 사회위험점수를 산출하였다. 사회위험점수는 표 2에서 제시한 바와 같이 재난유형별 재난문자 발생 빈도, 고령인구 취약점수, 독거노인 취약점수, 응급의료 취약점수를 동일가중치로 결합하여 산출하였다. 따라서 사회위험점수 상위 지역은 단순히 특정 재난문자가 많이 발송된 지역이 아니라, 재난유형별 위험 신호와 지역 취약성이 함께 높게 나타난 지역으로 해석할 수 있다.

폭염 유형에서는 경상북도 의성군이 96.10점으로 가장 높게 나타났으며, 대구광역시 군위군 83.00점, 경상남도 산청군 81.71점, 경상북도 영덕군 81.28점, 전라남도 곡성군 80.27점 순으로 나타났다. 이

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

는 폭염 재난문자 발생 빈도와 고령인구, 독거노인 가구, 응급의료 취약성이 함께 반영된 결과이다.

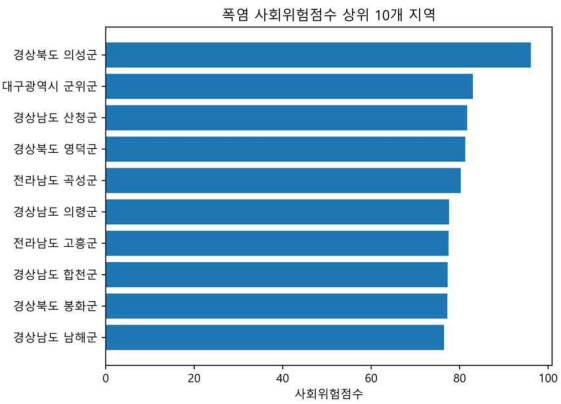


그림 3. 폭염 사회위험점수 상위 10개 지역

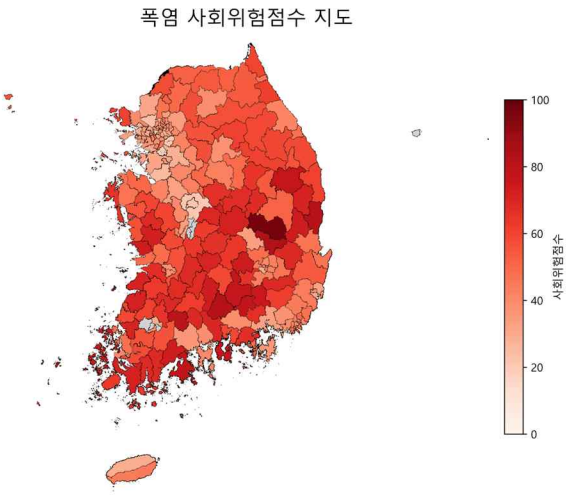


그림 4. 폭염 사회위험점수 지도

호우/침수 유형에서는 경상남도 산청군이 91.27점으로 가장 높게 나타났으며, 경상남도 합천군 83.78점, 전라남도 곡성군 78.84점, 경상북도 의성군 77.89점, 경상남도 의령군 77.70점 순으로 나타났다. 이는 동일한 지역이라도 재난유형에 따라 위험 수준이 다르게 나타날 수 있으며, 호우/침수 대응에서는 별도의 우선지역 선정이 필요함을 보여준다.

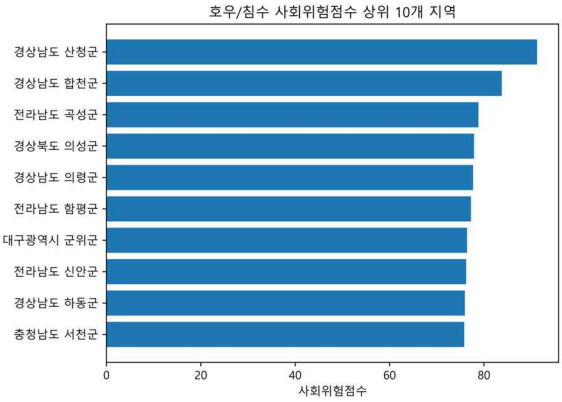


그림 5. 호우/침수 사회위험점수 상위 10개 지역

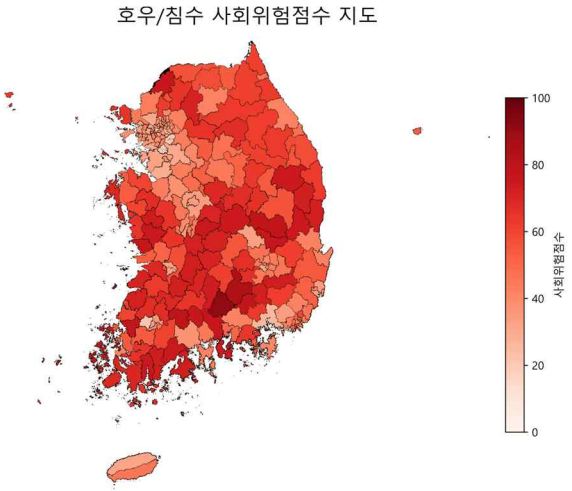


그림 6. 호우/침수 사회위험점수 지도

화재/산불 유형에서는 경상남도 산청군이 90.28점으로 가장 높게 나타났으며, 경상북도 의성군 87.15점, 경상남도 합천군 75.46점, 경상남도 남해군 75.27점, 경상남도 의령군 74.94점 순으로 나타났다.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

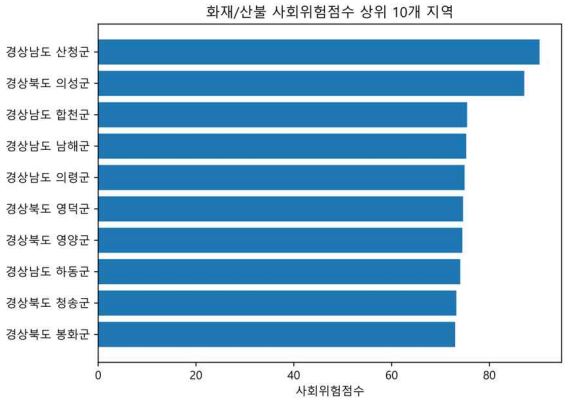


그림 7. 화재/산불 사회위험점수 상위 10개 지역

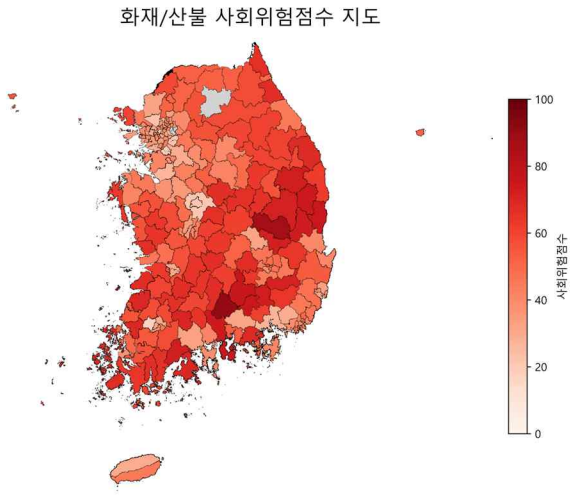


그림 8. 화재/산불 사회위험점수 지도

그림 3~8은 주요 재난유형별 사회위험점수 상위 지역과 공간적 분포를 나타낸다. 그림 3, 5, 7의 막대그래프는 폭염, 호우/침수, 화재/산불 유형별 사회위험점수 상위 10개 지역을 비교한 것이며, 그림 4, 6, 8의 지도는 각 재난유형에 대한 시군구별 사회위험점수의 공간적 분포를 나타낸 것이다. 각 지도는 0~100점의 동일한 점수 스케일을 적용하여 재난유형 간 비교가 가능하도록 하였으며, 색이 진할수록 해당 재난유형에 대한 사회위험점수가 높은 지역을 의미한다. 회색으로 표시된 지역은 해당 재난유형의 재난문자 발생 기록이 없거나 점수가 산출되지 않은 지역을 의미한다.

분석 결과, 재난유형별 사회위험점수 상위 지역은 일부 중복되기도 하였으나, 재난유형에 따라 서로 다른 양상도 나타났다. 특히 경상남도 산청군, 경상북도 의성군, 경상남도 합천군, 경상남도 의령군 등은 폭염, 호우/침수, 화재/산불 유형에서 반복적으로 상위권에 포함되었다. 이는 해당 지역들이 특정 재난유형에만 취약한 것이 아니라, 여러 재난유형에서 재난 신호와 지역 취약성이 함께 높게 나타나는 복합재난 취약지역일 가능성을 보여준다. 따라서 공공 사회위험 대응에서는 전체 재난문자 건수만을 기준으로 우선지역을 선정하기보다, 재난유형별 발생 신호와 지역 취약성을 함께 고려할 필요가 있다. 특히 폭염, 호우/침수, 화재/산불과 같이 발생 빈도와 대응 방식이 다른 재난유형은 각각 별도의 취약지역 분석과 대응 전략이 필요하다.

IV. 결론

본 연구는 긴급재난문자와 지역 취약지표를 결합하여 공공 사회위험 조기탐지 서비스 구조를 설계하고, 전국 시군구 단위의 재난유형별 사회위험점수를 산출하였다. 긴급재난문자는 폭염, 호우/침수, 화재/산불, 한파, 실종/배회 등 다양한 재난 및 안전 이슈를 포함하고 있어 지역별 위험 신호를 파악하는데 활용 가능한 공공 데이터이다. 본 연구에서는 2024년 1월 1일부터 2025년 12월 31일까지의 긴급재난문자를 분석 대상으로 설정하고, 문자 내용 기반 키워드 분류를 통해 재난유형을 구분하였다. 이후 재난문자 빈도를 전체 건수로만 집계하지 않고, 재난유형별로 분리한 뒤 시군구 단위로 재집계하여 지역별 재난 신호 지표를 산출하였다.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

분석 결과, 전체 재난유형별 긴급재난문자 빈도에서는 폭염, 호우/침수, 화재/산불 유형이 높은 비중을 차지하였다. 또한 재난유형별 재난문자 빈도와 고령인구비율, 독거노인가구비율, 응급의료기관 수를 결합한 결과, 재난유형에 따라 취약지역이 다르게 도출되었다. 폭염 유형에서는 경상북도 의성군, 대구광역시 군위군, 경상남도 산청군 등이 상위 지역으로 나타났으며, 호우/침수 유형에서는 경상남도 산청군, 경상남도 합천군, 전라남도 곡성군 등이 높은 점수를 보였다. 화재/산불 유형에서는 경상남도 산청군, 경상북도 의성군 등이 상위 지역으로 도출되었다. 이는 전체 재난문자 건수만을 기준으로 위험지역을 판단하기보다, 재난유형별 발생 신호와 지역 취약성을 함께 고려해야 함을 보여준다.

본 연구에서 제안하는 조기탐지 서비스의 목적은 특정 재난의 발생 시점을 정확히 예측하는 데 있지 않다. 긴급재난문자에서 반복적으로 나타나는 재난유형별 위험 신호와 지역 취약지표를 결합하여, 지자체가 사전에 대응 우선지역을 파악할 수 있도록 지원하는 데 목적이 있다. 예를 들어 폭염 사회위험점수가 높게 나타난 지역은 폭염 관련 재난문자 발생 빈도와 고령인구, 독거노인가구 등 지역 취약성이 함께 높은 지역으로 해석할 수 있다. 이러한 지역에서는 무더위쉼터 운영 점검, 독거노인 안부 확인, 방문 돌봄 강화, 냉방용품 지원, 폭염특보 시 맞춤형 재난문자 발송과 같은 대응 방안을 우선적으로 검토할 수 있다. 호우/침수 취약지역은 배수시설 점검, 침수취약도로 사전 통제, 저지대 거주민 대피 안내와 연결될 수 있으며, 화재/산불 취약지역은 산림 인접 지역 감시 강화, 소방 접근로 점검, 대피 경로 안내 등에 활용될 수 있다.

본 연구의 의의는 긴급재난문자를 단순한 알림 데이터가 아니라 지역별 재난 신호 데이터로 활용하고, 이를 지역 취약지표와 결합하여 재난유형별 대응 우선지역을 도출했다는 점에 있다. 또한 재난문자 수집, 전처리, 재난유형 분류, 지역 매핑, 취약지표 결합, 사회위험점수 산출, 시각화로 이어지는 데이터 파이프라인을 제시함으로써 공공데이터 기반 사회위험 분석의 활용 가능성을 확인하였다. 이는 지자체가 재난유형별 취약지역을 사전에 파악하고, 취약계층 보호와 재난 대응 자원 배분을 위한 기초자료로 활용될 수 있다.

다만 본 연구는 몇 가지 한계를 가진다. 첫째, 긴급재난문자는 실제 피해 규모를 직접 나타내는 자료가 아니므로, 재난문자 발생 빈도가 곧 실제 피해 수준을 의미한다고 해석하기는 어렵다. 둘째, 재난유형 분류가 키워드 기반 방식으로 수행되어 문맥에 따른 세부적인 의미를 충분히 반영하지 못할 수 있다. 셋째, 지역 취약지표가 고령인구비율, 독거노인가구비율, 응급의료기관 수로 제한되어 지역의 다양한 재난 대응 역량을 모두 반영하지는 못하였다. 향후 연구에서는 기상자료, 무더위쉼터, 침수흔적도, 생활인구, 소방·의료 접근성, 실제 재난 피해 자료 등을 추가로 결합하고, 자연어처리 기반 분류모델과 대시보드형 시각화를 적용하여 분석 정확도와 서비스 활용성을 높일 필요가 있다.

참고문헌

[1] 행정안전부, “긴급재난문자 데이터,” 공공데이터포털, 2026.

[2] 김민준, 이창규, 이상원, 강전영, 최진무, “인구감소와 고령화에 따른 노인 자연재난 위험지역 탐지 및 재난대응력 측정,” 대한공간정보학회지, 제31권, 제2호, pp. 25-34, 2023.

[3] 박순욱, 전혜윤, 김윤수, 이수원, “재난문자 분류를 위한 딥러닝 모델,” 한국컴퓨터정보학회논문지, 제26권, 제12호, pp. 1-9, 2021.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

[4] 주상임, 강현주, 오승희, “감성분석을 이용한 재난문자의 긴급성 평가,” 한국통신학회논문지, 제49권, 제2호, pp. 226-236, 2024.

[5] 김윤희, 김성희, “COVID-19 긴급재난문자 현황분석과 데이터 시각화를 통한 시사점,” 디지털콘텐츠학회논문지, 제22권, 제11호, pp. 1867-1875, 2021.

[6] 국립재난안전연구원, “안전취약계층 재난안전 대응 서비스 기술 개발,” 자체연구보고서, 2019.

[7] KOSIS, “고령인구비율,” 2024.

[8] KOSIS, “독거노인가구비율,” 2024.

[9] 공공데이터포털, “전국응급의료기관표준데이터,” 2026.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

공공데이터 기반 머신러닝을 활용한 전력 수요 예측 모델 비교 연구

지민석[O], 백수진[*]
[O]한신대학교 AI시스템반도체학과
[*]한신대학교 AI시스템반도체학과
e-mail: jms03050@hs.ac.kr[O], croso79@hs.ac.kr[*]

Comparative Study of Machine Learning-Based Power Demand Forecasting Models Using Public Data

Minseok Ji[O], Sujin Baek[*]
[O]Dept. of AI System Semiconductor Engineering, Hanshin University
[*]Dept. of AI System Semiconductor Engineering, Hanshin University

요 약

최근 전력 소비 증가와 탄소중립 정책 확산에 따라 안정적인 전력 수급과 효율적인 에너지 관리의 중요성이 커지고 있다. 특히 계절 변화와 생활 패턴 변화로 인해 전력 수요의 변동성이 증가하면서, 이를 더욱 정확하게 예측하기 위한 기술의 필요성이 높아지고 있다. 본 연구에서는 한국전력거래소의 전력수급실적 데이터와 기상청의 기상 데이터를 활용하여 일별 최대전력 수요를 예측하고, 머신러닝 모델별 성능을 비교 분석하였다. 연구에는 2024년부터 2025년까지 수집한 공공데이터를 활용하였으며, 데이터 전처리 및 feature engineering 수행 후 Linear Regression, Random Forest, XGBoost, LightGBM 모델을 적용하였다. 연구 결과, LightGBM 모델이 가장 우수한 예측 성능을 보였으며, 결정계수(R^2)는 약 0.825로 나타났다. 또한 feature importance 분석 결과, 평균기온과 상대습도가 전력 수요 예측에 중요한 영향을 미치는 변수로 확인되었다. 본 연구를 통해 공공데이터와 머신러닝 기반 전력 수요 예측의 활용 가능성을 확인하였으며, 향후 스마트 전력 운영 및 AI 기반 에너지 관리 시스템 연구에 활용될 수 있을 것으로 기대된다.

키워드: 전력 수요 예측, 머신러닝, LightGBM, Random Forest, 공공데이터

I. 서론

최근 AI 기술의 발전과 데이터센터 확산, 전기

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

차 보급 증가 등으로 인해 국내 전력 소비량이 지속적으로 증가하고 있다. 또한 탄소중립 정책과 에너지 효율 관리의 중요성이 강조되면서 안정적인 전력 수급 운영에 대한 관심이 높아지고 있다. 특히 계절 변화, 기온 변화, 생활 패턴 등에 따라 전력 수요가 크게 변동하기 때문에, 전력 공급의 안정성을 확보하기 위해서는 정확한 전력 수요 예측 기술이 필요하다.

기존의 전력 수요 예측 연구에서는 통계 기반 시계열 분석 기법이 주로 활용되었다. 그러나 최근에는 데이터의 규모와 복잡성이 증가하면서 비선형적인 패턴을 효과적으로 학습할 수 있는 머신러닝 기반 예측 기법이 활발하게 연구되고 있다. Random Forest, XGBoost, LightGBM과 같은 앙상블 기반 모델은 다양한 분야에서 우수한 예측 성능을 보이고 있으며, 전력 수요 예측 분야에서도 활용 연구가 증가하고 있다.

또한 공공데이터포털과 기상청 등에서 제공하는 공공데이터의 활용 범위가 확대되면서, 실제 환경 데이터를 기반으로 한 데이터 분석 및 예측 연구가 활발하게 진행되고 있다. 전력 수요는 기온, 습도, 강수량 등의 기상 요인과 밀접한 관련이 있기 때문에, 기상 데이터를 함께 활용할 경우 예측 정확도를 향상시킬 수 있다.

본 연구에서는 한국전력거래소의 전력수급실적 데이터와 기상청의 기상 데이터를 활용하여 일별 최대전력 수요를 예측하고자 한다. 이를 위해 날짜 정보와 기상 정보를 기반으로 feature engineering을 수행하였으며, Linear Regression, Random Forest, XGBoost, LightGBM 모델을 적용하여 성능을 비교 분석하였다. 또한 RMSE, MAE, R^2 지표를 활용하여 모델 성능을 평가하고, 특성 중요도(feature importance) 분석을 통해 전력 수요에 영향을 미치는 주요 변수를 분석하였다.

II. 관련 연구

2.1 전력 수요 예측 연구

전력 수요 예측은 안정적인 전력 공급과 에너지 관리 측면에서 중요한 연구 분야로 다루어지고 있다. 기존에는 통계 기반 시계열 분석 기법을 활용한 연구가 주로 이루어졌으나, 최근에는 비선형 데이터 학습이 가능한 머신러닝 기반 예측 기법의 활용이 증가하고 있다[1]. Random Forest와 XGBoost 같은 앙상블 모델은 복잡한 데이터 패턴을 효과적으로 학습할 수 있어 전력 수요 예측 분야에서도 활용되고 있다[2]. Random Forest는 여러 개의 결정 트리를 기반으로 학습하는 앙상블 모델로, 비선형 데이터에 대한 예측 성능이 우수한 특징을 가진다. 또한 과적합 문제를 완화할 수 있어 다양한 데이터 분석 분야에서 활용되고 있다.

최근에는 학습 속도와 성능 측면에서 장점을 가지는 LightGBM 모델을 활용한 연구도 증가하고 있다[3]. XGBoost와 LightGBM은 Gradient Boosting 기반 모델로, 이전 모델의 오차를 보완하는 방식으로 학습이 진행된다[4]. 특히 LightGBM은 leaf-wise 방식의 트리 구조를 사용하여 빠른 학습 속도와 높은 예측 성능을 제공하는 특징이 있다[5].

따라서 본 연구에서는 전력 수요 예측 성능 비교를 위해 Linear Regression, Random Forest, XGBoost, LightGBM 모델을 활용하였다. Linear Regression은 입력 변수와 출력 변수 사이의 선형 관계를 기반으로 하는 기본적인 회귀 모델로, 다른 모델과의 성능 비교를 위해 활용되었다.

2.2 공공데이터 활용 연구

최근 공공데이터포털과 기상청 등에서 제공하는 데이터를 활용한 데이터 분석 연구가

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

활발하게 이루어지고 있다. 전력 수요는 기온, 습도, 강수량 등의 기상 요인과 밀접한 관련이 있기 때문에, 기상 데이터를 함께 활용한 예측 연구가 다양한 분야에서 진행되고 있다[6]. 특히 공공데이터를 활용할 경우 실제 환경 데이터를 기반으로 분석이 가능하다는 장점이 있으며, 에너지 관리 및 스마트 전력 운영 분야에서도 활용 가능성이 높다.

기존 연구에서는 특정 모델 중심의 분석이 주로 이루어졌으나, 다양한 머신러닝 모델 간 성능 비교와 공공데이터 기반 분석에 대한 연구는 상대적으로 부족하였다. 본 연구에서는 공공데이터 기반 전력 데이터와 기상 데이터를 활용하여 다양한 머신러닝 모델의 성능을 비교 분석하였다.

III. 제안방법

3.1 데이터 수집

본 연구에서는 한국전력거래소에서 제공하는 전력수급실적 데이터와 기상청의 기상 데이터를 활용하였다. 전력 데이터는 일별 최대전력 정보를 포함하고 있으며, 기상 데이터는 평균기온, 상대습도, 강수량 등의 정보를 포함하고 있다. 연구에는 2024년 1월부터 2025년 12월까지 수집한 데이터를 사용하였으며, 날짜 정보를 기준으로 전력 데이터와 기상 데이터를 결합하였다. 데이터는 공휴일 및 계절 변화에 따른 전력 사용 패턴을 반영할 수 있도록 연 단위 기간을 포함하여 수집하였다. 전력 수요는 계절 변화와 기상 환경에 큰 영향을 받기 때문에, 전력 데이터와 기상 데이터를 함께 활용하여 예측 성능을 향상시키고자 하였다.

3.2 데이터 전처리

수집한 데이터에 대해 데이터 형식 변환 및 날짜 데이터 처리를 수행하였다. 전력 데이터와 기상 데이터는 서로 다른 형식으로 제공되었기 때문에, 동일한 날짜 기준으로 데이터를 통합하는 과정을 수행하였다. 날짜 데이터는 datetime 형식으로 변환한 후 분석에 활용하였으며, 데이터 결합 과정에서 날짜 기준 정렬을 수행하였다.

3.3 Feature Engineering

전력 수요 예측 성능 향상을 위해 날짜 정보를 기반으로 추가 변수를 생성하였다. 날짜 데이터로부터 월(month)과 요일(dayofweek) 정보를 추출하여 입력 변수로 활용하였다. 월 정보를 계절별 전력 사용 패턴을 반영하기 위해 사용하였으며, 요일 정보는 주중과 주말의 전력 사용량 차이를 반영하기 위해 활용하였다.

또한 평균기온, 상대습도, 강수량 데이터를 함께 사용하여 기상 환경이 전력 수요에 미치는 영향을 분석하고자 하였다. 최종적으로 월, 요일, 평균기온, 상대습도, 강수량 데이터를 입력 변수로 사용하였으며, 일별 최대전력을 예측 대상으로 설정하였다.

표 1. 입력 변수 구성

변수명	설명
month	월 정보
dayofweek	요일 정보
temperature	평균기온
humidity	상대습도
rainfall	강수량
max_power	일별 최대전력

3.4 머신러닝 모델

본 연구에서는 선형 모델과 앙상블 기반 모델 간 성능 차이를 비교하기 위해 Linear Regression, Random Forest, XGBoost, LightGBM 모델을 선정하였다. 또한 최근 다양한 데이터 분석 분야에서 활용되고 있는 Gradient Boosting 기반 모델의 전력

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

수요 예측 성능을 비교 분석하고자 하였다.

전체 데이터는 학습 데이터와 테스트 데이터로 분리하여 모델 학습 및 성능 평가를 수행하였다. 학습 데이터와 테스트 데이터의 비율은 8:2로 설정하였다.

3.5 성능 평가 지표

모델 성능 평가는 RMSE(Root Mean Squared Error), MAE(Mean Absolute Error), R^2(결정계수) 지표를 활용하였다. RMSE는 실제값과 예측값 차이의 제곱 평균 제곱근을 의미하며, 값이 작을수록 높은 예측 성능을 의미한다. MAE는 실제값과 예측값 차이의 절대값 평균으로 계산되며, 오차 크기를 직관적으로 확인할 수 있는 장점이 있다. R^2는 모델의 설명력을 나타내는 지표로, 1에 가까울수록 높은 예측 성능을 의미한다. 본 연구에서는 각 모델의 RMSE, MAE 및 R^2값을 비교하여 전력 수요 예측 성능을 분석하였다.

IV. 실험 결과 및 분석

4.1 월별 최대전력 변화 분석

월별 평균 최대전력 변화는 그림 1과 같다. 그래프를 통해 계절에 따라 전력 사용량이 변화하는 경향을 확인할 수 있었다. 특히 여름철과 겨울철 구간에서 평균 전력이 상대적으로 높게 나타났으며, 이는 냉방기기와 난방기기 사용량 증가의 영향을 받은 것으로 보인다.

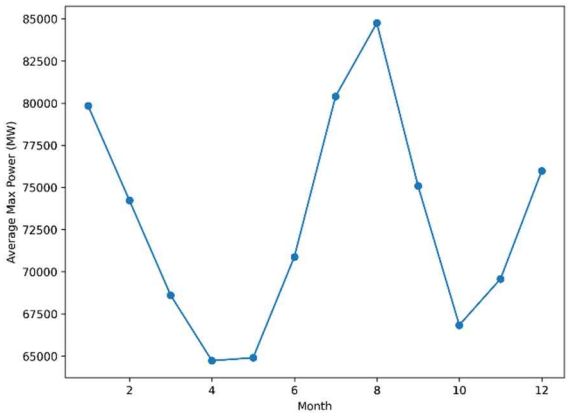


그림 1. 월별 평균 최대전력 변화

반면 봄과 가을 구간에서는 비교적 낮은 전력 사용량이 나타났다. 이를 통해 계절 변화와 기상 환경이 실제 전력 수요에 영향을 미치는 것을 확인할 수 있었으며, 기상 데이터를 활용한 전력 수요 예측의 필요성을 확인할 수 있었다.

4.2 모델별 성능 비교

본 연구에서는 Linear Regression, Random Forest, XGBoost, LightGBM 모델을 활용하여 전력 수요 예측 성능을 비교하였다. 모델 성능 평가는 RMSE, MAE, R^2 지표를 기반으로 수행하였으며, 각 모델의 성능 비교 결과는 표 2와 같다.

표 2. 머신러닝 모델별 성능 비교

Model	RMSE	MAE	R^2
Linear Regression	7733	6395	0.259
Random Forest	3818	2844	0.819
XGBoost	3958	2961	0.806
LightGBM	3754	2904	0.825

실험 결과, LightGBM 모델이 가장 우수한 예측 성능을 나타냈으며, 결정계수(R^2)는 약 0.825로 확인되었다. Random Forest와 XGBoost 모델 또한 비교적 높은 성능을 보였으나, Linear Regression은

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

상대적으로 낮은 성능을 나타냈다. 이는 전력 수요 데이터가 단순 선형 관계보다 비선형적인 패턴을 포함하고 있기 때문으로 분석된다.

특히 Random Forest, XGBoost, LightGBM과 같은 앙상블 기반 모델은 기상 데이터와 날짜 정보 사이의 복합적인 관계를 효과적으로 학습할 수 있어 높은 예측 성능을 보인 것으로 판단된다. 반면 Linear Regression은 변수 간 선형 관계만을 기반으로 학습하기 때문에 복잡한 전력 사용 패턴을 충분히 반영하지 못한 것으로 보인다.

4.3 실제값과 예측값 비교

LightGBM 모델의 실제값과 예측값 비교 결과는 그림 2와 같다. 대부분의 구간에서 예측값이 실제 전력 사용량과 유사한 패턴을 나타내는 것을 확인할 수 있었다. 일부 구간에서는 오차가 발생하였으나, 전체적으로 실제값의 변화를 안정적으로 추종하는 경향을 보였다.

특히 전력 사용량이 급격하게 증가하거나 감소하는 구간에서도 비교적 유사한 형태의 예측 결과를 나타냈으며, 이를 통해 LightGBM 모델이 전력 수요 패턴을 효과적으로 학습한 것으로 판단된다.

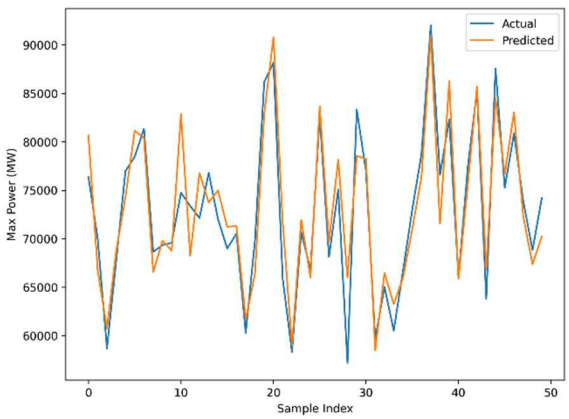


그림 2. LightGBM 모델의 실제값과 예측값 비교

4.4 Feature Importance 분석

LightGBM 모델의 feature importance 분석 결과는 그림 3과 같이 평균기온(temperature)과 상대습도(humidity)가 전력 수요 예측에 중요한 영향을 미치는 변수로 나타났다. 이는 여름철 냉방기기 사용 증가와 같은 계절적 요인이 실제 전력 소비량에 영향을 미치기 때문으로 분석된다.

또한 월(month) 정보 역시 일정 수준의 중요도를 나타냈으며, 계절 변화에 따른 전력 사용 패턴이 반영된 결과로 볼 수 있다. 반면 강수량(rainfall)의 중요도는 상대적으로 낮게 나타났으며, 전력 수요와의 직접적인 관계는 비교적 작은 것으로 확인되었다.

이를 통해 기상 데이터와 날짜 정보가 전력 수요 예측에 유의미한 변수로 활용될 수 있음을 확인할 수 있었다.

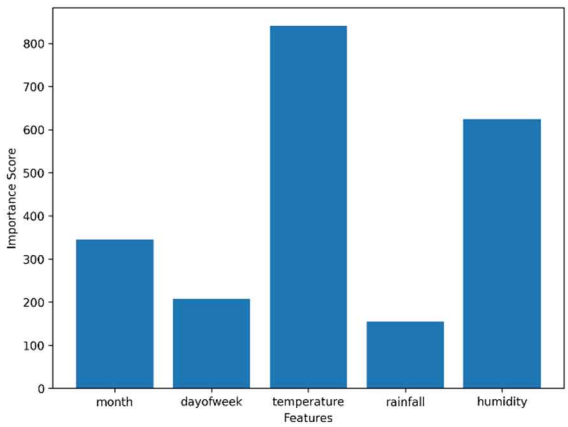


그림 3. LightGBM 모델의 특성 중요도 분석

V. 결론

본 연구에서는 한국전력거래소의 전력수급실적 데이터와 기상청의 기상 데이터를 활용하여 일별 최대전력 수요를 예측하고, 머신러닝 모델별 성능을 비교 분석하였다. 연구에는 2024년 1월부터 2025년 12월까지의 공공데이터를 활용하였으며, 날짜 정보와 기상 정보를 기반으로 feature engineering을 수행하였다.

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

전력 수요 예측을 위해 Linear Regression, Random Forest, XGBoost, LightGBM 모델을 적용하였으며, RMSE, MAE, R^2 지표를 활용하여 성능을 평가하였다. 실험 결과, LightGBM 모델이 가장 우수한 예측 성능을 나타냈으며, Random Forest와 XGBoost 모델 또한 비교적 높은 성능을 보였다. 반면 Linear Regression은 비선형적인 전력 사용 패턴을 충분히 반영하지 못해 상대적으로 낮은 성능을 보였다.

또한 feature importance 분석 결과, 평균기온과 상대습도가 전력 수요 예측에 중요한 영향을 미치는 변수로 확인되었다. 이는 계절 변화와 냉난방 사용량 증가가 실제 전력 소비에 영향을 미치기 때문으로 분석된다.

본 연구를 통해 공공데이터와 머신러닝 기반 전력 수요 예측의 활용 가능성을 확인할 수 있었으며, 향후 스마트 전력 운영 및 AI 기반 에너지 관리 시스템 연구에 활용될 수 있을 것으로 기대된다. 또한 본 연구 결과는 향후 전력 수급 계획 수립 및 에너지 사용량 관리 시스템 구축 과정에서 기초 자료로 활용될 수 있을 것으로 기대된다.

다만 본 연구에서는 2024년부터 2025년까지의 일별 공공데이터를 활용하였기 때문에 데이터 기간과 변수 구성에 한계가 존재한다. 또한 시간 단위 전력 데이터와 산업별 전력 사용 데이터 등을 함께 활용할 경우, 보다 세밀한 전력 수요 예측이 가능할 것으로 예상된다. 향후 연구에서는 더 다양한 기상 변수와 시간 단위 데이터를 활용하여 예측 성능을 향상시키는 추가 연구가 필요할 것으로 생각된다.

참고문헌

[1] 문형빈, 안재균, 이철용, "머신러닝을 이용한 제주도 전력수요 예측," 한국혁신학회지, 제18권, 제4호, pp. 49-72, November 2023.

[2] 김도준, "머신러닝 기반 앙상블 모델 비교 및 소프트 보팅을 이용한 전력 수요 예측," Proceedings of KIIT Conference, pp.1571-1574, November 2024.

[3] 윤성연, 오수민, 김지연, 박민서, "머신러닝 기반 전기 에너지 소비량 예측," 한국지능정보시스템학회, 제51권, 제2호, pp. 171-181, April 2025.

[4] Tianqi Chen and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.785-794, 2016.

[5] Guolin Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," Advances in Neural Information Processing Systems (NeurIPS), Vol.30, 2017.

[6] 차현종, 강아름, "기상정보를 활용한 머신러닝 기반의 전력수요 예측 모델," 한국콘텐츠학회 논문지, 제23권, 제2호, pp. 117-124, February 2023.

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

AI Agent 환경 내 능동형 접근제어 모델

홍현준* (고려대학교), 최부영ᵒ (한신대학교), 고병수 (한신대학교), 이세영 (서강대학교)
e-mail: hongkd9988@korea.ac.kr, ra020509@hs.ac.kr

Active Access Control Model in AI Agent Environment

Hyeonjun Hong* (Korea Univ.), Buyeong Choiᵒ (Hanshin Univ.), Byeongsu Ko (Hanshin Univ.), Seyeong Lee (Sogang Univ.)
(주저자, ᵒ 발표자)*

요 약 대규모 언어 모델(LLM) 기반 AI 에이전트가 빠르게 확산되면서, 기존 접근제어 모델이 가정해 온 결정론적이고 사람 중심의 정적(靜的) 권한 구조로는 대응하기 어려운 새로운 보안 위협이 나타나고 있다. AI 에이전트는 동일한 입력에도 다른 행동을 보일 수 있고(비결정성), 사람의 개입 없이 스스로 외부 도구와 API를 호출하며, 다수의 에이전트가 서로 권한을 위임한다. 본 논문은 이러한 환경에 특화된 능동형 접근제어 모델 ACAM(Access Control AI Model)을 제안한다. ACAM은 신원 중심 인증, 최소 권한의 동적 부여, 행동 기반 신뢰라는 세 가지 설계 원칙 위에서 신원·인증, 데이터 접근, 모델 API, 에이전트 오케스트레이션, 거버넌스·감사의 5계층 아키텍처와 실시간 위험 점수 엔진(RSE)을 결합한다. 이산 사건 시뮬레이션으로 8가지 AI 고유 위협 시나리오를 평가한 결과, ACAM은 평균 89.5%의 탐지율을 기록하여 기존 RBAC 대비 73.1%p 향상되었고, 에이전트 권한 상승 공격의 91%를 차단하였으며, 평균 대응 시간(MTTC) 1초 미만을 달성하였다. 또한 NIST AI RMF, OWASP LLM Top 10, EU AI Act 등 주요 규제 프레임워크와의 정렬을 확인하여 규제 산업에서의 적용 가능성을 제시한다.

키워드 : AI 에이전트 보안, 능동형 접근제어, 위험 점수 엔진, 멀티에이전트 시스템, 제로 트러스트

I. 서 론

대규모 언어 모델(LLM)을 기반으로 하는 AI 에이전트는 단순한 질의응답을 넘어, 스스로 목표를 세우고 외부 도구를 호출하며 다른 에이전트와 협력해 복잡한 작업을 수행하는 단계로 발전하고 있다. 이러한 자율성은 생산성을 크게 높이지만, 동시에 기존 보안 체계가 전제해 온 가정을 무너뜨린다.

전통적인 접근제어 모델인 RBAC(Role-Based Access Control)와 ABAC(Attribute-Based Access Control)는 (1) 행위 주체가 사람이고, (2) 권한이 사전에 정의된 역할이나 속성에 따라 정적으로 부여되며, (3) 동일한 조건에서 동일한 행동이 반복된다는 결정론적 전제 위에 설계되었다[1]. 그러나 AI 에이전트 환경에서는 이러한 전제가 모두 성립하지 않는다. 에이전트는 동일한 입력에도 확률적으로 다른 행동을 보이고(비결정성), 사람의 개입 없이 실행 시점에 도구와 API를 동적으로 선택하며, 멀티에이전트 구조에서는 한 에이전트가 다른 에이전트에게 권한을 위임한다. 이로 인해 프롬프트 인젝션, 에이전트 권한 상승, 위임 사슬을 통한 권한 오남용과 같은 새로운 공격 표면이 발생한다[4].

본 논문은 이러한 문제를 해결하기 위해 AI 에이전트 환경에 특화된 능동형 접근제어 모델 ACAM을 제안한다. 본 논문의 기여는 다음과 같다. 첫째, AI 에이전트의 고유 특성을 반영한 세 가지 설계 원칙을 정의한다. 둘째, 이를 구현하는 5계층 아키텍처와 실시간 위험 점수 엔진(RSE)을 설계한다. 셋째, 이산 사건 시뮬레이션을 통해 8가지 위협 시나리오에서 제안 모델의 효과를 정량적으로 검증한다.

II. 관련 연구

2.1 기존 접근제어 모델의 한계

RBAC는 사용자에게 역할을 부여하고 역할에 권한을 묶어 관리 복잡도를 낮춘 대표적인 접근제어 모델이다[1]. ABAC는 사용자·자원·환경의 속성을 조합해 보다 세밀한 정책을 표현한다. 그러나 두 모델 모두 권한이 사전에 정의되어 정적으로 유지되므로, 실행 중 행위 주체의 행동 변화를 반영하지 못한다. 또한 행위 주체를 사람으로 가정하기 때문에, 신원이 명확하지 않고 행동이 가변적인 AI 에이전트에는 그대로 적용하기 어렵다.

2.2 AI 에이전트 보안 동향

제로 트러스트 아키텍처[2]는 "절대 신뢰하지 않고 항상 검증한다"는 원칙에 따라 모든 접근 요청을 지속적으로 검증한다. NIST AI RMF[3], OWASP LLM Top 10[4], EU AI Act[5] 등은 AI 시스템의 위험을 식별·관리하기 위한 지침과 규제를 제시한다. 다만 이들은 주로 거버넌스 차원의 원칙과 점검 항목을 제공할 뿐, 실행 시점에 에이전트의 행동을 평가하고 권한을 동적으로 제어하는 구체적인 실행 모델은 충분히 다루지 않는다. 본 논문은 이러한 공백을 메우는 실행 계층의 접근제어 모델을 제안한다.

III. 제안 방법 : ACAM

3.1 설계 원칙

ACAM은 다음 세 가지 원칙 위에서 설계된다.

- 신원 중심 인증(Identity-First) : 모든 에이전트는 검증 가능한 고유 신원을 가지며, 모든 접근 요청은 이 신원을 기반으로 인증된다.
- 최소 권한의 동적 부여(Least Privilege Dynamism) : 권한은 사전에 고정되지 않고, 작업 단위로 필요한 범위만큼 한시적으로 부여된 뒤 회수된다.
- 행동 기반 신뢰(Behavioral Trust) : 신뢰는 고정된 역할이 아니라 실행 중 관찰된 행동에서 도출되며, 비정상 행동이 감지되면 즉시 하향 조정된다.

3.2 5계층 아키텍처

ACAM은 표 1과 같이 5개 계층으로 구성된다. 각 계층은 독립적으로 정책을 적용하면서도 실시간 위험 점수 엔진(RSE)을 통해 유기적으로 연동된다. 하위 계층에서 수집된 행동 신호는 RSE로 전달되어 위험 점수로 환산되고, 그 결과는 상위 계층의 권한 결정에 즉시 반영된다.

표 1. ACAM의 5계층 아키텍처

계층	주요 역할
신원·인증	에이전트 신원 발급·검증 및 요청 인증
데이터 접근	데이터 단위 접근 정책 및 민감도 통제
모델 API	모델 호출 범위·빈도 제한, 프롬프트 검증
오케스트레이션	멀티에이전트 권한 위임·협력 통제
거버넌스·감사	정책 관리, 로그 기록 및 규제 정렬

3.3 실시간 위험 점수 엔진(RSE)

RSE는 ACAM의 핵심 구성 요소로, 에이전트의 행동을 실시간으로 관찰해 위험 점수를 산출한다. 요청 빈도, 권한 범위의 변화, 평소 행동 패턴과의 편차, 위임 사슬의 깊이 등 여러 신호를 가중 합산하여 0에서 1 사이의 위험 점수를 계산한다. 점수가 임계값을 초과하면 요청을 차단하거나 추가 검증을 요구하고, 정상 범위이면 권한을 허용한다. 이러한 판단은 1초 미만에 이루어져 자율적으로 동작하는 에이전트의 처리 속도를 저해하지 않는다.

IV. 실험 및 결론

제안 모델의 효과를 검증하기 위해 이산 사건 시뮬레이션 환경에서 프롬프트 인젝션, 에이전트 권한 상승, 위임 사슬 오남용 등 8가지 AI 고유 위협 시나리오를 구성하여 평가하였다. 표 2와 같이 ACAM은 평균 89.5%의 탐지율을 기록하여 기존 RBAC(약 16.4%) 대비 73.1%p 향상되었으며, 에이전트 권한 상승 공격의 91%를 차단하였다. 또한 평균 대응 시간(MTTC)은 1초 미만으로 위협 탐지부터 대응까지의 지연을 최소화하였다.

표 2. ACAM과 RBAC의 성능 비교

지표	RBAC	ACAM
평균 탐지율	16.4%	**89.5%**
권한 상승 차단율	–	**91%**
대응 시간(MTTC)	–	**< 1초**

본 논문은 AI 에이전트 환경의 고유한 보안 위협에 대응하기 위해 신원 중심 인증, 최소 권한의 동적 부여, 행동 기반 신뢰의 원칙을 갖춘 능동형 접근제어 모델 ACAM을 제안하고, 시뮬레이션을 통해 그 효과를 확인하였다. ACAM은 NIST AI RMF, OWASP LLM Top 10, EU AI Act 등 주요 프레임워크와 정렬되어 규제 산업에서의 실용적 적용 가능성을 보인다. 향후 연구에서는 실제 멀티에이전트 운영 환경에 ACAM을 적용하여 위험 점수 모델을 정교화하고, 다양한 공격 유형에 대한 일반화 성능을 추가로 검증할 계획이다.

참고문헌

[1] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, “Role-Based Access Control Models,” IEEE Computer, vol. 29, no. 2, pp. 38-47, Feb. 1996.

[2] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, “Zero Trust Architecture,” NIST Special Publication 800-207, Aug. 2020.

[3] National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” NIST AI 100-1, Jan. 2023.

[4] OWASP, “OWASP Top 10 for LLM Applications 2025,” OWASP GenAI Security Project, 2025.

[5] European Parliament and Council, “Regulation (EU) 2024/1689 (Artificial Intelligence Act),” Official Journal of the EU, OJ L, 2024/1689, 12 Jul. 2024.

※ *본 연구는 [지원기관·과제명]의 지원을 받아 수행되었습니다. (해당 시 기재)*

지역사회 시니어 체육 프로그램 참여와 주관적 건강 및 생활 만족도의 인과관계 분석

노유경, 김유신
한신대학교 빅데이터융합학과
e-mail: yknoh1002@hs.ac.kr, yoosin25@hs.ac.kr

A Study on the Causal Relationship between Senior Physical Education Program and Health & Life Satisfaction

You Kyong Noh, Yoosin Kim
Dept. of Big Data Convergence, Hanshin University

요 약

고령화 시대를 맞아 시니어의 체육 활동 참여가 증가하고 있으나, 기존 단순 설문조사 방식으로는 운동과 건강 및 삶의 만족도 간의 명확한 인과관계를 규명하기 어렵다. 본 연구는 시니어 체육 프로그램 참여자들의 운동 빈도, 주관적 건강, 생활 만족도 변화를 장기간 추적하는 시간에 따른 데이터를 구축하고 AI 기술을 적용하여 실질적인 원인-결과 구조를 밝히고자 한다. 시간의 흐름에 따른 표본의 분산과 모평균을 기초적으로 추정하고, 이를 기반으로 도출된 인과관계 분석 결과를 시각화 하여 맞춤형 특수체육 및 노인 복지 정책 수립을 위한 과학적 근거를 제시하는 것을 목적으로 한다.

키워드: 시니어 체육 프로그램, 인과관계, 인공지능, 삶의 질

I. 서론

고령화 시대를 맞아 시니어의 지역사회 체육 프로그램 참여가 꾸준히 증가하고 있다. 능동적인 체육 활동은 노년기 삶의 질을 높이는 핵심 메커니즘이지만, 기존의 일회성 설문조사 방식으로는 운동과 건강 간의 명확한 인과관계를 증명하기 어렵다. 이에 본 연구는 장기간 수집된 시계열 데이터에 AI 알고리즘을 적용하여, 시니어 체육 활동이 건강 및 삶의 질에 미치는 실질적인 효과를 과학적으로 입증하고자 한다.

II. 관련 연구

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

2.1 국내 동향

선행 연구들은 노인의 능동적 여가 및 체육 활동이 삶의 의미체계를 구성하는 핵심 독립변인임을 보여준다. 또한 기존의 종단 연구와 패널 분석 결과들은 단편적인 조사를 넘어선 시간 흐름 데이터의 필요성을 강력히 지지하고 있다. 이는 시니어 체육 프로그램 참여와 삶의 질 변화 추이를 시계열적으로 추적하고자 하는 본 연구 설계의 명확한 학술적 근거가 된다

표1. 데이터 분석 모델에 따른 변수 유형

구분	독립변인	종속변인
패널 데이터 연구	노인 여가스포츠 활동빈도	일상생활 수행능력, 건강상태
시계열 추적 연구	시간에 따른 주관적 건강 상태	삶의 만족도 변화 추이
여가 역할 연구	능동적 체육·여가 활동	심리적 혜택, 궁극적 삶의 질

III. 제안방법

본 연구는 시니어 체육 프로그램의 효과를 증명하기 위해 3단계 방법론을 제안한다. 첫째, 참여자의 운동 빈도, 선호 종목, 건강 상태 및 만족도 변화를 장기간 수집해 시계열 데이터를 구축한다. 둘째, 수집된 데이터의 분산 등 기초 통계를 바탕으로 AI 인과 추론 알고리즘을 적용해 운동 참여와 삶의 질 간의 실질적 인과관계를 정밀하게 구조화한다. 셋째, 도출된 복잡한 분석 결과를 누구나 직관적으로 파악할 수 있도록 그래프와 차트 등으로 시각화한다

IV. 결론

본 연구에서 제안하는 AI 기반 시계열 데이터 분석 모델은 시니어 대상 체육 프로그램이 삶의 질 향상에 기여하는 바를 과학적이고 가시적으로 증명하는 데 크게 기여할 것이다. 나아가 이러한 실증적 분석 결과는 노년층의 신체적, 심리적 특성을 반영한 맞춤형 노인 복지 정책을 수립하고, 지역사회 내 능동적 체육 활동 참여를 장려하기 위한 핵심적인 근거 자료로 활용될 것으로 기대된다.

참고문헌

[1] 나항진, "삶의 질 향상을 위한 노인 여가의 역할에 관한 연구", Vol. 24, No. 1, pp. 53-70, 2004.

[2] 김지연, 홍준희, "노인의 생활체육 참여 정도가 주관적 건강, 노화태도, 삶 만족에 미치는 영향: 변인 간 관계를 중심으로", 한국스포츠심리학회지, Vol. 28, No. 2, pp. 11-23, 2017.

[3] [4] 허준수, 조승호, "노인들의 삶의 만족도에 대한 인과모형 연구: 유형별 사회참여활동의 매개효과를 중심으로", 한국콘텐츠학회논문지, Vol. 17, No. 1, pp. 88-101, 2017.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

프로야구 선수 평가 지표의 통계적 착시 현상 분석: 상대전적과 클러치 스탯을 중심으로

유준호*, 김유신
한신대학교 빅데이터융합학과
e-mail: dbwnsgh0306@naver.com, yoosin25@hs.ac.kr

Illusion of Small Sample Data in KBO: Based in Baseball Batter vs. Pitcher and Clutch Stats

Author*, Junho Yu, Yoosin Kim
Dept. of Big Data Convergence, Hanshin University

요 약

야구 중계 및 언론에서는 '상대전적'이나 '득점권 타율' 같은 상황별 지표가 선수를 평가하는 주요 근거로 활용되나, 이는 표본이 극도로 작은 '스몰 샘플'로 통계적 유의성이 부족하다. 야구는 인간이 수행하는 스포츠로 당일 컨디션 등 외부 노이즈가 개입되는데, 대수의 법칙이 작용하지 않는 스몰 샘플에서는 이러한 우연성이 결과를 왜곡하여 통계적 착시를 일으킨다. 본 연구는 데이터 분석 도구를 활용하여 KBO 및 MLB 데이터를 분석하고, 상황별 지표의 연도별 상관관계를 통해 통계적 일관성을 검증한다. 나아가 단순히 안타 여부라는 '결과'만 따지는 방식에서 벗어나 타구 속도, 발사각 등 역학적 '과정' 데이터를 결합한 인과관계 기반의 검증 방법론을 제안한다. 이를 통해 전통적 지표의 무비판적 소비를 비판하고, 데이터 사이언스 관점에서의 올바른 기록 해석 방향을 제시하고자 한다.

키워드: 프로야구데이터, 선수성적평가, 상대전적, 클러치스탯,

I. 서론

최근 프로야구 중계나 언론 보도를 살펴보면, 선수의 능력을 평가하거나 경기 결과를 예측하기 위해 '특정 투수 상대 타율(상대전적)'이나 '득점권 타율(클러치 스탯)'과 같은 상황별 세부 지표가 빈번하게 인용된다. 이러한 지표들은 경기의 서사를 극적으로 전달하는 데 효과적이기 때문에 대중적으로 널리 소비되고 있다.

그러나 데이터 사이언스와 통계학적 관점에서 볼 때, 현재 야구 현장에서 활용되는 상황별 지표들은 표본의 크기가 지나치게 작다는 치명적인 한계를 지닌다. 야구는 인간이 수행하는 스포츠이므로 당일의 생체 리듬, 심리적 압박감 등 통제 불가능한 외부 변수가 타석의 결과에 필연적으로 개입된다. 표본이 충분히 누적되는 정규 시즌 전체 데이터의 경우, 대수의 법칙에 의해 이러한 외부 노이즈들이 상쇄되며 선수의 고유한 기량 값으로 수렴 및 안정화된다. 하지만 스몰 샘플 환경에서는 단 한두 번의 컨디션 저하나 우연한 행운이 전체 비율 지표를 심각하게 왜곡하여 '통계적 착시'를

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

일으킨다.

II. 관련 연구

2.1 상황별 지표의 우연 변동성(Random Variance)과 클러치 능력의 허구성 (Law, 2017) 데이터 분석가 키스 로(Keith Law)는 저서 『스마트 베이스볼』을 통해 전통적인 상황별 스탯이 극도로 적은 표본의 문제를 넘어, 근본적으로 '우연 변동'에 의해 왜곡된 수치에 불과함을 지적했다. 그는 이른바 '클러치 히터'로 불리는 특정 타자들의 실제 데이터를 추적하여 이들의 득점권 성적이 일관되지 않음을 실증적으로 보여주었다. 특히 해당 저서는 Tango 등(2007)의 통계적 연구인 『The Book』을 인용하며, 타격 스탯 내에 클러치 효과를 입증할 변수를 적용하더라도 유의미한 특별한 능력이 발견되지 않았음을 강조한다. 결론적으로 평상시와 클러치 상황에서 특정 선수가 보여줄 것이라 기대되는 타격 성적은 동일하며, 위기 상황에서 기량이 상승한다는 믿음은 통계적 허상이라는 것이다.

2.2 맞대결 데이터의 차원 축소 및 예측 한계 연구 (O'Connell, 2024) O'Connell(2024)은 특정 타자와 투수의 맞대결 타율을 예측하기 위해 '일반화된 연결 행렬 분해(GLMF)' 기법을 활용한 연구를 진행했다. 이 연구는 투타 맞대결 데이터의 극단적인 희소성(Sparsity) 때문에 과거 상대전적 결과만으로는 미래 성적을 예측할 수 없음을 지적했다.

2.3 타격 지표의 우연 변동성과 연도별 재현성 부족에 관한 통계적 연구 (Albert, 2008) 통계학자 Jim Albert(2008)는 스포츠 통계 학술지를 통해 타자들의 '몰아치기(Streaky)'나 '클러치 타격' 성적이 통계적 유의성을 가지는지 검증했다. 연구에 따르면 한 시즌 동안 관찰되는 타자들의 극단적인 단기 성적은 수많은 사람이 동시에 동전을 던질 때(toss coins) 우연히 앞면이 연속으로 나오는 현상과 같은 다중성 문제로 설명된다. 특히 선수의 타격 능력이 상황에 관계없이 일정하다고 가정하는 단순 모델만으로도 한 시즌 내에 발생하는 몰아치기 변동성의 대부분이 수학적으로 설명됨을 입증했다. 나아가, 특정 연도에 극단적인 성적을 기록한 선수가 여러 시즌에 걸쳐(across seasons) 그 특성을 일관되게 보여주는 사례는 찾기 어렵다고 결론지었다.

III. 제안방법

본 연구는 선행 연구들의 통계적 비판을 바탕으로, KBO 및 MLB 데이터를 활용하여 전통 스탯의 한계를 실증하고 이를 보완할 수 있는 '인과관계 기반 검증 파이프라인'을 제안한다. 접근 도구로는 분석의 범용성과 재현 가능성을 확보하기 위해 널리 활용되는 통계 데이터 분석 소프트웨어를 활용한다.

첫째, 베이스볼 서번트(Baseball Savant)를 통해 다년간의 투타 매치업 데이터와 상황별 스탯을 정형 데이터 형태로 수집하여 분석 도구에 로드한다. 이와 함께 타구 속도(Exit Velocity), 선구안(Chase Rate), 투구 무브먼트 등 타석의 '과정(Process)'을 설명하는 트래킹 데이터를 결합한다.

둘째, 데이터 분석 도구의 필터링 및 그룹화 기능을 활용하여 데이터를 논리적 인과관계 유무에 따라 분류한다. 상대전적은 '투수의 구종별 강점과 타자의 코스별 약점이 일치하는 인과관계 그룹'과 '단순 무작위 그룹'으로 나누며, 클러치 스탯은 '심리적 압박으로 인해 선구안 등 과정 메커니즘이 실제로 붕괴된 그룹'과 '과정은 동일하나 운이 없어 결과만 나쁜 그룹'으로 분할한다.

셋째, 분석 도구의 통계 연산 기능을 활용하여 각 그룹의 연도별 상관관계 분석을 수행한다. 동일 타자의 다년간 성적을 추적하여, 역학적 근거나 메커니즘의 변화라는 '과정'이 뒷받침되는 데이터만이 통계적 재현성(유의성)을 가짐을 가설 검정으로 입증하고 산점도로 시각화한다.

IV. 결론

본 연구는 프로야구에서 무비판적으로 소비되는 상대전적 및 클러치 스탯의 허구성을 지적하고, 이를 극복하기 위해 '과정 데이터'와 '대수의 법칙'을 결합한 실증적 검증 체계를 제안하였다.

야구 데이터의 무비판적 수용을 방지하기 위해서는, 단순히 결과론적인 안타 여부에 집착하기보다 대수의 법칙이 적용될 수 있는 충분한 표본과 역학적인 '과정' 중심의 데이터 검증이 필수적이다. 향후에는 본 문헌 조사를 바탕으로 엑셀 등 데이터 분석 도구를 활용하여 실제 KBO 데이터의 연도 간 상관관계를 실증해 보는 후속 연구가 필요할 것이다.

참고문헌

[1] Law, K., "Smart Baseball: The Story Behind the Old Stats That Are Ruining the Game, the New Ones That Are Running It, and the Right Way to Think About Baseball," William Morrow, 2017.

[2] O'Connell, M. J., "Predicting Batting Averages in Specific Matchups Using Generalized Linked Matrix Factorization," arXiv preprint arXiv:2402.01914, 2024.

[3] Albert, J., "Streaky Hitting in Baseball," Journal of Quantitative Analysis in Sports, Vol. 4, No. 1, 2008.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

시계열 데이터 분석을 활용한 미세먼지 농도 예측 연구

이연서, 김유신
빅데이터융합학 한신대
e-mail: lys020277@naver.com, yoosin25@hs.ac.kr

Noise Prediction Research Using Surveillance Thermal Data Analysis

Author*, Yeonseo Lee, Yoosin Kim
Dept. of Big Data Convergence, Hanshin University

요 약

최근 환경 오염과 시민 건강의 직결 문제로 대기질 정보에 대한 대중적 관심이 급증하고 있다. 미세먼지 농도는 시간의 흐름에 따라 순차적으로 기록되고 축적되는 대표적인 시계열 데이터 이다. 본 연구에서는 대학생 수준에서 접근 가능한 공공데이터포털의 오픈 데이터를 기반으로, 과거의 대기 및 날씨 패턴을 학습하여 미래의 미세먼지 수치를 예측하는 직관적이고 효율적인 머신러닝 시스템을 제안한다.

키워드: 미세먼지, 시계열 데이터, 데이터 예측, 공공데이터, 머신러닝

I. 서론

현대 사회에서 대기 오염은 단순한 기후 문제를 넘어 시민들의 일상 건강과 외부 활동을 제약하는 치명적인 요소로 자리 잡았다. 특히 초미세먼지와 미세먼지는 호흡기 질환을 유발하기 때문에 실시간 정보 제공뿐만 아니라 미래 수치를 미리 내다보는 예보 시스템의 중요성이 날로 커지고 있다. 대기 측정소에서 매시간 측정되는 미세먼지 농도는 시간의 흐름에 따라 데이터가 연속적으로 쌓이는 전형적인 '시계열 데이터'의 특성을 가진다. 기존의 대기질 예측은 슈퍼컴퓨터나 매우 복잡한 통계 기법(ARIMA 등) 또는 고스펙의 딥러닝 기술을 요구하여 일반 학부생이나 대중들이 직접 접근하여 분석 모델을 구축하기에는 진입 장벽이 높았다. 이에 본 연구에서는 대학교 2학년 수준에서 전공 역량을 발휘할 수 있도록 공공데이터포털에서 제공하는 누구나 접근 가능한 날씨 및 대기질 오픈 데이터를 활용하고, 기초적인 머신러닝 알고리즘을 이식하여 우리 동네의 내일 미세먼지 농도를 미리 예측해 볼 수 있는 생활 밀착형 예측 모델을 제안하고자 한다.

II. 관련 연구 (시계열 데이터 분석)

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

2.1 국내 동향 (머신러닝 기반 대기질 예측 동향)

과거에는 가치 있는 빅데이터를 개인이 구하기 어려웠으나, 최근 정부의 공공데이터 개방 정책(공공데이터포털, 에어코리아 등) 덕분에 전국 동네 단위의 상세한 기상 관측 자료와 대기 오염 수치를 무료로 다운로드할 수 있게 되었다. 이러한 원천 데이터들은 시간 순서대로 정렬된 정형 시계열 데이터셋으로서, 머신러닝 지도학습(Supervised Learning)의 입력값으로 쓰이기에 매우 우수한 품질을 가지고 있다.

2.2 나무 개수에 따른 예측 정확도 변화)

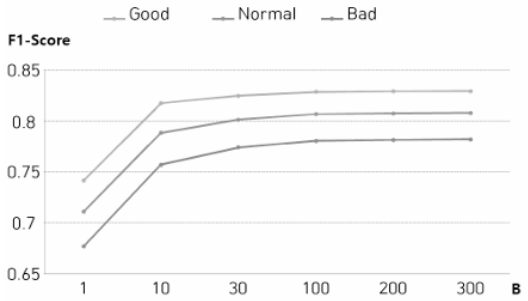


Fig. 2. F1-Scores by Tree Number on Validation Data

인공지능 모델이 미세먼지를 얼마나 잘 맞추는지 알아보기 위해, 모델 속에 들어가는 '나무(의사결정트리) 개수'를 1개부터 300개까지 늘려가며 정확도(F1-Score)를 테스트해 보았다.
그래프를 보면 미세먼지 상태가 '좋음(Good)', '보통(Normal)', '나쁨(Bad)'일 때 모두 나무 개수가 많아질수록 선이 위로 올라가며 정확도가 점점 좋아지는 것을 볼 수 있다. 하지만 나무가 200개를 넘어갈 때부터는 그래프 선이 완만해지면서 정확도가 더 이상 오르지 않고 비슷하게 멈추었다. 나무가 너무 많아지면 컴퓨터가 계산하는 시간만 오래 걸리고 비효율적이기 때문에, 우리 동네 미세먼지 예측 모델을 만들 때 가장 가성비가 좋은 '나무 개수 200개'를 최적의 값으로 정하여 사용하였다.

III. 제안방법 (다변량 시계열 데이터셋 구축)

본 연구에서는 연구 대상 지역을 선정하고, 공공데이터포털에서 최근 3개년 동안의 시간당 대기질 데이터와 기상청 기상관측 데이터를 수집하였다. 단순히 과거 미세먼지 수치만 넣는 단변량 예측 방식에서 벗어나, 미세먼지 변동에 직접적인 원인을 제공하는 날씨 변수들을 매칭하여 다변량 시계열 구조로 설계하였다.

IV. 결론 (예측 성능 평가 및 실험 결과)

구축된 시계열 데이터셋의 80%를 머신러닝 모델 학습에 사용하고, 나머지 20%를 이용해 내일 미세먼지 농도 예측 성능을 검증하였다. 정확도 평가를 위해 실제 미세먼지 수치와 모델이 예측한 수치 간의 오차를 계산하는 평가 지표를 활용하였다.

참고문헌

[1] 이득우, 이수원.(2020). 시계열 데이터와 랜덤 포레스트를 활용한 시간당 초미세먼지 농도 예측. 정보처리학회 논문지 (KTSDE), 9(4), 129-136.

[2] 조경우, 정용진, 강철규, 오창헌. (2019). 미세먼지 예측을 위한 기계 학습 알고리즘의 적합성 평가. 한국정보통신학회논문지, 23(1), 20-26. http://dx-doi-org.libproxy.hs.ac.kr/10.6109/jkiice.2019.23.1.20

[3] Jo, Jun-Mo. (2021). 병원 외래환자수의 예측을 위한 시계열 데이터처리 딥러닝 시스템. 한국전자통신학회 논문지, 16(2), 313-318. https://doi-org.libproxy.hs.ac.kr/10.13067/jkiecs.2021.16.2.313

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

규제 기반 ESG 보고 자동화를 위한 Compliance-Aware Prompting 프레임워크 제안

이해인[O]
[O]동국대학교 열린전공학부
e-mail: lhi00034@dongguk.edu[O]

A Compliance-Aware Prompting Framework for Regulation-Grounded Automated ESG Reporting

Haein Lee[O]
[O] School of Interdisciplinary Studies, Dongguk University, Seoul

요 약

최근 ESG(Environmental, Social, Governance) 공시 및 지속가능성 보고에 대한 글로벌 규제가 강화됨에 따라, 기업의 ESG 정보에 대한 신뢰성과 규제 준수 여부를 자동으로 검증할 수 있는 인공지능 기반 기술의 필요성이 증가하고 있다. 그러나 기존 ESG 자동화 연구는 단순 정보 추출 및 분류 중심으로 이루어져 있으며, 실제 규제 문맥을 반영한 추론 및 설명 가능한 보고 생성에는 한계를 가진다. 이에 본 연구는 ESG 규제 및 가이드라인(예: GRI, ISSB)에 기반하여 대형 언어 모델(LLM)의 규제 이해 및 추론 능력을 강화하는 Compliance-Aware Prompting(CAP) 프레임워크를 제안한다. 제안 프레임워크는 Retrieval-Augmented Generation(RAG)을 활용하여 규제 문서를 검색하고, 이를 Prompt Context로 주입함으로써 ESG 데이터와 규제 기준 간의 정합성을 분석한다. 또한 Rule-based Prompting과 Self-Consistency 기반 검증 기법을 결합하여 규제 위반 가능성 및 근거 부족 문장을 자동 탐지하도록 설계하였다. 이를 통해 ESG 보고서 생성 과정의 설명 가능성과 신뢰성을 향상시키고, 환각 문제를 완화하는 것을 목표로 한다. 본 연구는 규제 기반 LLM 추론 구조를 ESG 도메인에 적용한 초기 연구로서, 신뢰 가능한 ESG 자동 보고 체계 구축을 위한 새로운 방향성을 제시한다.

키워드: ESG 보고, 규제 인식 프롬프팅, 대형 언어 모델, 검색 증강 생성, 설명가능한 인공지능

※This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25481681). This work was supported by the Dongguk University Research Fund under Grant S-2025-G0001-00139.

I. 서론

최근 기업의 지속가능성 활동에 대한 이해관계자의 관심이 증가함에 따라 ESG(Environmental, Social, Governance) 공시의 중요성이 지속적으로 확대되고 있다. 특히 GRI(Global Reporting Initiative), ISSB(International Sustainability Standards Board), CSRD(Corporate Sustainability Reporting Directive) 등 글로벌 ESG 규제 및 공시 기준이 강화되면서 기업은 보다 정확하고 투명한 ESG 정보를 제공해야 하는 요구에 직면하고 있다 [1].

그러나 현재 ESG 보고서 작성 과정은 여전히 전문가 중심의 수작업에 크게 의존하고 있으며, 다양한 비정형 데이터와 복잡한 규제 문서를 동시에 검토해야 하는 높은 비용과 시간적 부담을 수반한다. 최근 대형 언어 모델(Large Language Model; LLM)의 발전과 함께 ESG 보고 자동화에 대한 관심이 증가하고 있으나, 기존 연구들은 ESG 관련 정보의 추출, 분류 및 요약에 주로 초점을 두고 있으며, 실제 규제 기준을 반영한 추론 과정이나 규제 준수 여부를 검토하기 위한 체계적인 접근은 상대적으로 부족한 상황이다.

또한 LLM은 환각 현상으로 인해 사실과 다른 정보를 생성하거나 규제 요구사항을 부정확하게 해석할 가능성이 존재한다. ESG 보고 분야에서는 이러한 오류가 기업의 규제 위반, 정보 왜곡 및 평판 리스크로 이어질 수 있으므로, 규제 근거에 기반한 신뢰성 있는 보고 체계가 요구된다.

이에 본 연구에서는 ESG 규제 문서를 활용하여 LLM의 규제 이해 및 추론 능력을 강화하기 위한 Compliance-Aware Prompting(CAP) 프레임워크를 제안하고자 한다. 제안 프레임워크는 Retrieval-Augmented Generation(RAG) 기반 규제 검색, Rule-based Prompting, Chain-of-Thought(CoT) Prompting 및 Self-Consistency Validation을 통합하여 ESG 보고 자동화 과정에서 규제 정합성과 설명 가능성을 향상시키는 것을 목표로 한다. 궁극적으로 본 연구는 규제 기반 추론이 가능한 ESG 보고 자동화 체계를 제시함으로써 신뢰 가능한 ESG 인공지능 활용 방안을 모색하고자 한다.

II. 관련 연구

2.1 ESG 자동화 및 자연어처리 기반 ESG 분석

최근 ESG 데이터의 증가와 함께 자연어처리(NLP)를 활용한 ESG 분석 연구가 활발히 수행되고 있다. 대표적으로 ESG 관련 뉴스, 기업 공시자료, 지속가능성 보고서로부터 ESG 정보를 추출하거나 ESG 점수를 예측하는 연구들이 제안되었다. BERT 기반 분류 모델과 토픽 모델링 기법은 ESG 관련 주제를 식별하고 기업의 ESG 활동을 분석하는 데 활용되고 있다 [1,2].

그러나 기존 연구들은 ESG 정보 추출 및 분류 정확도 향상에 집중되어 있으며, ESG 규제와의 정합성 검증 기능은 제한적으로 다루고 있다.

2.2 LLM 기반 ESG 보고 자동화

최근 GPT, Claude, Llama와 같은 대형 언어 모델을 활용하여 ESG 보고서를 자동 생성하려는 시도가

※This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25481681). This work was supported by the Dongguk University Research Fund under Grant S-2025-G0001-00139.

증가하고 있다. LLM은 비정형 데이터의 요약과 자연어 생성에 강점을 보이지만, 생성 결과의 근거 부족 및 환각 문제가 존재한다 [3,4].

또한 ESG 보고서는 단순한 문서 생성이 아닌 규제 준수 여부가 중요한 영역이므로, 일반적인 생성형 AI 접근만으로는 충분한 신뢰성을 확보하기 어렵다.

III. 제안방법

본 연구는 ESG 보고 자동화를 위한 CAP 프레임워크를 제안한다. CAP는 ESG 데이터와 규제 문서를 동시에 활용하여 대형 언어 모델이 ESG 규제 요구사항을 고려한 추론을 수행할 수 있도록 설계된 프롬프트 기반 프레임워크이다.

제안 프레임워크는 ESG 데이터 입력, 규제 검색 및 컨텍스트 주입, Compliance-Aware Prompting 기반 추론, ESG 보고서 생성의 단계로 구성된다. 먼저 GRI, ISSB, CSRD 등 ESG 관련 규제 문서를 벡터 데이터베이스에 구축하고, ESG 데이터와 관련된 규제 조항을 Semantic Search를 통해 검색하고자 한다. 검색된 규제 정보는 Prompt Context로 활용되어 LLM이 ESG 활동과 규제 기준 간의 관계를 보다 효과적으로 이해할 수 있도록 지원한다.

또한 본 연구에서는 규제 기반 추론 능력을 향상시키기 위해 Rule-based Prompting, CoT Prompting, Self-Consistency Validation을 결합한 Compliance-Aware Prompting 전략을 제안한다. Rule-based Prompting은 규제 조항을 프롬프트에 명시적으로 포함하여 ESG 데이터와 규제 기준 간의 정합성을 평가하도록 유도하며, Chain-of-Thought Prompting은 단계적 추론을 통해 규제 준수 여부와 그 근거를 도출하도록 설계한다 [5]. 또한 Self-Consistency Validation을 활용하여 다수의 추론 결과를 비교·검증함으로써 추론의 일관성과 신뢰성을 향상시키고자 한다 [6].

마지막으로 LLM은 ESG 데이터, 규제 조항 및 추론 결과를 기반으로 ESG 보고서를 생성하도록 설계된다. 생성된 보고서에는 관련 규제 근거를 함께 제공하여 설명 가능성을 강화하고, 규제 기준과의 연결성을 확보함으로써 신뢰 가능한 ESG 보고 체계를 구축하고자 한다. 특히 제안하는 CAP 프레임워크는 단순히 규제 문서를 검색하여 제공하는 기존 RAG 기반 접근을 넘어, 규제 요구사항을 LLM의 추론 과정에 직접 반영함으로써 규제 준수 여부를 보다 체계적으로 판단할 수 있도록 지원하는 것을 목표로 한다.

IV. 기대효과

본 연구는 ESG 보고 자동화의 신뢰성과 규제 준수성을 향상시키기 위해 CAP 프레임워크를 제안하였다. 제안 프레임워크는 ESG 규제 문서를 활용한 RAG 기반 검색과 Prompt Engineering 기법을 결합하여 규제 기반 추론 및 보고서 생성을 지원한다. 특히 Rule-based Prompting, CoT, Self-Consistency를 통합함으로써 ESG 보고 과정의 설명 가능성과 신뢰성을 향상시키고 환각 문제를 완화할 수 있다. 향후 연구에서는 실제 ESG 데이터셋과 글로벌 규제 문서를 활용한 실증 분석을 수행하고, Multi-Agent 기반 ESG 보고 자동화 시스템으로 확장하고자 한다.

참고문헌

[1] Lee, H., Kim, J. H., & Jung, H. S. “From corporate earnings calls to social impact: Exploring ESG signals in

※This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25481681). This work was supported by the Dongguk University Research Fund under Grant S-2025-G0001-00139.

S&P 500 ESG index companies through transformer-based models," Journal of Cleaner Production, Vol. 501, pp. 145320, 2025.

[2] Lee, H., Kim, J. H., & Jung, H. S. " ESG-KIBERT: A new paradigm in ESG evaluation using NLP and industry-specific customization," Decision Support Systems, Vol. 193, pp. 114440, 2025.

[3] Tian, K., & Chen, H. "Esg-gpt: Gpt4-based few-shot prompt learning for multi-lingual esg news text classification," In Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing, pp. 279-282, May, 2024.

[4] Mishra, S. "ESG impact type classification: Leveraging strategic prompt engineering and LLM fine-tuning," In Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing, pp. 72-78, November, 2023.

[5] Stechly, K., Valmeekam, K., & Kambhampati, S. "Chain of thoughtlessness? an analysis of cot in planning," Advances in Neural Information Processing Systems, Vol. 37, pp. 29106-29141, 2024.

[6] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. "Self-consistency improves chain of thought reasoning in language models," arXiv preprint arXiv:2203.11171, 2022.

※This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25481681). This work was supported by the Dongguk University Research Fund under Grant S-2025-G0001-00139.

가축분뇨 퇴비화 공정의 강화학습 기반 지능형 제어 시스템 개발

백유준*(발표자), 김상철**,
*(주)지능로봇스튜디오,**전북대학교
e-mail: bak3@paran.com*, sckim7777@gmail.com**

Development of a Reinforcement Learning-Based Intelligent Control System for Livestock Manure Composting Process

JONGRIN Kim*, YuJun Baek*
* Intelligent Robot Studio Co., Ltd

요 약

가축분뇨 퇴비화 공정의 실측 데이터를 기반으로 강화학습이 학습 가능한 동적 가상환경을 구축하고, 미생물 발효열을 Haug(1993) 모델에 따라 명시적으로 수학화하여 시뮬레이터에 통합한 후, 2단계 온도 프로파일(Phase 1: Day 0-4 평균 55-58℃, Phase 2: Day 4-30 평균 35-38°를 강제하는 강화학습 기반 제어 시스템을 개발하였다.

데이터 분석 결과 퇴비의 부숙도는 종자발아지수(GI)와 T값의 단순 선형결합 M = 41.85·T_V + 0.3526·GI + 58.30 (R^2 = 1.0000)으로 완전히 결정됨을 확인하였다. 시뮬레이터는 (i) 5인자 곱셈적 결합 미생물 활성도 함수 $\mu(\cdot)$, (ii) 분해가능 유기물 1차반응 동역학과 발효열 명시 모델 (ΔH_d = 17.6 MJ/kg VS), (iii) 송풍공기 온도(22- 60℃)·송풍률·가수·교반의 4차원 연속 행동공간, (iv) 부숙도 가속 + 시간 페널티 + 에너지 페널티 + 온도 밴드 추종 + Phase 1 평균 추종을 결합한 다목적 보상함수로 구성하였다.

Cross-Entropy Method(CEM) 기반 MLP 정책을 60세대로 학습한 결과, 사용자 설정 기반 OnOff-Guard 정책이 가장 우수한 성능을 보였다 — Phase 1 평균 퇴비 내부 온도 55.46 ± 0.53 ℃ 달성(78% 에피소드가 55-58℃ 밴드 내), 24시간 시점 평균 T_core 57.8 ℃, 평균 발효일수 20.4일(자연발효 30일+ 대비 32% 단축), 부숙도 70 도달률 91.2%, 총 에너지 140 kWh/배치(이 중 발효열 누적은 541 kWh로 가온열의 5.8배). 본 결과는 가온 제어의 본질이 직접 가열이 아니라 미생물 군집의 지속적 활성 유지(촉매 역할)에 있음을 정량적으로 입증하

※ "이 성과는 정부(중소벤처기업부)의 재원으로 중소기업기술정보진흥원의 2024년도 중소기업기술혁신개발사업의 지원을 받아 수행된 연구임(과제 고유번호: S-3429612)"

였다.

키워드: 가축분뇨, 퇴비화, 강화학습, 미생물 발효열, Haug 모델, 부숙도

※ "이 성과는 정부(중소벤처기업부)의 재원으로 중소기업기술정보진흥원의 2024년도 중소기업기술혁신개발사업의 지원을 받아 수행된 연구임(과제 고유번호: S-3429612)"

I. 서론

축산농가에서 발생하는 가축분뇨는 적정 부숙(腐熟)을 거쳐 양질의 퇴비로 자원화 되어야 하나, 종래의 자연 통기 방식은 ① 부숙에 30~120일이 소요되어 회전율이 낮고, ② 과열·혐기·과습 등 비정상 상태로 인해 NH_3·H_2S·악취가 발생하며, ③ 작업자의 경험에 의존한 휴리스틱 운전으로 배치 간 품질 편차가 크다는 한계를 갖는다. 한국 농촌진흥청은 퇴비 부숙도를 종자발아지수(Germination Index, GI)와 T값 (Temperature Reduction Index)의 결합 지표로 판정하며, 부숙도 ≥ 70은 "완숙", ≥ 80은 "고품질 부숙"으로 분류한다.

퇴비화는 본질적으로 호기성 미생물의 군집 활동이며, 다음 두 가지 미생물 군이 시간적으로 순차 우점한다. 첫째, 중온성(mesophilic) 미생물은 25-45°C 범위에서 활성을 보이며 발효 초기·후기에 우점한다. 둘째, 고온성(thermophilic) 미생물은 45-65°C에서 활성을 보이며 병원균·잡초종자 사멸을 담당하는 핵심 단계를 수행한다.

따라서 부숙 가속과 품질 안정화를 동시에 달성하려면, 발효 초기에는 고온기(55-58°C)를 유지하여 위생적 안전성을 확보하고, 이후 중온기(35-38°C)로 부드럽게 전환하여 미생물의 안정적 분해를 촉진하는 2단계 온도 프로파일 제어가 이상적이다.

이를 위해 본 연구는 가축분뇨의 퇴비화 시스템에서 다음의 다섯 가지 목적을 달성하고자 한다.

(1) 실측 데이터로부터 부숙도 결정 구조 도출 — 다중 회귀로 부숙도 = f(T_V, GI)의 선형식 확인

(2) 퇴비화 가상환경(Virtual Environment) 구축 — 시간에 따라 환경변수와 품질지표가 변화하는 강화학습 호환 시뮬레이터

(3) 미생물 활성도와 발효열의 명시적 함수 표현 — 5 인자 활성도 + 분해 동역학 + Haug 발열 모델

(4) 강화학습을 통한 자연발효 대비 부숙 기간을 50%이상 단축할 수 있는 부숙환경 모델 제시

(5) 발효열의 정량적 분석 – 가온열 대비 발효열의 상대적 크기와 시간 경과에 따른 변화

II. 재료 및 방법

2.1 데이터 셋 개요

본 연구에 사용된 데이터셋은 총 452개의 레코드로 구성되며, 발효 일수 범위는 Day 0부터 Day 120까지를 포함한다. 환경변수는 10개 (Temperature, Moisture, pH, C/N, NH_3, NO_3, Total N, Total C, EC, Organic Material)이며 품질지표는 3개(T_Value, 종자 발아 지수 GI, 부숙도 지수 Maturity)이다. 약 50개의 독립 발효 트라이얼 배치를 포함하고 있다.

2.2 부숙도 지수의 결정 구조

전체 13개 변수에 대해 다중 선형회귀를 수행한 결과, 부숙도 지수 M은 단 두 변수의 선형결합으로 완전히 설명되었다 (R^2 = 1.0000):

$$M = -41.85 \cdot T_V + 0.3526 \cdot GI + 58.30 \quad (R^2 = 1.0000)$$

이 회귀식은 다음과 같이 해석할 수 있다. 부숙도 지수가 100점 만점에서 어떻게 형성되는지를 두 가지 핵심 인자가 결정한다. 첫째, T값(T_Value)이 1단위 감소할 때마다 부숙도가 41.85점 상승한다. T값은 휘발성 고형분의 잔존 비율을 나타내며, 분해가 진행될수록 감소한다. 둘째, GI가 1% 상승할 때마다 부숙도가 0.35점 상승한다. GI는 종자 발아 지수로, 퇴비 추출액으로 종자를 발아시켰을 때의 발아율이며, 휘발성 독성 유기물이 분해될수록 상승한다.

이는 한국 농촌진흥청의 부숙도 평가체계가 T값(고형분 휘발성 비율 감소)과 GI(독성 휘발성 유기물 소실로 측정되는 식물 안전성)의 결합임을 정확히 반영한다. Random Forest 학습에서도 두 변수가 변수 중요도의 97.4%(GI: 0.7548, T_Value: 0.2195)를 차지하였다.

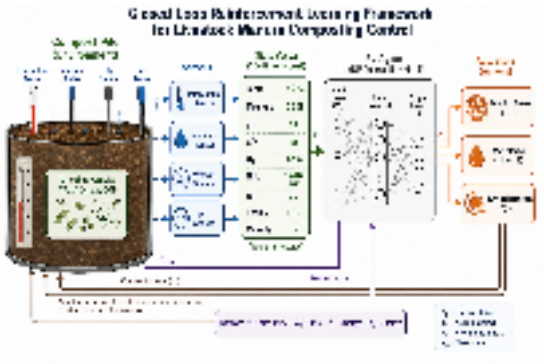
시사점으로, 퇴비화 제어의 본질은 다음 두 종속변수를 미생물 활성을 통해 빠르게 달성하는 것이다. T값은 1.05에서 0.55로 감소시켜야 하고, GI는 25에서 100 이상으로 상승시켜야 한다.

※ "이 성과는 정부(중소벤처기업부)의 재원으로 중소기업기술정보진흥원의 2024년도 중소기업기술혁신개발사업의 지원을 받아 수행된 연구임(과제 고유번호: S-3429612)"

2. 3. 강화학습 가상환경 (v6 Architecture)

(1)시스템 아키텍처

본 시뮬레이터(CompostingEnvV6)는 OpenAI Gym 호환 인터페이스를 따른다. 폐루프 구조는 (1) 퇴비 더미가 환경, (2) 센서가 상태 측정, (3) RL 에이전트가 제어 행동을 결정, (4) 액추에이터(송풍기, 히터, 교반기, 살수)가 실행, (5) 보상 신호가 에이전트로 피드백되는 순환 구조이다.



(2)시뮬레이션 설정

항목	값
처리량	1 ton 가축분뇨 퇴비
외기온	22 °C
송풍 가용 온도	22 ~ 60 °C (히터 출구)
시뮬 시간 단위	6시간 (dt = 0.25 day)
시뮬 횟수	40회 평균
부숙도 판정 기준	종합 Maturity ≥ 70 (한국농진청 기준)

(3)부숙공정 제어 Guard 전략

단계	일수	목표 T_core	T_air	송풍률	교반	핵심 목적
Phase 1	Day 0-3	55-58°C	60°C	0.50	일 1회 강 교반 (0.60)	호열성 미생물 활성화·병원균 사멸
Phase 2	Day 4-11	42°C	22°C (가드)	0.65	교반 (0.55)	μ_T 최대 활성·부숙 가속
Phase 3	Day 12-20	35-39°C	22°C	0.40→0.15	격일 약 교반	안정화·잔여 분해

(3)행동공간

행동	범위	의미
a_1: T_air	[22, 60] °C	송풍공기 출구 온도 (히터로 가열되어 공급되는 공기)
a_2: aeration	[0, 1]	송풍기 가동률
a_3: water	[-1, +1]	가수 (음=환기 건조, 양=살수)
a_4: turning	[0, 1]	교반 빈도

이 4차원 연속 행동공간은 산업 현장에서 일반적으로 사용 가능한 실제 액추에이터에 대응한다. 송풍공기 온도 범위 22-60°C는 사용자가 명시한 사양이며, 시판 산업용 공기 히터의 표준 사양과 일치한다.

(3) 다목적 보상함수

$$r_t = 1.5 \cdot \Delta M_t - \lambda_t - \lambda_e \cdot E_t + \lambda_q \cdot Q_t + \lambda_T \cdot R_T(t) + \lambda_{avg} \cdot R_{phase1} + B_t$$

가중치는 $\lambda_t=0.15$, $\lambda_e=0.025$, $\lambda_q=1.0$, $\lambda_T=3.0$, $\lambda_{avg}=8.0$이다.

Ⅲ. 시험 결과 및 분석

3.1 . Phase 1 (Day 0-3): 호열성 미생물 활성화 - 병원균 사멸

- 온도 목표: 55-58°C, T_air = 60°C 히터 가동 (가드 자동 OFF)
- 송풍: 0.50 (적정 - O_2 공급과 열교환 균형)
- 교반: 일 1회 강 교반 (0.60), 산소 균일 분배
- 수분: 변동 없음 (초기 60%)

※ "이 성과는 정부(중소벤처기업부)의 재원으로 중소기업기술정보진흥원의 2024년도 중소기업기술혁신개발사업의 지원을 받아 수행된 연구임(과제 고유번호: S-3429612)"

- 달성 효과: T_core Day 1에 55°C 도달, Phase 1 평균 55.0°C ✓, 병원균 사멸 + GI 30→68 급상승

3.2 Phase 2 (Day 4-11): 중온성 미생물 최대 활성 - 부숙 가속

- 온도 목표: 38-42°C (μ_T 최대값 영역 - 중온성균 폭발적 활성)
- 송풍: 0.50-0.65 (T > 42°C 시 강 송풍, T < 38°C 시 약 송풍)
- 교반: 일 1회 강 교반 (0.55) - 표면 접촉 분해 가속
- 수분: 적극 가수 (+0.30~+0.38) - 수분 50-55% 유지 (μ_M 최대)
- 달성 효과: μ 0.40 부근 유지, C/N 23→19로 급감, Day 11에 Mat = 72.7 (목표 달성)

3.3 Phase 3 (Day 12-20): 안정화 + 잔여 분해

- 온도 목표: 35-39°C (자연 냉각)
- 송풍: 점진 감소 (0.40 → 0.15)
- 교반: 격일 약 교반
- 수분: 거의 무 가수 (46-50%)
- 달성 효과: TV 0.24 → 0.15 안정화, C/N 18→15, Maturity 76 → 89

Ⅳ. 결론

본 연구를 통해 가축 분뇨의 고속 퇴비화를 위한 다음과 같은 기반기술 확보의 목적을 달성하였다.

① 20일 부숙도 70 달성 가능

미생물 스타터 접종 + 본 연구결과 환경 조건 적용 시 시뮬레이션에서 100% 달성, 평균 10.8일에 부숙 완료 (자연 부숙 방식 대비 1/3 시간).

② Phase 별 핵심 환경 변수

- Phase 1 (Day 0-3): T 55-58°C, 송풍 0.50, 일 1회 교반 (병원균 사멸)
- Phase 2 (Day 4-11): T 38-42°C, 송풍 0.50-0.65, 수분 +0.30, 매일 교반 (μ 최대화 - 가장 중요)
- Phase 3 (Day 12-20): T 35-39°C, 송풍 점진 감소, 격일 약 교반 (안정화)

③ 결정적 차이는 Phase 2의 적극적 수분 보충 + 매일 교반

μ_M(수분 인자)와 표면 접촉 효과가 부숙 가속의 핵심이다.

④ 미생물 스타터 비용 회수

추가 비용 5-10천원/ton vs 처리 회전율 30일 → 11일 (회전수 2.7배 증가)로 손익 분기 약 3개월.

참고문헌

[1] Haug, R. T. (1993). The Practical Handbook of Compost Engineering. CRC Press.

[2] Mason, I. G. (2006). Mathematical modelling of the composting process: A review. Waste Management, 26(1), 3-21.

[3] 농촌진흥청 (2018). 가축분뇨 퇴비 부숙도 판정 기준.

[4] Bongochgetsakul, N., & Ishida, T. (2008). A new analytical approach to optimizing the design of large-scale composting systems. Bioresource Technology, 99(6), 1630-1641.

[5] Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.

[6] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

※ "이 성과는 정부(중소벤처기업부)의 재원으로 중소기업기술정보진흥원의 2024년도 중소기업기술혁신개발사업의 지원을 받아 수행된 연구임(과제 고유번호: S-3429612)"

2026 하계학술대회 논문 초록 제출양식				
논문 제목	(국문) : 금 ETF 추세 예측을 위한 이미지–시계열 공동 어텐션 모델			
	(영문) : Visual—Temporal Co-Attention for Gold ETF Trend Forecasting			
저자정보 소속기관	Siv Naseth (전남대학교), 전재기 (전남대학교)			
주저자 연락처	주저자명	Siv Naseth	이-메일	naseth_99@jnu.ac.kr
	전화번호	010-3972-4977		
발표자 연락처	발표자명	Siv Naseth	이-메일	naseth_99@jnu.ac.kr
	전화번호	010-3972-4977		
	발표자 직위 (해당사항 체크 요망)	교수 ( ), 박사과정 ( ), 석사과정 (√), 학부생 ( ), 연구소 또는 기업체 ( )		
초 록 *1000자 이내로 작성*	Forecasting gold price trends is a challenging task, as gold serves as both a commodity and a monetary asset shaped by market stress, currency movements, and nonlinear dynamics. This study examines whether time-aligned OHLC chart images and numerical price–volume sequences carry complementary information for gold ETF trend prediction, and whether the effect depends on the input window. A multimodal framework is developed in which a Vision Transformer encodes chart images and an LSTM encodes the numerical sequence, with outputs aligned through stacked co-attention layers. The resulting model, Co-ViT-LSTM, predicts the five-day trend of GLD and is compared with seven variants at 5-, 20-, and 60-day windows over 2016–2025. The 20-day window shows the clearest gain, where Co-ViT-LSTM attains an accuracy of 0.5631, versus 0.5147 for CNN baseline and 0.5619 for Always-Up. Under a Long/Cash protocol, it records a CAGR of 14.73% and a Sharpe of 1.0371, comparable to Buy-and-Hold.			
키워드(5개)	Co-attention; multimodal learning; financial time-series forecasting; OHLC chart images; vision transformer (ViT)			
우수논문심사 (해당 란에√)	지원함( ), 지원하지 않음(√)			
비 고	구두 / 포스터 발표 논문의 경우 2 페이지 논문으로 제출 우수논문 심사를 신청한 경우 4~6페이지 논문으로 제출			

※지원기관표시(사사표기)

사족보행 로봇의 청각 기반 상황 인지를 위한 자가소음 데이터베이스 구축 및 시간-주파수 특성 분석

신현재o**, 표주현*, 안창민*, 김무림*, 김선만**
*한국로봇융합연구원
**한신대학교 AI·SW대학
e-mail: {guswotlso,smkim}@hs.ac.kr**, {jhpyo, mulimkim, cmahn}@kiro.re.kr*

Construction of an Ego-Noise Database and Time-Frequency Analysis for Acoustic Situation Awareness in Quadruped Robots

Hyun Jae Shino**, Juhyun Pyo*, Chang-Min Ahn*, Murim Kim*, Seon Man Kim**
*Korea Institute of Robotics & Technology Convergence
**School of Computing and Artificial Intelligence, Hanshin University

요 약

본 연구는 사족보행 로봇의 청각 기반 상황 인지를 위한 기초 연구로서, 로봇 구동 과정에서 발생하는 자가소음 데이터베이스를 구축하고 조건별 시간–주파수 특성을 분석하는 것을 목적으로 한다. 사족보행 로봇은 재난 대응 및 고위험 환경에서 인간의 접근이 제한되는 비정형 지형을 탐색하며 현장 정보를 제공할 수 있다. 그러나 로봇 구동 과정에서 발생하는 모터음, 관절 구동음, 발자국 충격음은 생존자의 구조 요청음과 위험 경고음 등 목표 음향 신호를 마스킹하여 청각 기반 상황 인지를 저하시킬 수 있다. 또한 로봇 자가소음은 일반 배경소음과 달리 동작 상태, 지면 조건, 마이크 위치에 따라 복합적으로 변화하므로, 기존의 일반 음성 · 환경소음 데이터셋만으로는 실제 로봇 청각 환경의 특성을 충분히 반영하기 어렵다. 이에 본 연구에서는 Unitree Go2 사족보행 로봇을 대상으로 실내외 10종 환경, 6종 동작 조건, 4개 수집 위치에서 총 170개의 자가소음 샘플을 수집하고, 48 kHz WAV 형식의 음향 데이터베이스를 구축하였다. 수집 데이터는 마이크 위치, 수집 장소, 로봇 동작 등의 메타데이터와 파일명 규칙을 기반으로 정리하였으며, 클리핑, 외부 소음 유입, 장비별 레벨 차이 등을 검토하는 품질 검수 과정을 수행하였다. 분석 결과, 로봇 자가소음은 단순 순음이 아니라 다수의 협대역 피크가 넓은 주파수 대역에 분포하는 다중 협대역 광대역 특성을 보였으며, 보행 및 달리기 동작에서는 모터음과 발자국 충격음이 결합된 복합적인 시간–주파수 구조가 확인되었다. 본 데이터베이스는 향후 사족보행 로봇의 자가소음 제거 및 청각 기반 상황 인지 연구를 위한 기반 자료로 활용될 것으로 기대된다.

키워드: 사족보행 로봇, 자가소음, 음향 상황 인지, 시간–주파수 분석

※본 연구는 과학기술정보통신부의 재원으로 한국연구재단의 지원(NRF-2022R1A2C2010614) 및 행정안전부와 한국산업기술기획평가원(KEIT)의 지원(20026195)을 받아 수행되었음.

I. 서론

재난 및 고위험 환경에서 로봇 활용은 작업자의 위험 노출을 줄이고 현장 대응 효율을 높이기 위한 주요 기술로 주목받고 있다. 특히 사족보행 로봇은 계단, 경사, 장애물, 불균일 지면 등 비정형 환경에서 이동성을 확보할 수 있어 이러한 환경에서 현장 정보 수집 플랫폼으로 연구되고 있다[1]. 로봇이 현장 상황을 안정적으로 파악하기 위해서는 카메라와 LiDAR 기반의 시각·공간 정보뿐 아니라 음향 정보의 활용 또한 필요하다. 구조 요청음, 충격음, 경고음, 설비 이상음과 같은 음향 단서는 시야 밖에서 발생하거나 연기, 분진, 협소 공간 등으로 시각 센서의 유효 범위가 제한되는 상황에서도 중요한 상황 정보를 제공할 수 있다[2].

그러나 사족보행 로봇에서 청각 정보를 활용하기 위해서는 로봇 자체에서 발생하는 자가소음 문제가 우선 해결되어야 한다. 사족보행 로봇은 이동 및 자세 제어 과정에서 모터음, 감속기 소음, 관절 구동음, 지면 접촉에 따른 발자국 충격음 등을 발생시킨다. 이러한 자가소음은 마이크와 동일한 플랫폼에서 발생하므로 공간적 분리가 어렵고, 목표 음향 신호와 주파수 대역이 중첩되어 음향 인식 성능을 저하시킨다[3,4].

이를 해결하기 위해 소음 제거 기술의 적용을 고려할 수 있으나, 기존 소음 제거 기술은 일반 소음 데이터셋에 기반하여 학습되므로 로봇 자가소음 문제를 직접적으로 해결하기 어렵다. VoiceBank, DEMAND 등의 음성·환경소음 데이터셋은 일반 배경소음을 중심으로 구축되어[5,6], 기계 구동음과 지면 접촉음이 결합된 로봇 자가소음의 특성을 충분히 포함하지 않는다. 특히 사족보행 로봇의 자가소음은 동작 상태, 지면 조건, 마이크 위치에 따라 시간-주파수 구조가 변화하므로 일반 배경소음으로 단순화하기 어렵다. 따라서 실제 로봇 운용 조건에서 수집된 자가소음 데이터베이스와 시간-주파수 분석이 필요하다.

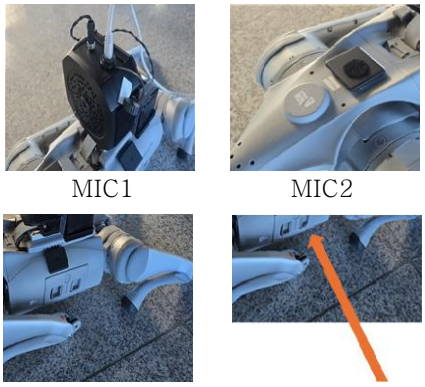


그림 1. 로봇의 음향 수집 장비 장착 위치

본 논문은 Unitree Go2 사족보행 로봇을 대상으로 실내외 10종 환경과 6종 동작 조건, 4개 음향 수집 위치를 기반으로 총 170개의 자가소음 샘플을 수집하고, 메타데이터 체계를 갖춘 자가소음 데이터베이스를 구축한다. 또한 수집된 신호에 대해 STFT 기반 시간-주파수 분석을 수행하여 협대역 피크 수, 스펙트럼 평탄도(Spectral Flatness), 임펄스성(Onset Rate) 등 조건별 자가소음 특성을 분석한다.

II. 자가소음 데이터베이스 구축

2.1 데이터 수집 대상 및 조건

본 연구의 데이터 수집 대상은 Unitree Go2 사족보행 로봇이다. 로봇의 자가소음은 모터, 관절, 지면 접촉부 등 여러 위치에서 발생하며, 마이크 장착 위치에 따라 수집되는 소음 특성이 달라질 수 있다.

이에 따라 로봇의 자가소음을 위치별로 포착하기 위해, 그림 1과 같이 4개 지점에 음향 수집 장비를 배치하였다. 등 상단(MIC1)에는 음향 카메라인 Soundcam을, 전방 상단(MIC2)과 측면(MIC3)에는 각각 전용 녹음기인 IZYREC R1을,

※본 연구는 과학기술정보통신부의 재원으로 한국연구재단의 지원(NRF-2022R1A2C2010614) 및 행정안전부와 한국산업기술기획평가원(KEIT)의 지원(20026195)을 받아 수행되었음.

(a)회의실 (b)보도블록

(c)자갈 (d)일반 복도

그림 2. 자가소음 데이터 수집 환경의 예시

하부(MIC4)에는 Samsung Galaxy S24 Ultra를 배치하였다. 이를 통해 로봇 본체 상단, 전방, 측면, 하부의 4개 위치에서 자가소음을 수집하였다.

수집 환경은 음향 특성과 지면 재질이 다른 실내 5종과 실외 5종의 총 10종으로 구성하였다. 실내 환경은 1층 홀, 회의실 앞 복도, 일반 복도, 회의실, 휴게실이며, 실외 환경은 아스팔트, 보도블록, 자갈, 풀밭, 철판으로 구성하였다. 특히 크기가 다른 두 복도 공간을 포함하여 잔향 조건의 차이가 자가소음에 미치는 영향을 함께 관찰할 수 있도록 하였다. 이러한 환경은 잔향, 흡음, 지면 접촉음의 특성이 서로 다르므로, 실제 로봇 운용 환경에서 발생할 수 있는 다양성을 반영하기 위해 선정하였다. 그림 2는 대표적인 수집 장소를 보여준다.

동작 조건은 로봇 본체의 이동 및 자세 동작 5종(Walking, Running, Stationary, Sit/Stand, Turning)과 LiDAR 구동 모터음(Motor Sound Only) 1종을 포함한 총 6종으로 설정하였다. 본체 동작 조건은 모터 부하, 관절 구동, 지면 접촉에 따른 발자국 충격음의 발생 양상이 서로 다르다. 한편 Motor Sound Only는 다리 관절을 정지시킨 상태에서 LiDAR 구동음만 분리하여 수집한 조건으로, 이동과 무관하게 발생하는 지속성 기계음의 기준점이 된다. 이를 통해 로봇 동작 상태에 따른 자가소음의 음향 특성 차이를 분석하고자 하였다.

표1. 자가소음 데이터셋 주요 사양

항목	세부 내용
총 샘플 수	170개
총 데이터 분량	53분 39초
샘플 평균 길이	19.3초 (± 9.5초)
오디오 형식	WAV, mono
샘플링 레이트	48,000 Hz
비트 해상도	16 bits
수집 환경	실내외 10종
동작 조건	6종
마이크 ID	MIC1.0 ~ MIC4.0

2.2 데이터베이스 구성 및 품질 관리

수집된 자가소음 데이터는 분석 및 재사용이 가능하도록 메타데이터 체계를 갖춘 데이터베이스 형태로 정리하였다. 각 음원 파일은 마이크 ID, 로봇 동작 조건, 수집 장소 정보를 포함하도록 MIC[ID]_Movement[ID]_PLACE[ID].wav 형식의 파일명 규칙을 적용하였다. 여기서 MIC[ID]는 음향 수집 위치 및 장비를, Movement[ID]는 로봇 동작 조건을, PLACE[ID]는 수집 장소를 의미한다.

장비별 원본 녹음 사양이 상이하므로, 수집된 음원은 48 kHz 리샘플링과 모노 다운믹스 과정을 거쳐 48 kHz, 16 bit, mono WAV 형식으로 통일하였다. 또한 클리핑 발생 여부와 외부 소음 유입을 검토하였으며, 기준에 부합하지 않는 구간은 제외하거나 재녹음하였다. 이러한 검수 과정을 거쳐 최종 데이터베이스는 총 170개 오디오 샘플, 53분 39초 분량으로 확정되었으며, 샘플당 평균 길이는 19.3초(±9.5초)로 동작 조건에 따라 녹음 길이에 차이가 있다. 표 1은 구축된 데이터베이스의 주요 사양을 나타낸다.

※본 연구는 과학기술정보통신부의 재원으로 한국연구재단의 지원(NRF-2022R1A2C2010614) 및 행정안전부와 한국산업기술기획평가원(KEIT)의 지원(20026195)을 받아 수행되었음.

III. 시간-주파수 특성 분석

3.1 분석 방법 및 지표 정의

수집된 자가소음 데이터의 음향적 특성을 확인하기 위해 STFT 기반 시간-주파수 분석을 수행하였다. 사족보행 로봇의 자가소음은 모터 구동음, 관절 구동음, 지면 접촉에 따른 발자국 충격음이 함께 포함된 신호이므로, 시간 영역 파형만으로는 동작 상태에 따른 주파수 구조의 변화를 충분히 파악하기 어렵다. STFT는 48 kHz 신호에 대해 2048 크기 FFT를 표준 설정으로 적용하였으며, hop 길이는 512 샘플(약 10.67 ms)로 설정하였다. 이때 주파수 해상도는 약 23.4 Hz이다. 이를 통해 얻은 스펙트로그램으로 시간에 따른 주파수 성분의 변화를 확인하고, 협대역 피크 수, Spectral Flatness, Onset Rate를 중심으로 자가소음의 구조적 특성을 분석하였다.

협대역 피크 수는 특정 주파수 대역에 집중된 기계음 성분의 개수를 확인하기 위한 지표로, 스펙트럼에서 주변 대비 일정 수준 이상 솟아오른 국부 최대점을 피크로 검출하여 그 개수를 산출하였다. Spectral Flatness는 신호가 백색소음처럼 평탄한 구조인지 또는 특정 주파수 성분이 두드러지는 구조인지를 나타내는 지표로, 0에서 1 사이의 값을 가지며 1에 가까울수록 평탄한(잡음에 가까운) 스펙트럼을, 0에 가까울수록 특정 주파수에 에너지가 집중된 피크 중심 구조를 의미한다. Onset Rate는 단위 시간당 음향 이벤트의 발생 횟수로, 발자국 충격음과 같이 시간적으로 급격히 발생하는 성분의 빈도를 확인하기 위해 사용하였다.

3.2 전체 자가소음 특성

수집된 170개 자가소음 데이터를 분석한 결과, 사족보행 로봇의 자가소음은 단순한 순음 또는 백색소음 형태가 아니라 다수의 협대역 성분과 광대역 성분이 혼재된 복합적인 구조를 보였다. 100–8000 Hz 범위에서 협대역 피크는 평균 11.2개로 나타났으며, 이는 자가소음이 여러 개의 주파수 피크를 가지며, 각 피크가 명확히 분리되어 있음을 의미한다.

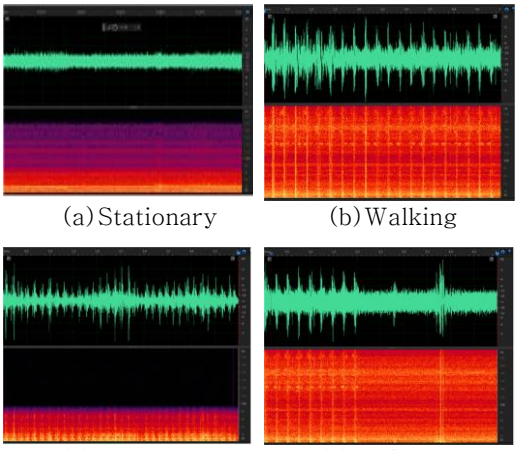

(a) Stationary (b) Walking
(c) Running (d) Sit/Stand

그림 3. 동작 조건별 자가소음의 파형 및 스펙트로그램

또한 Spectral Flatness는 평균 0.041로 낮게 나타났다. 이는 로봇 자가소음이 백색소음처럼 주파수 전 대역에 평탄하게 분포하는 소음이 아니라, 특정 주파수 대역에 에너지가 집중된 피크 중심 구조를 가진다는 것을 보여준다. Onset Rate는 평균 6.8회/sec로 나타났으며, 이는 로봇 구동 과정에서 발자국 충격음이나 구동계 동작 같은 반복적인 시간적 이벤트가 포함되어 있음을 의미한다.

이상의 세 지표를 종합하면, 사족보행 로봇의 자가소음은 단일한 모터 순음이나 일반적인 배경소음으로 설명하기 어렵다. 본 연구에서 수집된 자가소음은 여러 협대역 피크가 넓은 주파수 범위에 분포하고, 동작에 따라 충격성 성분이 함께 나타나는 다중 협대역-광대역 복합 구조를 갖는 것으로 확인되었다.

3.3 동작 조건별 자가소음 특성

동작 조건별 분석 결과, 자가소음의 음향 특성은 로봇의 동작 상태에 따라 뚜렷한 차이를 보였다. 본 절에서는 본체 동작 중 대표적인 네

※본 연구는 과학기술정보통신부의 재원으로 한국연구재단의 지원(NRF-2022R1A2C2010614) 및 행정안전부와 한국산업기술기획평가원(KEIT)의 지원(20026195)을 받아 수행되었음.

가지 조건(Stationary, Walking, Running, Sit/Stand)에 대한 파형과 스펙트로그램을 그림 3에 제시한다.

그림 3(a)의 Stationary 동작에서는 로봇이 이동하지 않는 상태에서도 시간축 전반에 걸쳐 지속적인 에너지가 유지되며, 스펙트로그램에서 수평 방향의 협대역 성분이 관찰된다. 이는 지면 접촉음보다는 모터 및 구동계에서 발생하는 지속성 기계음이 주요 성분임을 보여준다.

반면 그림 3(b)의 Walking 동작에서는 파형과 스펙트로그램에서 반복적인 에너지 변화가 나타나며, 이는 보행 과정에서 발생하는 관절 구동음과 지면 접촉에 따른 발자국 충격음이 주기적으로 결합된 결과로 해석된다.

그림 3(c)의 Running 동작에서는 빠른 발자국 접촉으로 인해 시간축에서 반복적인 에너지 변화가 더 조밀하게 나타나며, 보행 조건보다 충격성 이벤트의 발생 빈도가 증가하는 양상을 보인다.

그림 3(d)의 Sit/Stand 동작에서는 보행과 같은 주기적 패턴보다는 특정 시점에서 큰 에너지 변화가 나타나며, 이는 자세 전환 과정에서 여러 관절이 동시에 구동되면서 발생하는 비주기적 구동음 특성을 반영한다.

이러한 결과는 사족보행 로봇의 자가소음이 단순한 배경소음이 아니라, 동작 조건에 따라 다르게 결합되는 복합적인 시간-주파수 신호임을 보여준다.

IV. 결론

본 연구에서는 Unitree Go2 사족보행 로봇을 대상으로 자가소음 데이터베이스를 구축하고 시간-주파수 특성을 분석하였다. 데이터는 실내외 10종 환경, 6종 동작 조건, 4개 음향 수집 위치에서 수집되었으며, 최종적으로 총 170개 오디오 샘플과 53분 39초 분량으로 구성하였다. 분석 결과, 사족보행 로봇의 자가소음은 단순한 순음 또는 백색소음이 아니라 다중 협대역-광대역 복합 구조를 갖는 것으로 나타났다. 전체 샘플에서 100-8000 Hz 범위의 협대역 피크는 평균 11.2개, Spectral Flatness는 평균 0.041, Onset Rate는 평균 6.8회/sec로 측정되었다. 특히 보행 및 달리기 동작에서는 발자국 충격음에 의한 충격성 이벤트가 더 빈번하게 나타났다.

본 데이터베이스와 분석 결과는 향후 로봇 자가소음 제거 모델 학습 및 청각 기반 상황 인지 기술 개발을 위한 기초 자료로 활용될 수 있다.

참고문헌

[1] P. Biswal and P. K. Mohanty, “Development of quadruped walking robots: A review,” Ain Shams Engineering Journal, vol. 12, no. 2, pp. 2017–2031, 2021.

[2] C. Rasc□n and I. Meza, “Localization of sound sources in robotics: A review,” Robotics and Autonomous Systems, vol. 96, pp. 184–210, 2017.

[3] Z. Cao, B. Nie, Y. Zhang, and Y. Gao, “Minimizing acoustic noise: Enhancing quiet locomotion for quadruped robots in indoor applications,” arXiv preprint arXiv:2506.23114, 2025.

[4] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura, "Ego noise suppression of a robot using template subtraction," in Proc. IEEE/RSJ IROS, pp. 199–204, 2009.

[5] C. Valentini-Botinhao, “Noisy speech database for training speech enhancement algorithms and TTS models,” University of Edinburgh, School of Informatics, Centre for Speech Technology Research (CSTR), 2017.

[6] J. Thiemann, N. Ito, and E. Vincent, “The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings,” Proceedings of Meetings on Acoustics, vol. 19, no. 1, 035081, 2013.

※본 연구는 과학기술정보통신부의 재원으로 한국연구재단의 지원(NRF-2022R1A2C2010614) 및 행정안전부와 한국산업기술기획평가원(KEIT)의 지원(20026195)을 받아 수행되었음.

학사 규정 질의응답 모델의 오류 검토·교정 학습을 위한 합성 데이터셋 구축

박소영[O], 김선만[*]
한신대학교 AI·SW대학
e-mail:kong02931@hs.ac.kr[O], smkim@hs.ac.kr[*]

A Correction-Aware Synthetic Dataset for Training Language Models in Academic-Regulation Question Answering

So Yeong Park[O], Seon Man Kim[*]
School of Computing and Artificial Intelligence, Hanshin University

요 약

대학 학사 규정은 학생의 학업 이수와 학사 절차의 기준을 명시한 공식 문서이지만, 조항 간 관계가 복잡하고 예외 조건이 많아 사용자가 필요한 정보를 정확히 해석하기 어려운 경우가 있다. 또한 공지된 규정에 대해서도 유사 문의가 학기 단위로 반복되는 현상은, 정보 공개만으로는 학생의 정확한 규정 해석이 충분히 보장되지 않을 수 있음을 시사한다. 본 연구는 학사 규정 기반 질의응답 시스템 개발을 위한 선행 작업으로, 오류 검토 및 교정 절차를 포함한 학사 규정 합성 데이터셋 구축 과정을 정리한다. 특히 범용 대규모 언어 모델이 규정 원문과 다른 내용을 답변에 포함할 가능성을 고려하여, 모델 학습용 합성 데이터셋은 정답 응답과 함께 오류 응답에 대한 교정 예시를 포함하도록 구성하였다. 한신대학교 규정 관리집의 HWP 형식 문서를 기반으로 2종의 대규모 언어 모델을 활용한 자동 검토와 수작업 검토를 병행했으며, 8개 규정 영역에 대해 약 1만 건의 데이터를 구축했다. 추가적으로, 대규모 언어 모델의 zero-shot 정답·오류 응답 판별과 소형 언어 모델 대상 예비 파인튜닝을 수행하여 구축 데이터셋의 활용 가능성을 예비적으로 검토하였다. 예비 평가에서는 오류 검토 단계가 교정 성능을 좌우하는 병목임을 확인했다. 본 연구 결과는 담당자 검토를 거쳐 학생 안내를 보조하는 학사규정 질의응답 모델 개발로 확장될 수 있다.

키워드: 학사 규정 질의응답, 합성 데이터, 파인튜닝용 데이터셋, 소형 언어 모델, 도메인 특화 모델 학습

※본 연구는 과학기술정보통신부의 재원으로 한국연구재단 (NRF-2022R1A2C2010614)의 지원을 받아 수행되었음

I. 서론

대규모 언어모델(Large Language Model, LLM)은 문서 기반 질의응답과 행정 안내 등에 널리 활용되지만, 생성된 답변은 표현이 자연스럽더라도 근거 문서와 일치하지 않을 수 있다[1]. 특히 답변의 근거가 특정 도메인 문서로 한정되는 경우, 적용 대상·요건·수치·기한·예외 조건의 누락이나 왜곡은 잘못된 규정 적용으로 이어진다. 따라서 이러한 영역에서는 근거에 맞는 답변을 생성하는 능력뿐 아니라, 생성된 답변이 근거와 충돌하는지 검토하고 교정하는 능력이 함께 요구된다.

학사 규정은 학적 변동, 수강신청, 졸업, 장학 등을 판단하는 공식 기준이다. 그러나 조·항·호와 단서 조항에 따라 적용 조건이 달라지고, 동일한 질의라도 학생의 상황에 따라 답이 달라진다. 이처럼 규정이 공개되어 있어도 필요한 정보를 정확히 해석하기 어려워 유사한 문의가 반복된다. 한편 학사 행정 환경에서는 질의, 답변, 그 근거, 수정 과정이 학습에 쓸 형태로 축적되기 어렵다. 따라서 기존 상담 이력을 모으는 것만으로는 어떤 답변이 근거와 충돌하는지 판단하고 교정하는 과정을 학습하기에 충분하지 않다.

이에 본 연구는 학사 규정 도메인에서 언어모델이 근거에 맞는 답변 생성과 오류 응답 교정을 함께 학습하도록 데이터 구조를 설계하고, 파인튜닝으로 그 가능성과 한계를 평가한다. 구체적으로 정오 판단에 필요한 규정 정보를 조항 단위로 정리하고, 초안 답변·검토 결과·최종 답변이 대응되는 데이터를 구성하였다. 평가는 직접 답변 생성, 오류 검토·교정, 통합 추론의 세 조건에서 수행하였으며, 그 결과 교정의 병목이 답변을 작성하는 단계가 아니라 오류를 식별하는 검토 단계에 있음을 확인하였다.

II. 관련 연구

규정·법령 질의응답은 단일 조항의 사실 검색을 넘어, 여러 조항을 종합하고 단서 조건을 판단해야 한다는 점에서 일반 개방형 질의응답과 구분된다. LegalBench는 법률 질의응답을 여러 추론 유형으로 분류해, 이 영역의 난점이 사실 검색이 아니라 조건적 추론에 있음을 보였다[2]. 학사 규정 질의응답에서는 이러한 조건 의존성이 더 두드러져, 조항 위치·적용 대상·예외 조건에 따라 동일 질의의 정답이 달라진다.

검색 증강 생성은 외부 저장소에서 관련 조항을 검색해 생성의 사실성을 높이는 대표적 접근이다[3]. 그러나 규정 질의응답에서는 조항을 검색하는 것만으로 충분하지 않다. 답변은 적용 대상·기한·수치·예외 조건을 모두 반영해야 하며, 하나라도 누락·왜곡되면 근거와 불일치한 응답이 된다. 이때 조항을 올바르게 반영하는 능력은 조항을 검색하는 능력과 구분되므로, 검색과 별개의 학습 신호가 요구된다.

도메인 특화 학습에서는 고품질 지도 데이터를 확보하기 어렵다. Self-Instruct는 모델이 생성한 instruction-output 쌍을 선별해 활용하지만[4], 답변이 근거와 일치하는지 검증하는 구조는 없다. 추론을 단계적으로 생성하는 방식[5]이나 출력을 스스로 검토·수정하는 자기교정 방식[6]도 제안되었으나, 후자는 검토와 교정을 한 모델이 함께 수행하므로 규정 질의응답에서 오류 초안이 어떤 규정 조건과 충돌하는지 식별하는 학습 신호를 별도로 제공하지 못한다. 본 연구는 이 지점에 초점을 두어 검토와 교정을 분리하고, 오류 식별 능력을 독립적으로 평가함으로써, 자기교정 구조가 규정 도메인에서 한계를 보이는 지점을 진단한다.

III. 제안 방법

3.1 학사 규정 판단 정보 구조화

학사 규정 답변의 정합성은 조항 원문만으로 판정하기 어렵고, 각 조건이 답변에 어떻게 반영되었는지를 함께 확인해야 한다. 또한 정오 판정이 조항 단위로 추적되려면 판정에 필요한 조건이 명시적으로 분리되어 있어야 한다. 이를 위해 각 조항마다 정오 판정에 필요한 조건을

※본 연구는 과학기술정보통신부의 재원으로 한국연구재단 (NRF-2022R1A2C2010614)의 지원을 받아 수행되었음

표 1. 규정 판단 정보 구조화 예시 (휴학규정, 제29조 제3항)

필드	내용
규정명	학사에 관한 내규(2025.12.16. 기준)
조항 위치	제29조 제3항
적용 대상	수업일수 1/4선 경과 후 휴학을 신청하는 재학생
요건	학과장 의견서 제출 및 교무혁신처장의 허가
수치·기한	질병 사유는 4주 이상의 진단 필요
예외조건	특별사유에 의한 허가는 중간고사 이전 신청자에 한함

규정명, 조항 위치, 적용 대상, 요건, 수치·기한, 예외 조건의 여섯 항목으로 정규화한 규정 판단 정보를 구성하였다. 이 정보는 정답을 구성할 때는 충족 조건이 되고, 오류 응답을 구성할 때는 어떤 조건이 왜곡되었는지를 통제하는 기준이 된다. 단일 조항을 구조화한 예시는 표 1과 같다.

3.2 오류 검토·교정 학습 데이터 구성

학습 데이터는 PASS, FAIL, OOS의 세 라벨로 구성하였다. PASS는 정상 질의에 대해 모델이 규정과 일치하는 답변을 생성하고 유지하는 능력을 학습시키기 위한 사례로 답변 초안이 규정 판단 정보와 일치한다. FAIL은 근거와 충돌하는 오류가 있어 교정이 필요한 경우, OOS는 질의가 규정 범위를 벗어나거나 제공된 근거만으로 확정 답변을 내기 어려운 경우이다.

각 사례는 초안 답변, 검토 결과, 최종 답변의 세 요소로 구성하였다. 검토 결과는 초안을 유지·교정·보류할지 판단한 것으로, 오류 유형과 규정 근거와의 충돌 지점, 교정 방향을 함께 담는다. 최종 답변은 검토 결과를 반영해 산출한다. 각 질의는 언어모델로 생성하고, 정답과 검토 근거는 규정 원문을 기준으로 직접 작성하였다. 정답과 오류의 생성이 규정 원문과 규칙에 근거하므로, 정오 판정과 오류 삽입 내역은 조항 단위로 추적된다. FAIL의 오류 초안은 정답에서 단 하나의 조건만 위반하는 통제된 오류로 설계하여, 모델이 회복해야 할 조건을 분명히 하였다. 수치·적용 대상·조항 번호의 왜곡은 규칙적 치환으로 생성하고(예: 수강신청 상한 21학점을 19학점으로 변경), 요건 누락이나 예외 조건 오적용처럼 해석이 필요한 유형은 직접 작성하였다.

구축된 학습 데이터가 규정 근거와 라벨 구조에 맞게 구성되었는지 확인하기 위해, 자동 규칙 검증과 인적 검수를 수행하였다. 자동 검증 단계에서는 기준 답변에 기록된 근거 조항, 수치, 라벨이 원문 규정과 일치하는지 확인했다. 또한 초안 답변과 최종 답변의 역할이 명확히 구분되는지도 함께 점검했다. PASS 사례의 경우 최종 답변이 기준 답변과 일치하는지를 확인하였고, FAIL 사례의 경우 초안 답변에 포함된 오류가 최종 답변에서 규정 근거에 맞게 교정되었는지를 검토했다. 자동 검증을 통과한 사례에 대해서는 규정 원문과 기준 답변을 대조하는 인적 검수를 수행하였다. 라벨 적합성, 오류 유형, 최종 답변의 정오 판정이 엇갈린 사례는 원문 규정을 기준으로 재검토한 뒤 합의하여 결정했다.

구축된 총 9,958건의 데이터 중 PASS는 6,557건, FAIL은 2,914건, OOS는 487건이다. 해당 전체 데이터는 학습 8,355건, 검증 717건, 평가 886건으로 나누었다. 분할은 동일한 원 질의에서 파생된 PASS·FAIL·OOS 사례를 하나의 단위로 묶어 수행하였다. 이를 통해 원 질의에서 나온 사례가 서로 다른 분할에 동시에 포함되지 않도록 하였다.

※본 연구는 과학기술정보통신부의 재원으로 한국연구재단 (NRF-2022R1A2C2010614)의 지원을 받아 수행되었음

표 2. 평가 설정별 오류 검토·교정 성능

학습 포맷	검토신호	정확도	FAIL 교정률
직접 답변 생성	없음	0.8892	0.5464
오류 검토 교정	정답 검토	0.9938	0.9897
통합 추론	학습된 검토	0.8431	0.2990

3.3 검토·교정 모듈 구성 및 평가 설계

검토와 교정을 한 모델이 함께 수행하면 최종 응답의 오류가 검토 실패에서 비롯된 것인지, 교정 실패에서 비롯된 것인지 분리해 해석하기 어렵다. 본 연구에서는 검토와 교정은 별도 모듈로 구성하였다. 검토 모듈은 질의와 초안을 입력받아 오류 여부와 오류 문구, 교정 방향을 출력하고, 교정 모듈은 질의·초안·검토 결과를 입력받아 최종 답변을 출력한다. 교정 모듈은 Qwen2.5-7B-Instruct[7], 검토 모듈은 Llama-3.1-8B[8]를 각각 4bit 양자화 기반 QLoRA로 파인튜닝하였으며[9, 10], 주요 학습 파라미터는 동일하게 설정하였다. 후술하는 세 평가 조건(표 2)은 모두 동일한 교정 모듈(Qwen2.5-7B-Instruct)을 대상으로 하며, 교정 모듈에 주어지는 검토 신호의 출처만 달리한 것이다. 검토 모듈(Llama-3.1-8B)은 이 중 통합 추론 조건에서만 검토 신호 생성에 사용된다.

오류 식별과 교정 중 어느 단계가 성능을 제약하는지 진단하기 위해, 교정 모듈에 주어지는 검토 신호의 출처를 달리한 세 조건을 비교하였다. 세 조건은 모두 초안을 입력으로 받고 검토 신호에서만 차이가 난다. 직접 답변 생성은 검토 근거 없이 라벨만으로 교정하고, 오류 검토·교정은 오류 위치와 교정 방향이 정확히 주어진 정답 검토를 함께 제공하며, 통합 추론은 학습된 검토 모듈이 생성한 검토를 사용한다.

평가 지표는 전체 정확도와 FAIL 교정률이다. 전체 정확도는 PASS·FAIL·OOS 전 사례에서 최종 응답이 해당 라벨에 맞는 올바른 처리를 한 비율이다. 정오는 자동 판정하였으며, 기준 답변이 정해진 경우 완전일치로, 오류 교정이 필요한 경우 오류 구간의 복원 여부로 판정하였다. 세 조건 모두 초안을 받아 교정하므로, FAIL 교정률은 동일하게 오류 초안의 해당 구간이 규정 근거에 맞게 복원된 비율로 정의한다.

IV. 결과 및 분석

표 2는 세 조건의 평가 결과이다. 교정 성능은 검토 신호의 정확도에 따라 크게 갈렸다. 정답 검토가 주어진 오류 검토·교정 조건에서는 FAIL 교정률이 0.9897로 가장 높게 나타났다. 이는 검토 정확도가 충분히 확보될 경우, 교정 수행 능력이 효과적으로 발휘될 수 있음을 시사한다.

반면 검토 근거 없이 라벨만 쓰는 직접 답변 생성은 0.5464, 학습된 검토를 쓰는 통합 추론은 0.2990으로 가장 낮았다. 통합 추론에서도 PASS 유지(0.995)와 OOS 보류(1.000)는 안정적이어서, 저하는 FAIL 교정에 선택적으로 나타났다.

통합 추론 조건의 낮은 교정률은 교정 단계 자체보다 검토 단계의 불안정성에서 비롯된 것으로 해석된다. 통합 추론에서는 모델이 오류 초안을 스스로 검토한 뒤 교정까지 수행하므로, 검토 단계의 오류가 최종 답변에 그대로 누적되기 때문이다. 이를 직접 확인하기 위해 통합 추론 조건에서 검토 모듈의 출력을 분석한 결과, 검토 모듈은 오류 초안의 49.0%만 탐지하였고 탐지한 사례 중 오류 문구를 정확히 지목한 비율은 61.1%에 그쳤다. 이때 PASS 유지가 0.995로 높게 유지된 점은 잘못된 검토로 정상 답변을 과도하게 수정하는 과교정이 크지 않음을 의미하므로, FAIL 교정률 저하의 주원인은 과교정이 아니라 검토 단계의 낮은 오류 탐지율로 인한 교정 누락으로 분해된다. 직접 답변 생성의 전체 정확도(0.8892)가 동일 조건의 FAIL 교정률(0.5464)보다 높게 나타나는 것은 평가셋에서 PASS 사례가 다수를 차지하여 전체 정확도가 FAIL 교정 성능을 직접 반영하지 않기 때문이다.

※본 연구는 과학기술정보통신부의 재원으로 한국연구재단 (NRF-2022R1A2C2010614)의 지원을 받아 수행되었음

V. 결론

본 연구는 학사 규정 도메인에서 언어모델이 규정 근거에 부합하는 답변을 생성하고, 오류가 포함된 응답을 교정할 수 있도록 학습 데이터 구조를 설계하였다. 이를 위해 PASS, FAIL, OOS 사례를 구분하여 데이터셋을 구축하고, 직접 답변 생성, 오류 검토·교정, 통합 추론의 세 가지 조건에서 모델의 성능을 비교하였다.

실험 결과, 정답 검토가 입력으로 제공되는 오류 검토·교정 조건에서는 FAIL 교정률이 0.9897임을 확인하였다. 이는 오류 위치와 교정 방향이 명확히 주어질 경우, 모델이 규정 근거를 반영해 대부분의 오류 초안을 교정할 수 있음을 보여준다. 반면 학습된 검토 모듈의 출력을 사용하는 통합 추론 조건에서는 FAIL 교정률이 0.2990에 그쳤다. PASS 유지와 OOS 보류가 비교적 안정적으로 나타난 것과 달리, FAIL 교정 성능만 선택적으로 저하되었다는 점에서 통합 추론의 병목은 최종 답변 생성보다 검토 단계에 있는 것으로 해석된다.

세부 분석에서도 이러한 경향이 확인되었다. IV장에서 보였듯이, 통합 추론에서 검토 모듈은 오류 초안의 절반가량만 탐지하였고 탐지한 사례에서도 오류 문구 지목 정확도가 제한적이었다. 이는 통합 추론의 낮은 교정률이 검토 단계의 오류 탐지와 지목 실패에서 비롯되었음을 뒷받침한다. 이는 오류를 발견하고 특정하는 검토 신호가 충분히 확보되지 않으면, 모델의 교정 능력이 실제 추론 과정에서 제대로 발휘되기 어렵다는 점을 보여준다.

특히 통합 추론의 교정률이 직접 답변 생성 조건보다 낮게 나타나 검토 기반 교정 구조의 한계를 드러낸다. 부정확한 검토는 교정을 돕는 보조 신호로 작용하기보다, 오히려 불필요하거나 잘못된 수정을 유도할 수 있다. 따라서 오류 검토 단계는 전체 교정 성능과 신뢰성을 좌우하는 핵심 구성 요소로 보아야 한다.

본 연구는 검토 단계의 오류 식별 성능을 높이고 학습 파라미터 최적화를 거쳐 실제 학교 환경에서 학생 안내를 보조하는 학사 규정 질의응답 모델 개발로 확장될 수 있다.

참고문헌

[1] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, Article 248, 2023.

[2] N. Guha et al., "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models," Advances in Neural Information Processing Systems, 2023.

[3] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems, 2020.

[4] Y. Wang et al., "Self-Instruct: Aligning Language Models with Self-Generated Instructions," Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023.

[5] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Advances in Neural Information Processing Systems, 2022.

[6] A. Madaan et al., "Self-Refine: Iterative Refinement with Self-Feedback," Advances in Neural Information Processing Systems, 2023.

[7] Qwen Team, "Qwen2.5 Technical Report," arXiv preprint arXiv:2412.15115, 2024.

[8] A. Grattafiori et al., "The Llama 3 Herd of Models," arXiv preprint arXiv:2407.21783, 2024.

[9] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," International Conference on Learning Representations, 2022.

[10] T. Dettmers et al., "QLoRA: Efficient Finetuning of Quantized LLMs," Advances in Neural Information Processing Systems, 2023.

※본 연구는 과학기술정보통신부의 재원으로 한국연구재단 (NRF-2022R1A2C2010614)의 지원을 받아 수행되었음

공중화장실 안전위생 통합관제용 AI 환경센서 개발

저자 홍아름, 정우진, 김유신
한신대학교 빅데이터융합학과
e-mail: armihong00@airdeep.co.kr, cck0829@airdeep.co.kr, yoosin25@hs.ac.kr

IoT Sensor system for the public restroom security

Author
Areum Hong, Woojin Jeong, Yoosin Kim
Dept. of Big Data Convergence, Hanshin University

요 약

공중화장실은 불특정 다수가 이용하는 공공공간이면서도 사생활 보호가 요구되어 CCTV 설치가 제한되는 대표적인 안전관리 사각지대이다. 기존 공중화장실 안전관리 체계는 비상벨, CCTV, 인력 순찰, 단일 감지기 중심으로 운영되어 왔으나, 이용자의 직접 호출에 의존하거나 특정 이벤트만 감지하는 방식이어서 흡연, 악취·오염, 기물 파손, 폭력, 배회, 화재 등 다양한 안전·위생 문제를 통합적으로 판단하는 데 한계가 있다. 이에 본 연구에서는 공기질 센서(PM, TVOC 등), 음향 센서, 재실 및 동작감지 센서 등 비식별 복합 센서 데이터를 활용하고, Edge AI 기반 온디바이스 분석 구조를 적용한 공중화장실 안전위생 통합관제용 AI 환경센서 시스템을 제안한다. 제안 시스템은 현장 단말에서 수집 데이터를 실시간으로 분석하여 흡연 및 전자담배, 악취·오염, 파손·폭력, 배회, 화재 등 이상상황을 판단하고, 관리자 앱 및 웹 기반 통합관제 시스템과 연계하여 실시간 알림과 운영 현황 모니터링을 지원한다.

키워드: 공중화장실, 흡연탐지, 환경센서, IoT, AI

I. 서론

최근 학교, 공공시설, 공동주택, 병원, 사무공간 등 다중이용시설을 중심으로 실내 안전과 위생 관리에 대한 요구가 증가하고 있다. 특히 공중화장실은 불특정 다수가 이용하는 공간이면서도 사생활 보호가 요구되어 CCTV 설치가 제한되는 대표적인 관리 사각지대이다. 이러한 공간에서는 흡연, 악취 및 오염, 기물 파손뿐 아니라 폭력, 성범죄 등 안전사고가 발생할 수 있으나, 영상 감시 기반의 기존 안전관리 체계만으로는 이를 상시 감지하거나 즉각 대응하는 데 한계가 있다.

공중화장실에서의 안전 문제 역시 지속적으로 제기되고 있다. 2016년 강남역 공중화장실 살인사건[1], 2024년 부산역 화장실 폭행사건[2], 2025년 여중생 집단 성폭행 사건[3] 등은 화장실과 같은 공용공간에서도 강력범죄가 발생할 수 있음을 보여준다. 그러나 기존 비상벨 기반 시스템은 이용자가 직접 버튼을 누르거나 음성으로 구조를 요청해야 작동하는 방식이므로, 실제 위급상황에서 피해자가 즉시 신고하기 어려운 경우에는 골든타임 내 대응이 제한될 수 있다. 또한 기존 IoT 안전관리 시스템은 클라우드 또는 서버 중심 처리 구조에 의존하는 경우가 많아, 통신 장애나 네트워크 지연 상황에서 실시간 감지 및 대응 성능이 저하될 수 있다.

따라서 화장실과 같이 영상 감시가 제한되는

공간에서는 사생활 침해 우려 없이 이상상황을 감지할 수 있는 비영상 기반 지능형 안전관리 기술이 필요하다. 본 연구에서는 PM, TVOC 등 공기질 센서와 음향 센서, 재실 및 동작감지 센서 등 복합 센서 데이터를 하나의 디바이스에 통합하고, Edge AI 기반의 온디바이스 판단 구조를 적용한 공중화장실 안전위생 통합관제용 AI 환경센서 개발 방안을 제안한다. 제안 시스템은 흡연 및 전자담배, 파손·폭력, 성폭행, 배회, 악취·오염 등 다양한 위험 및 위생 상황을 비영상 방식으로 감지하고, 현장에서 즉시 분석·판단하여 관리자에게 실시간 알림을 제공하는 것을 목표로 한다. 이를 통해 기존 CCTV 및 비상벨 중심의 사후 대응형 관리 체계를 보완하고, 공중화장실의 안전·위생 문제를 사전 예방형 관리 체계로 전환할 수 있는 가능성을 제시하고자 한다.

II. 관련 연구

2.1 국내 동향

국내 공중화장실 및 공공시설 안전관리 분야에서는 기존의 비상벨, CCTV, 인력 순찰 중심의 사후 대응 방식에서 벗어나 IoT 기반 안전장치와 음원감지 기술을 결합한 지능형 안전관리 방식으로 점차 고도화되고 있다. 특히 화장실, 탈의실 등 사생활 보호가 요구되는 공간은 CCTV 설치가 제한되기 때문에, 영상정보를 사용하지 않고 이상상황을 감지할 수 있는 비영상 기반 센서 기술의 필요성이 증가하고 있다.

그러나 2020년 이후 국내 공중화장실 안전관리 분야의 도입 사례를 살펴보면, 현재 확산 중인 서비스는 비상벨, 이상음원 감지, IoT 연계형 안전장치 중심으로 구성되는 경향이 있다. 행정안전부는 공중화장실의 안전한 사용환경 조성을 위해 비상벨 등 안전관리시설 설치를 포함한 선도사업을 추진하고 있으며, 지자체에서도 스마트 IoT 비상벨, AI 안심비상벨, 음원감지기 등을 설치하는 사례가 확대되고 있다[4]. 예를 들어, 의왕시는 공중·개방화장실에 스마트 IoT 비상벨을 설치하여 화장실 내 비명소리 등 이상 음원이 감지되거나 이용자가 버튼을 누르는 경우 관할 경찰서 112상황실과 자동 연계되도록 운영하고 있다[5]. 또한 여수시는 공중화장실 79개소에 IoT 기반 AI 안심비상벨 및 음원감지기를 설치하여 공중화장실 범죄 예방과 안전관리 기능을 강화하고 있다[6]. 이러한 사례는 국내 공중화장실 안전관리 서비스가 현재 비상 호출, 이상음원 감지, 경찰 신고 연계 중심으로 도입되고 있음을 보여준다.

다만 이러한 방식은 위급상황 발생 시 이용자의 직접 호출에 의존하거나, 특정 음원 또는 단일 이벤트를 중심으로 위험상황을 판단하는 구조에 가깝다. 또한 비상벨 및 음성인식 기반 안전장치가 설치되어 있더라도 유지관리 상태와 감지 기준에 따라 실제 위급상황에서 정상 작동이 보장되지 않을 수 있다. 실제로 경기도의 공중화장실 비상벨 유지관리 실태 점검에서는 비상벨 작동 불량, 경찰 또는 관리기관 미연결, 이상음원 미감지, 설치 장소 부적정 등 다수의 부적합 사례가 확인된 바 있다[7]. 이는 단순히 비상벨이나 음원감지기를 설치하는 것만으로는 공중화장실 내 다양한 안전·위생 문제에 대한 실시간 대응성을 충분히 확보하기 어렵다는 점을 보여준다. 따라서 흡연, 전자담배, 악취, 위생오염, 기물 파손, 폭력 상황 등 다양한 위험요인을 하나의 시스템에서 통합적으로 판단하기 위해서는 공기질·음향·동작 등 복합 센서 데이터를 기반으로 이상상황을 감지할 수 있는 비영상 AIoT 안전관리 기술이 필요하다.

최근 기술 발전 방향은 단일 센서 기반 감지에서 공기질, 입자상물질, TVOC, 소음, 재실, 동작, 조도 등 다양한 데이터를 함께 분석하는 멀티센서 융합형 지능형 IoT 서비스로 이동하고 있다. Kim 등은 14종 환경센서를 활용한 지능형 Edge Computing 기반 실내 다중환경 센서 시스템을 제안하고, PM2.5, TVOC, CO 등 실내공기질 요소를 예측하는 구조를 제시하였다[8]. 이는 수집 데이터를 클라우드로 전송한 뒤 분석하는 방식에서 나아가, 현장 단말에서 환경 데이터를 수집·분석·예측하는 Edge AI 구조가 실내 안전·환경 모니터링 분야에서 중요한 기술 방향으로 활용될 수 있음을 보여준다.

따라서 국내 공공화장실 안전관리 기술은 비상 호출 및 음원감지 중심의 부분적 안전장치에서, 비영상 멀티센서와 Edge AI를 활용한 통합 안전·위생 관리 서비스로 확장될 필요가 있다. 본 연구에서 제안하는 AI 환경센서는 PM, TVOC 등 공기질 센서와 음향, 재실, 동작감지 센서를 복합적으로 활용하여 흡연, 전자담배, 악취·오염, 파손·폭력, 배회 등 다양한 상황을 통합적으로 감지하는 것을 목표로 한다. 이는 기존 비상벨 및 단일 감지기 중심의 사후 대응형 시스템을 보완하고, 공중화장실 안전관리를 사전 예방 및 실시간 대응 중심으로 전환할 수 있는 기술적 접근이다.

표1. 기존 공중화장실 안전관리 기술과 제안 기술 비교

구분	기존 기술의 한계	제안 기술의 특징
비상벨	이용자가 직접 호출해야 작동하는 수동 대응 방식	센서 기반 이상상황 자동 감지 및 관제 연동
CCTV	화장실 등 사생활 보호 공간에는 설치 제한	카메라 없이 비식별 센서 데이터 기반 감지
단일 감지기	흡연, 악취, 비상상황 등 개별	공기질·음향·재실·동작 센서 기반 통

	기능 중심	합 감지
서버 중심 IoT	네트워크 장애 시 대응 지연 가능	Edge AI 기반 현장 분석 및 즉시 판단
인력 순찰	상시 감시가 어렵고 사후 대응 중심	실시간 모니터링 및 데이터 기반 관리 가능

III. 제안방법

본 연구에서는 CCTV 설치가 제한되는 공중화장실 환경에서 안전 및 위생 이상상황을 비영상 방식으로 감지하기 위해, 복합 환경센서와 온디바이스 AI를 결합한 AIoT 기반 통합관제 시스템을 제안한다. 제안 시스템은 공기질 센서, 음향 센서, 재실 및 동작감지 센서를 단일 디바이스에 통합하고, 현장 단말에서 수집된 데이터를 Edge AI 기반으로 분석하여 흡연, 전자담배, 악취·오염, 파손·폭력, 배회, 화재 등 다양한 위험상황을 실시간으로 판단하는 구조를 갖는다.

제안 시스템은 크게 센서 데이터 수집부, Edge AI 분석부, AIoT 디바이스부, 통합관제 시스템부, 알림 및 연계부로 구성된다. 센서 데이터 수집부는 PM, TVOC 등 공기질 데이터와 음향, 재실, 동작감지 데이터를 수집한다. Edge AI 분석부는 수집된 데이터를 기반으로 공중화장실 내 이상상황을 실시간으로 분석하고 이벤트 유형을 분류한다. AIoT 디바이스부는 네트워크 연결 상태와 관계없이 현장 단말에서 1차 판단이 가능하도록 경량 AI 모델을 탑재한다. 통합관제 시스템부는 시설별 이벤트 발생 이력, 빈도, 시간대별 위험도, 운영 현황을 관리하며, 알림 및 연계부는 관리자 앱, 웹 대시보드, 보안·신고 체계와 연동하여 실시간 대응을 지원한다.

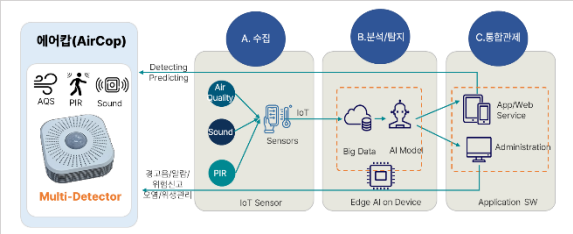
핵심 기능 중 하나인 흡연탐지 기능은 PM, TVOC 등 공기질 센서 데이터의 변화 패턴을 기반으로 수행된다. 일반적으로 흡연 및 전자담배 사용 시 미세먼지와 휘발성유기화합물 농도 변화가 발생하므로, 단일 시점의 센서값뿐 아니라 일정 시간 동안의 증가량, 지속 시간, 회복 패턴을 함께 고려한다. 특히 화장실 환경에서는 향수, 방향제, 청소용품, 스프레이 등으로 인해 흡연과 유사한 공기질 변화가 발생할 수 있으므로, 단순 임계값 기반 판단만으로는 오탐 가능성이 존재한다. 이에 본 연구에서는 흡연 이벤트와 비흡연 이벤트의 센서 변화 패턴을 구분하고, 연초 및 전자담배 유형까지 분류할 수 있는 AI 기반 판단 구조를 적용한다.

흡연탐지 모델은 공기질 데이터 수집, 전처리, 라벨링, 기초 통계 분석, 모델 학습 및 검증 절차를 통해 구성된다. 수집 데이터는 흡연/비흡연 상태와 연초/전자담배 유형으로 구분하여 라벨링하고, PM, TVOC 등 주요 센서값의 변화 특성을 분석한다. 이후 지도학습 기반 분류 모델을 활용하여 흡연 여부를 판단하며, 모델 성능은 정확도, 정밀도, 재현율, F1-score 등을 기준으로 평가한다. 선행 흡연탐지 연구에서는 PM2.5, PM10, tVOC 등의 공기질 센서값이 흡연/비흡연 구분에 유의미한 차이를 보였으며, SVM 등 기계학습 기반 분류 모델을 통해 95% 이상의 흡연탐지 정확도를 확보할 수 있음을 확인하였다[9].

또한 제안 시스템은 흡연탐지 외에도 음향 및 동작 기반 이상상황 감지 기능을 포함한다. 비명, 충격음, 파손음 등 이상 음향 패턴과 재실·동작 변화 데이터를 함께 분석하여 폭력, 파손, 배회 등 위험상황 발생 가능성을 판단한다. 이때 카메라 영상이나 음성 녹음과 같은 식별성 데이터를 사용하지 않고, 비식별 센서 데이터의 패턴만을 활용함으로써 사생활 침해 우려를 최소화한다. 감지된 이벤트는 현장 디바이스에서 1차 분석된 뒤 관리자에게 실시간 알림으로 전송되며, 필요 시 통합관제 시스템 및 외부 대응 체계와 연동될 수 있다. 제안 시스템의 주요 기능은 아래와 같다.

- **흡연 탐지 서비스:** 공기질 센서(PM, TVOC 등)를 활용하여 공중화장실 내 흡연(전자 담배 포함) 여부를 실시간 감지
- **위생상태 감지 서비스:** 공기질 및 환경 변화 데이터를 기반으로 악취 및 오염 상태를 감지하여 위생 상태를 관리
- **이상행동 감지 서비스:** 배회, 장시간 체류 등 비정상적인 행동 패턴을 감지하여 위험 가능성을 사전에 인지
- **위급상황 감지 서비스:** 파손, 폭력, 충격 등 위급 상황을 음향 및 환경 데이터를 통해 감지
- **엣지AI 기반 분석 서비스:** 디바이스 내 온디바이스 AI를 활용하여 센서 데이터를 실시간 분석하고 이벤트를 판단
- **실시간 알림 및 통합관제 서비스:** 위험 상황 발생 시 관리자에게 실시간 알림을 제공하고, 통합 관제 시스템을 통해 상태를 모니터링
- **데이터기반 분석 및 운영지원서비스:** 수집된 데이터를 기반으로 위험 패턴 부석 및 운영효율화를 지원

그림1. 제안 AI 환경센서 제품 및 시스템 구성도



IV. 결론

본 연구에서는 CCTV 설치가 제한되는 공중화장실 환경에서 안전 및 위생 이상상황을 비영상 방식으로 감지하기 위한 AI 환경센서 기반 통합관제 시스템을 제안하였다. 제안 시스템은 PM, TVOC 등 공기질 센서와 음향, 동작, 재실, 조도 등 비영상 센서 데이터를 복합적으로 활용하여 흡연, 전자담배, 악취·오염, 파손·폭력, 배회, 화재 등 다양한 위험 및 위생 상황을 실시간으로 감지하며, Edge AI 기반 온디바이스 분석 구조를 통해 서버 중심 처리 방식의 응답 지연과 네트워크 의존성을 완화하고 현장 단말에서 즉각적인 1차 판단이 가능하도록 하였다.

이를 통해 제안 시스템은 이용자의 직접 호출에 의존하는 비상벨, 사생활 보호 공간에 설치가 제한되는 CCTV, 특정 이벤트만 감지하는 단일 감지기 중심의 기존 안전관리 방식이 가진 한계를 보완할 수 있다. 또한 카메라 영상이나 음성 녹음 없이 비식별 센서 데이터만으로 흡연·악취·화재·위급상황 등을 자동 감지함으로써 사생활 침해 우려를 줄이는 동시에, 이벤트 발생 시간·위치·유형·빈도 등의 데이터를 축적하여 시설별 위험도 분석과 순찰·청소 우선순위 도출 등 데이터 기반의 사전 예방형 관리체계로 전환하는 데 기여할 수 있다.

다만 본 연구는 시스템의 구조와 핵심 기능을 설계 수준에서 제안한 단계로, 실제 공중화장실 환경에서의 장기 운영 및 실증 데이터를 아직 확보하지 못했다는 한계를 가진다. 따라서 향후 연구에서는 환경 특성이 상이한 다양한 공중화장실에 시범 설치하여 장기 운영 데이터를 수집하고, 이를 바탕으로 탐지 정확도, 오탐률, 응답속도, 운영 안정성을 정량적으로 검증할 필요가 있다.

특히 공중화장실은 짙은 방향제, 향수, 락스 등 청소용품에서 발생하는 휘발성유기화합물 및 입자상물질로 인해 흡연과 유사한 센서 변화가 빈번하게 나타나므로, 오탐(False Positive) 저감이 시스템 신뢰성 확보의 핵심 과제가 된다. 그러나 선행 연구의 흡연탐지 모델은 단일 시점의 데이터를 기반으로 하여 흡연의 시계열적 변화 특성을 충분히 반영하지 못하였고, 차량과 노이즈 구성이 상이한 공중화장실 환경에 대한 검증도 이루어지지 않았다. 따라서 향후 연구에서는 기존의 트리 기반 및 SVM 모델과 더불어 CNN, RNN(LSTM) 등 시계열 패턴을 반영한 경량 딥러닝 모델을 함께 구축하고, 청소용품 사용이 혼재된 공중화장실 환경에서 수집한 데이터를 기반으로 각 모델의 정확도, 정밀도, 재현율, F1-score를 정량적으로 비교·분석하여 환경에 최적화된 탐지 모델을 선정하고자 한다. 나아가 통합관제 시스템 및 외부 신고·보안 연계 체계와의 실증을 통해 학교, 역사, 터미널 등 다양한 다중이용시설로 확장 가능한 AIoT 안전위생 관리 서비스로 발전시킬 수 있을 것으로 기대된다.

참고문헌

[1] SBS 뉴스, “[리포트+] ‘강남역 살인 사건’ 1년…공중화장실 이제는 안전할까?”, https://news.sbs.co.kr/news/endPage.do?news_id=N1004199828, 2017.05.17.
[2] 부산지방법원, “[형사] 부산역 여자화장실에서 일면식도 없는 피해자를 폭행하여 살인미수 범행을 저지른 피고인에게 징역 12년과 10년간 위치추적 전자장치 부착을 명한 사례”, 우리법원 주요판결, 우리법원 주요판결 - 상세보기 | 부산지방법원, 2024.07.18.
[3] 데일리안, “공중화장실서 14살 여중생 집단 성폭행, 가해자들은 뻔뻔했다”, 공중화장실서 14살 여중생 집단 성폭행, 가해자들은 뻔뻔했다, 2026.01.05.
[4] 행정안전부, “안전하고 편리한 사용환경 조성”, 공중화장실 정책자료, 업무안내, 행정안전부> 업무안내> 자치혁신실> 안전하고 편리한 공중화장실> 안전하고 편리한 사용환경 조성, 2024.
[5] 보안뉴스, “의왕시, 공중화장실에 스마트 IoT 비상벨 설치”, 의왕시, 공중화장실에 스마트 IoT 비상벨 설치, 2023.08.22.
[6] 여수시, “여수시, 공중화장실 범죄 예방 위한 스마트 안전관리 강화”, 보도자료, 여수시, 공중화장실 범죄 예방 위한 스마트 안전관리 강화 < (전체) < 여수주요소식 < 여수소식 < 열린시정 - 여수시청, 2026.01.22.
[7] 연합뉴스, “전원 꺼지고 비명 감지 못하고…공중화장실 비상벨 ‘관리비상’”, 전원 꺼지고 비명 감지 못하고…공중화장실 비상벨 '관리비상' | 연합뉴스, 2024.01.11.
[8] M. Kim, T. Kim, S. Park, and K. Lee, “An Indoor Multi-Environment Sensor System Based on Intelligent Edge Computing,” Electronics, vol. 12, no. 1, 137, 2023.
[9] 김유신, “공기질 IoT 센서를 이용한 차량 흡연탐지 시스템,” Journal of Internet Computing and Services, vol. 26, no. 1, pp. 55-63, 2025.

인공지능 역량은 기대와 우려를 동시에 높이는가?: 일반국민과 정보취약계층 6개 집단 비교

이명하*O
*O순천향대학교 교육혁신원
e-mail:shiftcraft0142@gmail.com *O

Does AI Competency Raise Both Expectations and Concerns?: A Comparison of the General Population and Six Digitally Vulnerable Groups

Lee, MyeongHa*O
Institute for Education Innovation, Soonchunhyang University*O

요 약

디지털 역량은 흔히 기술 수용을 촉진하고 불안을 완화하는 요인으로 여겨진다. 그러나 「2024 디지털정보격차 실태조사」의 일반국민과 6개 정보취약계층 균형표본(N=4,900)을 분석한 결과, 이러한 관점이 AI 기술 인식에는 그대로 적용되지 않을 수 있는 것으로 나타났다. 먼저 AI 기술에 대한 기대와 우려는 서로 상충하기보다 정적 상관(r=.665)을 보이며 공존하였다. 나아가 디지털기술 태도와 AI 서비스 이용경험을 통제한 뒤에도, AI 역량이 높을수록 기대와 우려를 동시에 갖는 양가성이 유의하게 증가하였다(β=.375, p<.001). 이 효과는 모든 집단에서 나타났으나 장애인(β=.505)에서 가장 크고 북한이탈주민(β=.234)에서 가장 작아, 그 영향이 집단에 따라 다르게 나타났다. 요컨대 AI 역량은 불안을 단순히 낮추기보다 기대와 우려를 함께 정교하게 인식하는 '정보화된 양가성'과 관련되며, 정보취약계층 디지털 정책은 활용 교육과 함께 신뢰·안전·위험 이해를 다루어야 한다.

키워드: 인공지능 역량, 양가적 태도, 기대와 우려, 정보취약계층, 디지털정보격차 실태조사

I. 서론

인공지능(AI)이 빠르게 확산되면서, AI를 어느 정도 활용하고 어떻게 인식하는가가 디지털 정보격차의 새로운 쟁점으로 부각되고 있다. 기존 정보격차 연구는 주로 접근과 활용 능력의 차이에 주목해 왔으며, 정보 접근성이 높을수록 격차가 줄고 활용 능력이 높아지는 것으로 보고되었다[1]. 이는 디지털 능력을 키우면 기술에 대한 불안도 줄어든다는 생각으로 이어지기 쉽다. 그러나 AI는 편리함과 함께 개인정보 유출, 차별, 오남용 같은 우려도 불러온다. 실제로 일반 대중은 AI에 대해 기대와 우려를 동시에 갖는 것으로 보고된다[2]. 이처럼 한 대상에 긍정과 부정 감정이 함께 강하게 나타나는 상태를 양가성(ambivalence)이라 하며, 두 감정이 모두 강하고 비슷할수록 커진다[3]. 주목할 점은 AI를 잘 아는 사람일수록 장점과 위험을 모두 더 구체적으로 인식할 수 있다는 것이다. AI에 대한 지식과 지식에 대한 확신이 AI의 위험·편익 지각과 관련된다는 연구[4]는, AI를 아는 정도가 AI 인식을 단순히 긍정 또는 부정의 한 방향으로만 이끌지 않을 수 있음을 시사한다. 다만 AI 역량이 기대와 우려를 동시에 높이는지는 직접 검토된 바 없다. 본 연구는 이러한 가능성을 '정보화된 양가성'으로 명명하고 실증적으로 검토한다. 특히 정보취약계층은 AI 확산의 혜택과 위험에 동시에 노출되어 있어, AI 역량이 태도에 미치는 영향이 집단에 따라 다르게 나타날 가능성이 있다. 이에 본 연구는 「2024 디지털정보격차 실태조사」를 활용하여 AI 역량이 양가적 태도에 미치는 영향과 그 집단 간 차이를 비교한다. 연구질문은 다음과 같다. 첫째, AI에 대한 기대와 우려는 함께 나타나는가? 둘째, AI 역량은 양가성을 높이는가? 셋째, 그 영향은 집단에 따라 어떻게 다른가?

II. 연구방법

본 연구는 한국지능정보사회진흥원의 「2024 디지털정보격차 실태조사」 원자료를 활용하였다. 일반국민과 6개 정보취약계층(고령층, 장애인, 저소득층, 농어민, 결혼이민자, 북한이탈주민)을 대상으로 하되, 700명을 초과하는 집단은 무작위추출하고, 700명인 집단은 전수를 사용하여 총 4,900명의 균형표본을 구성하였다.

독립변수인 AI 역량은 AI 기기 식별·활용·평가 등을 측정하는 12개 문항(Cronbach's α=.945)으로 측정하였으며, 점수가 높을수록 역량이 높음을 의미한다. 종속변수인 양가성은 조사의 AI 인식 문항 중 긍정 인식 5개와 부정 인식 4개를 각각 기대(α=.846)와 우려(α=.803)로 명명하여 결합한 것이다. 양가성은 Thompson 등[3]의 방식에 따라 기대와 우려의 평균에서 두 값의 차이의 절대값을 뺀 값으로 산출하였으며, 두 인식이 모두 높고 서로 비슷할수록 큰 값을, 한쪽으로 치우칠수록 작은 값을 가진다. 통제변수로는 디지털기술 태도(4개 문항, α=.844)와 AI 서비스 이용경험(이용=1, 미이용=0)을 포함하였다.

분석은 상관분석과 위계적 회귀분석으로 진행하였다. 회귀분석은 1단계에 통제변수를, 2단계에 AI 역량을 투입하였다. 이어 그 효과가 집단에 따라 어떻게 다른지 비교하기 위해 일반국민과 6개 정보취약계층 각각에 대해 동일한 회귀분석을 실시하였다. 모든 분석에는 SPSS를 이용하였다.

III. 연구결과

기술통계를 살펴보면 4점 척도에서 AI 역량의 평균은 2.55(SD=0.65), 기대는 2.85(SD=0.56), 우려는 2.85(SD=0.63), 양가성은 2.49(SD=0.64)였으며, 주요 변수의 왜도와 첨도는 일반적으로 수용 가능한 범위에 있었다. 먼저, AI 기술에 대한 기대와 우려는 유의한 정적 상관을 보여(r=.665, p<.001), 두 인식이 상충하기보다 함께 나타남을 확인하였다. AI 역량 또한 기대(r=.556), 우려(r=.473), 양가성(r=.482)과 모두 유의한 정적 상관을 보였다(p<.001). AI 역량이 양가성에 미치는 영향을 위계적 회귀분석으로 검토한 결과, 통제변수만 투입한 1단계의 설명력은 17.0%였으나 AI 역량을 추가한 2단계에서 26.9%로 증가하였다(ΔR^2=.099, p<.001).

통제변수를 투입한 뒤에도 AI 역량은 양가성에 유의한 정적 영향을 미쳤으며(β=.375, p<.001), 다중공선성 문제는 없었다(VIF≤1.42). 즉 디지털기술 태도와 AI 서비스 이용경험을 통제한 후에도 AI 역량이 높을수록 기대와 우려를 동시에 더 크게 갖는 것으로 나타나, 이는 AI 역량이 기술에 대한 우려를 단순히 낮추는 방향으로만 작용하지 않을 수 있음을 보여준다. 집단별 분석에서 AI 역량의 효과는 모든 집단에서 유의하였으나(p<.001), 효과의 크기는 집단별로 다르게 나타났다(표 1). AI 역량의 표준화 회귀계수는 장애인에서 가장 컸고(β=.505), 북한이탈주민에서 가장 작았다(β=.234). 이는 AI 역량과 양가성의 관계가 모든 집단에서 공통적으로 나타났으나, 그 효과의 크기는 집단별 특성에 따라 달라질 수 있음을 시사한다.

표 1. 집단별 AI 역량이 양가성에 미치는 효과

집단	β	ΔR^2
장애인	.505	.197
저소득층	.430	.141
결혼이민자	.417	.124
농어민	.330	.067
일반국민	.324	.082
고령층	.275	.059
북한이탈주민	.234	.042

주. 모든 계수는 p<.001 수준에서 유의함.

IV. 결론

본 연구는 AI 역량이 AI에 대한 양가적 태도에 미치는 영향을 일반국민과 6개 정보취약계층을 대상으로 비교하였다. 분석 결과 AI 역량이 높을수록 기대와 우려가 함께 크게 갖게 되는 '정보화된 양가성'이 모든 집단에서 확인되었으며, 그 정도는 장애인에서 가장 크고 북한이탈주민에서 가장 작았다. 이는 디지털 능력이 기술에 대한 불안을 반드시 낮추는 방향으로만 작용하지 않으며, 역량이 높을수록 기대와 우려를 함께 정교하게 인식할 수 있음을 시사한다. 따라서 정보취약계층 대상 디지털 포용 정책은 활용 교육뿐 아니라 신뢰·안전·위험 이해를 함께 다루고, 집단별로 효과의 크기가 다른 만큼 차별화된 지원 전략을 마련할 필요가 있다. 다만 본 연구는 횡단자료에 기반해 인과관계를 단정하기 어렵다는 한계가 있으며, 향후 종단자료를 통한 검증과 집단 간 차이의 원인에 대한 후속 연구가 요구된다.

참고문헌

[1] 권선희, "디지털 정보격차에 대한 취약계층의 계층별 분석," 지식경영연구, Vol. 23, No. 4, pp. 1-19, Dec. 2022. DOI: 10.15813/kmr.2022.23.4.001.

[2] 김길수, "인공지능의 신뢰에 관한 연구," 한국자치행정학보, Vol. 34, No. 3, pp. 21-41, Sep. 2020. DOI: 10.18398/kjlgas.2020.34.3.21.

[3] M. M. Thompson, M. P. Zanna, and D. W. Griffin, "Let's not be indifferent about (attitudinal) ambivalence," in Attitude Strength: Antecedents and Consequences, R. E. Petty and J. A. Krosnick, Eds. Lawrence Erlbaum, pp. 361-386, 1995. DOI: 10.4324/9781315807041-14.

[4] N. Said, A.-E. Potinteu, I. R. Brich, J. Buder, H. Schumm, and M. Huff, "An artificial intelligence perspective: How knowledge and confidence shape risk and benefit perception," Computers in Human Behavior, Vol. 149, 107855, Jun. 2023. DOI: 10.1016/j.chb.2023.107855.

군집분석을 활용한 공공 흡연시설 설치 우선지역 선정 : 서울시 사례를 중심으로

하시윤[O], 백수진[*]
[O]한신대학교 AI시스템반도체학과
[*]한신대학교 AI시스템반도체학과
e-mail: tofumuckbang@hs.ac.kr[O], croso79@hs.ac.kr[*]

Selection of Priority Area for Public Smoking Facilities Using Cluster Analysis : A Case Study of Seoul

Siyoon Ha[O], Sujin Baek[*]
[O]Dept. of AI System Semiconductor Engineering, Hanshin University
[*]Dept. of AI System Semiconductor Engineering, Hanshin University

요 약

최근 금연구역 확대 정책에 따라 공공 흡연시설의 효율적 입지 선정 중요성이 증가하고 있다. 이에 본 연구에서는 서울시 S-DoT 보행량 데이터, 금연구역 데이터 및 공공 흡연시설 데이터를 활용하여 공공 흡연시설 설치 우선지역을 도출하였다. 방문자수, 금연구역수, 금연구역 면적 합, 흡연시설수 및 흡연시설 부족도를 기반으로 분석 데이터셋을 구축하고, K-Means, 계층적 군집분석 및 DBSCAN을 적용하여 군집화 성능을 비교하였다. 분석 결과 계층적 군집분석이 가장 우수한 성능을 나타냈으며, K-Means는 유사한 수준의 성능과 높은 해석 가능성을 보였다. 이에 K-Means 결과를 최종 분석에 활용하였으며, 양천구 목3동, 관악구 신원동, 양천구 신월1동, 송파구 잠실본동 등이 공공 흡연시설 설치 우선 검토 지역으로 도출되었다. 본 연구는 공공데이터 기반의 군집분석을 통해 공공 흡연시설 설치 우선지역을 객관적으로 제시하였다는 점에서 의의를 가진다.

키워드: 공공데이터 분석, 흡연시설, 군집분석, K-Means, 입지선정

I. 서론

최근 금연구역 확대 정책이 지속적으로 추진됨에 따라 공공 흡연시설의 적정 입지 선정에 대한 중요성이 증가하고 있다. 그러나 공공 흡연시설은 지역별 수요를 고려한 객관적 기준보다 행정적 판단에 의해 설치되는 경우가 많아 효율적인 설치 방안 마련이 요구되고 있다. 또한 공공 흡연시설 부족은 흡연자의 이용 불편뿐만 아니라 비지정 장소 흡연으로 이어질 수 있어 체계적인 관리가 필요하다 [1].

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

최근 공공데이터의 개방 확대와 데이터 분석 기술의 발전에 따라 도시 문제 해결을 위한 데이터 기반 정책 수립이 활발히 이루어지고 있다. 특히 유동인구, 금연구역, 공공시설 등의 데이터를 활용하면 지역별 특성을 정량적으로 분석할 수 있으며, 이를 통해 보다 합리적인 공공 흡연시설 설치 방안을 제시할 수 있다.

이에 본 연구는 서울시 공공데이터를 활용하여 공공 흡연시설 설치 우선지역을 도출하고자 한다. 이를 위해 방문자수, 금연구역수, 금연구역면적합 및 흡연시설수 데이터를 통합하고, 흡연시설부족도 지표를 생성하였다. 또한 K-Means, 계층적 군집분석 및 DBSCAN을 적용하여 군집화 성능을 비교하고, 공공 흡연시설 설치 정책에 활용 가능한 우선 검토 지역을 제안하고자 한다.

II. 관련 연구

2.1 MCLP 기반 입지 선정 연구

MCLP(Maximal Covering Location Problem)는 제한된 시설 수를 배치할 때 최대 수요를 커버할 수 있는 위치를 찾는 대표적인 시설 입지 최적화 기법이다. 선행연구에서는 금연구역과 담배꽁초 쓰레기통 주변 격자를 제외하고, 담배꽁초 무단투기 적발지점, 담배소매점 및 유동인구 데이터를 활용하여 격자별 수요를 산출하였다. 이후 흡연구역 간 최소 거리 제한을 고려한 MCLP 모델을 적용하여 제한된 수의 흡연구역이 최대 수요를 커버할 수 있는 최적 입지를 도출하였다[2].

2. 2 설치매력도 및 Random Forest 기반 예측

Random Forest는 다수의 의사결정트리를 결합하여 예측을 수행하는 앙상블 학습 기법이다. 선행연구에서는 다양한 공공시설의 설치 위치 및 활용 데이터를 기반으로 설치매력도를 산출하고, 이를 목표 변수로 설정하여 Random Forest 모델을 학습하였다. 또한 변수 중요도 분석과 GIS 기반 시각화를 통해 공공시설 설치 우선지역을 도출하였다[3].

III. 제안방법

3.1 데이터 수집

본 연구에서는 서울시 공공 흡연시설 설치 우선지역을 도출하기 위해 유동인구 데이터, 금연구역 데이터 및 공공 흡연시설 데이터를 수집하였다. 유동인구 데이터는 서울 열린데이터광장에서 제공하는 S-DoT 보행량 데이터를 활용하였으며, 계절적 변동과 특정 시기의 영향을 완화하기 위해 2024년 3월부터 2025년 12월까지의 자료를 수집하였다. 또한 월별 대표성을 확보하기 위해 각 월의 2주차 데이터를 기준으로 분석을 수행하였다.

금연구역 데이터는 서울시 및 25개 자치구에서 제공하는 금연구역 현황 데이터를 활용하였으며, 공공 흡연시설 데이터는 서울시에서 제공하는 실외 공공 흡연시설 현황 데이터를 활용하였다.

수집된 데이터는 행정동 단위로 통합하였으며, 방문자수, 금연구역수, 금연구역면적합 및 흡연시설수를 최종 분석 변수로 구성하였다.

3.2 데이터 전처리 및 변수 생성

수집된 데이터는 제공 기관별로 행정구역 표기 방식이 상이하여 자치구명을 기준으로 통일한 후 통합하였다. 금연구역 데이터는 서울시 25개 자치구 자료를 통합하여 자치구별 금연구역 수와 금연구역 면적을 집계하였으며, 공공 흡연시설 데이터는 자치구별 공공 흡연시설 수를 기준으로 정리하였다. 이후 자치구 단위의 금연구역 정보와 공공 흡연시

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

설 정보를 행정동 단위 유동인구 데이터와 결합하여 최종 분석 데이터셋을 구축하였다. 구축된 데이터셋은 총 556개의 관측치와 9개의 변수로 구성되며, 결측치는 존재하지 않는 것으로 확인되었다.

본 연구에서는 공공 흡연시설 설치 우선지역 도출을 위해 방문자수, 금연구역수, 금연구역면적합 및 흡연시설수를 주요 변수로 구성하였다. 또한 유동인구와 흡연시설 공급 수준을 함께 고려하기 위해 흡연시설부족도 변수를 생성하였다. 이를 통해 공공 흡연시설 공급이 상대적으로 부족한 지역을 식별하고자 하였다.

3.3 K-Means 군집분석

본 연구에서는 분할 기반 군집화 기법인 K-Means를 적용하여 행정동별 특성을 분석하였다. 분석에는 방문자수, 금연구역수, 금연구역면적합, 흡연시설수 및 흡연시설부족도를 활용하였다.

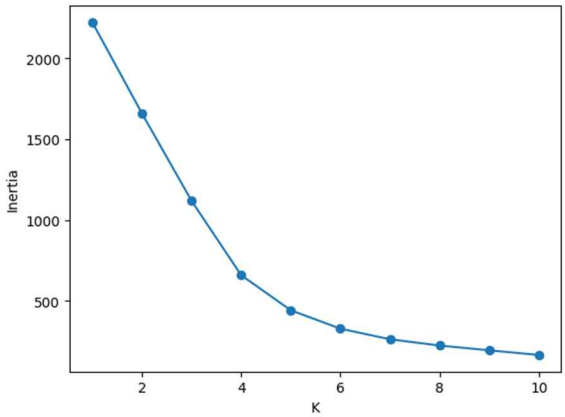


그림 1. Elbow Method 결과

군집 수 결정은 <그림 1>과 같이 Elbow Method를 활용하였다. 분석 결과 군집 수가 증가할수록 Inertia가 감소하였으며, 감소 폭이 완만해지는 지점을 고려하여 최종 군집 수를 4개로 설정하였다.

표1. K-Means 군집별 특성

cluster	방문자수	금연구역수	금연구역면적합	흡연시설부족도
0	38.966781	381.218750	2.233557e+06	3.144531
1	220.882237	552.188940	6.394881e+05	1.834101
2	69.925396	6722.283019	2.130894e+05	0.000000
3	53.527127	495.000000	9.082040e+05	40.000000

<표 1>의 분석 결과 Cluster 1은 방문자수와 흡연시설부족도가 높게 나타나 공공 흡연시설 공급이 부족한 지역군으로 확인되었다. 반면 Cluster 3은 흡연시설수가 가장 많고 흡연시설부족도가 가장 낮아 공공 흡연시설 공급이 상대적으로 충분한 지역군으로 나타났다. 또한 Cluster 2는 금연구역수가 매우 높은 특수 군집으로 확인되었으며, Cluster 0은 평균적인 특성을 보이는 일반 군집으로 분류되었다.

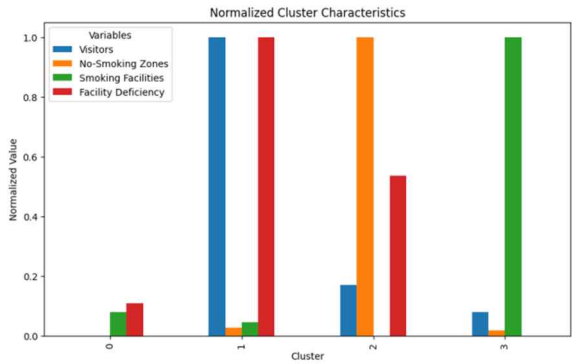


그림 2. K-Means 군집별 특성(정규화)

<그림 2>은 정규화된 값을 활용하여 군집별 특성을 비교한 결과이다. 군집 간 특성 차이가 명확하게 나타났으며, 특히 Cluster 1은 높은 방문자수와 흡연시설부족도를 보여 공공 흡연시설 설치 우선 검토 군집으로 확인되었다.

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

3.4 계층적 군집분석

본 연구에서는 계층 기반 군집화 기법인 계층적 군집분석(Hierarchical Clustering)을 적용하여 행정동별 특성을 분석하였다. 분석에는 방문자수, 금연구역수, 금연구역면적합, 흡연시설수 및 흡연시설부족도를 활용하였으며, K-Means와의 비교를 위해 군집 수는 4개로 설정하였다. 분석 결과 총 556개의 행정동 데이터가 4개의 군집으로 분류되었다. 군집별 특성을 살펴보면, 높은 방문자수와 흡연시설부족도를 나타내는 군집과 흡연시설 공급 수준이 상대적으로 높은 군집이 구분되었으며, 금연구역 수가 많은 특수 군집도 확인되었다.

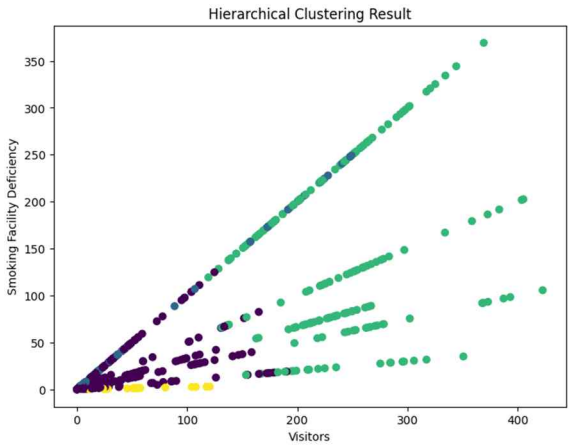


그림 3. 계층적 군집분석 결과 시각화

<그림 3>는 계층적 군집분석 결과를 시각화한 것이다. 분석 결과 군집별 특성 차이가 비교적 명확하게 나타났으며, 유동인구 규모와 흡연시설 공급 수준에 따라 행정동이 구분되는 경향을 확인할 수 있었다.

3.5 DBSCAN 분석

본 연구에서는 밀도 기반 군집화 기법인 DBSCAN(Density-Based Spatial Clustering of Applications with Noise)을 적용하여 행정동별 특성을 분석하였다. 분석에는 방문자수, 금연구역수, 금연구역면적합, 흡연시설수 및 흡연시설부족도를 활용하였다.

분석 결과 총 7개의 군집과 1개의 이상치 군집이 도출되었다. 일부 군집은 높은 방문자수와 흡연시설 부족도를 나타냈으며, 특정 군집은 금연구역 수가 높게 나타나는 등 군집별 특성 차이가 확인되었다.

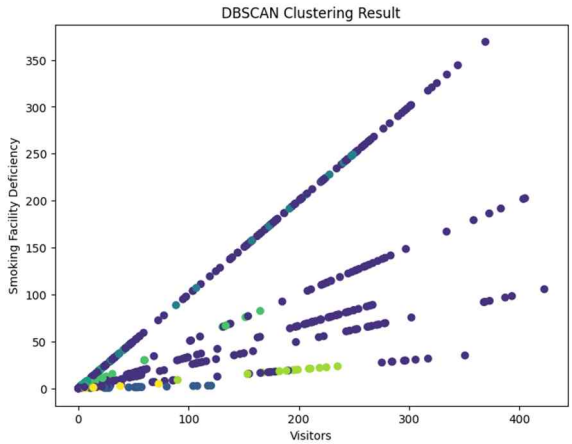


그림 4. DBSCAN 분석 결과 시각화

<그림 4>는 DBSCAN 분석 결과를 시각화한 것이다. 대부분의 행정동은 밀도 기반 군집으로 분류되었으며, 일부 데이터는 이상치로 식별되었다.

3.6 군집화 기법 비교 및 공공 흡연시설 설치 우선지역 도출

본 연구에서는 공공 흡연시설 설치 우선지역 도출을 위해 K-Means, 계층적 군집분석 및 DBSCAN을 적용하고 결과를 비교하였다.

표2. 군집화 기법 성능 비교

군집화 기법	Silhouette Score ↑	CH Index ↑	DB Index ↓
K-Means	0.458	347.551	0.757
Hierarchical Clustering	0.462	348.337	0.750
DBSCAN	0.082	105.116	0.929

<표 2>의 분석 결과 계층적 군집분석은

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

Silhouette Score, CH Index 및 DB Index에서 가장 우수한 성능을 나타냈으며, K-Means는 이에 근접한 성능을 보였다. 반면 DBSCAN은 상대적으로 낮은 군집 품질을 나타냈다.

비록 계층적 군집분석이 가장 우수한 성능을 보였으나, K-Means는 군집별 특성을 직관적으로 해석할 수 있으며 정책적 활용성이 높다는 장점을 가진다. 따라서 본 연구에서는 군집 품질과 결과 해석의 용이성을 고려하여 K-Means 분석 결과를 최종 결과로 활용하였다.

표3. 공공 흡연시설 설치 우선 후보지역(상위 5개)

순위	행정동	자치구	평균 흡연시설부족도
1	목3동	양천구	308.27
2	목3동(삼)	양천구	304.49
3	신원동	관악구	266.44
4	신월1동	양천구	247.99
5	잠실본동	송파구	244.44

K-Means 분석 결과를 기반으로 행정동별 평균 흡연시설부족도를 산출한 결과, 양천구 목3동, 양천구 목3동(삼), 관악구 신원동, 양천구 신월1동 및 송파구 잠실본동 등이 공공 흡연시설 설치 우선 검토 지역으로 도출되었다. 해당 지역은 유동인구에 비해 공공 흡연시설 공급 수준이 낮은 것으로 나타나 향후 공공 흡연시설 설치 정책 수립 시 우선적으로 고려될 필요가 있다.

IV. 결론

본 연구에서는 서울시 공공데이터를 활용하여 공공 흡연시설 설치 우선지역을 도출하고자 하였다. 이를 위해 유동인구 데이터, 금연구역 데이터 및 공공 흡연시설 데이터를 수집·통합하였으며, 방문자수, 금연구역수, 금연구역면적합, 흡연시설수 및 흡연시설부족도를 활용하여 군집분석을 수행하였다.

K-Means, 계층적 군집분석 및 DBSCAN을 적용하여 비교한 결과, 계층적 군집분석이 가장 우수한 군집 성능을 보였으나 K-Means 역시 유사한 수준의 성능을 나타냈다. 또한 K-Means는 군집별 특성을 직관적으로 해석할 수 있어 정책적 활용성이 높은 것으로 판단되었다.

분석 결과 양천구 목3동, 양천구 목3동(삼), 관악구 신원동, 양천구 신월1동 및 송파구 잠실본동 등이 공공 흡연시설 설치 우선 검토 지역으로 도출되었다. 해당 지역은 유동인구에 비해 공공 흡연시설 공급 수준이 낮은 것으로 나타나 향후 공공 흡연시설 설치 정책 수립 시 우선적으로 고려될 필요가 있다. 목3동과 목3동(삼)은 데이터 수집 시점에 따른 행정동 표기 차이를 반영한 것이다.

본 연구는 공공데이터를 활용하여 공공 흡연시설 설치 필요 지역을 데이터 기반으로 도출하였다는 점에서 의의를 가진다. 그러나 민간 흡연시설, 실제 흡연자 수 및 흡연 민원 데이터가 반영되지 못한 한계가 존재한다. 향후 연구에서는 다양한 공간정보와 흡연 관련 데이터를 추가적으로 활용하여 보다 정교한 입지선정 모델을 구축할 필요가 있다.

참고문헌

[1] 이근원, 김주연, 장영호, "흡연자의 행태분석을 통한 흡연위치 및 흡연공간의 유형에 관한 연구," 한국공간디자인학회 논문집, Vol. 15, No. 8, pp. 417-428, Dec. 2020.

[2] 김건형, 박다인, 정유찬, "MCLP를 활용한 흡연구역 최적 입지 선정: 강남구를 대상으로," 대한지리학회 학술대회논문집, Vol. 2025, No. 6, pp. 315-316, June 2025.

[3] 김명서, 김예나, 박선유, 신기환, 주종화, "공공시설 우선 설치 예측을 위한 통합 머신러닝 기반 분석 시스템," Proceedings of KIIT Conference, Vol. 2025, No. 6, pp. 1178-1182, June 2025.

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

기상 데이터 기반 돼지 구제역 발생 위험 분류 모델

조원희[O], 백수진[*]
[O]한신대학교 AISW학
[*]한신대학교 AI시스템반도체학
e-mail:{ovo8709, croso79}@hs.ac.kr[*]

A Weather Data-Driven Classification Model for Foot-and-Mouth Disease Outbreak Risk in Pigs

Wonhee Cho[O], Sujin Back[*]
[O]Department of AISW, HanShin University
*Department of AI System Semiconductor Engineering, Hanshin University

요 약

구제역은 축산업에 막대한 경제적 피해를 초래하는 제1종 가축전염병으로, 선제적 방역을 위한 위험 예측 기술의 중요성이 증가하고 있다. 이에 본 연구에서는 농림축산검역본부의 구제역 발생 이력 데이터와 기상청 기상 관측 데이터를 결합하여 기상 조건 기반 구제역 발생 이진 분류 모델을 개발하였다. 최저기온, 최대풍속, 최대풍향, 평균습도, 강수량 및 일조시간을 입력 변수로 활용하고, Logistic Regression, Random Forest, XGBoost를 비교 분석하였다. 실험 결과 Random Forest가 F1-Score 0.7692, ROC-AUC 0.9054로 가장 우수한 성능을 나타냈다. 또한 Sigmoid 기반 확률 보정을 적용한 결과, 위험 등급별 실제 발생 비율은 낮음 7.9%, 중간 28.6%, 높음 86.8%로 나타나 등급 체계의 타당성을 확인하였다. 본 연구는 공공데이터 기반의 구제역 발생 위험 예측 및 확률 기반 위험 등급을 제시함으로써 선제적 방역 의사결정 지원 가능성을 제시하였다.

키워드: 구제역, 기상데이터, 머신러닝, 위험 등급 분류, 확률 보정

I. 서론

구제역은 소, 돼지 등 우제류 동물에게 발생하는 급성 바이러스성 전염병으로, 국내 제 1종 가축전

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

염병이다. 전파 속도가 매우 빠르고 경제적 피해가 막대하여 축산업 전반에 치명적인 위협이 된다. 2010~2011년 국내 구제역 대유행 당시 약 350만 마리의 가축이 살처분되었으며, 직간접적 경제 피해는 약 2조원에 달하였다.

현행 방역 체계는 발생 신고 이후의 사후 대응에 집중되어 있어 신고까지 평균 2~3일이 소요되는 동안 바이러스가 인근 농장으로 이미 확산된 상태일 가능성이 높다. 이에 따라 발생 이전에 위험을 정량적으로 분류하고 선제적으로 대응할 수 있는 머신러닝 기반 이진 분류 모델이 필요하다.

분류 모델의 predict_proba() 출력은 클래스 판별을 위한 점수에 해당하나, 확률 보정을 통해 실제 발생 빈도와 일치하는 신뢰 확률로 변환할 수 있다. 따라서 본 연구는 보정된 확률값을 위험 등급 분류에 활용하는 방역 의사결정 체계를 제안한다.

II. 관련 연구

구제역 바이러스와 환경 요인의 관계에 대한 연구는 활발히 진행되어 왔다. Munsey et al.(2019)은 우간다의 구제역 발생 데이터를 분석하여 강수량과 기온이 주요 발생 요인임을 규명하였다[1]. Frontiers in AI(2024)에 게재된 연구는 2011~2022년 우간다 구제역 데이터와 기상 데이터를 결합하여 7개 머신러닝 모델의 예측 성능을 비교하였다[2]. 바이러스 생존 특성 측면에서 구제역은 저온, 고습한 환경에서 생존율이 높고 자외선에 취약한 것으로 알려져 있다. 또한 공기 중 전파가 가능하여 풍속 및 풍향이 전파 경로에 영향을 미치며, 강수는 바이러스를 토양에 침투, 이동시키는 역할을 한다.

그러나 국내 데이터를 활용한 기상 기반 구제역 발생 확률 예측 연구는 매우 제한적이며, 발생/미발생 레이블이 포함된 공개 데이터셋 또한 적다. 본 연구는 이러한 공백을 메우고자 한다.

III. 연구 방법

3.1 데이터 수집

본 연구에서는 국가가축방역통합시스템(KAHIS)의 구제역 발생 데이터와 기상청 기상자료개방포털의 기상 데이터를 활용하였다. 구제역 발생 데이터는 1990~2025년 국내 돼지 축종 발생 이력 285건을 사용하였으며, 기상 변수는 최저기온, 최대풍속, 최대풍향, 평균습도, 강수량, 일조시간의 6개 변수로 구성하였다.

3.2 미발생 데이터 구축 및 전처리

구제역 발생 여부를 이진 분류하기 위해 발생일 기준 발생 기록이 없는 관측소 데이터를 미발생(0)으로 레이블링하였다. 다른 가축 질병 발생일 데이터는 제외하였으며, 클래스 불균형 완화를 위해 발생:미발생 비율을 1:3으로 구성하였다. 또한 행정구역과 인접 기상관측소를 매핑하여 데이터를 병합하였으며, 결측치는 보완 및 대체 과정을 통해 처리하였다. 이상치는 기상 현상의 자연스러운 극값으로 판단하여 제거하지 않았다.

3.3 분류 모델 설계

모델의 분류 과제는 6개의 기상 변수를 입력받아 구제역 발생 여부를 이진 분류하는 것이다. 비교 모델로 Logistic Regression, Random Forest, XGBoost를 사용하였으며, 주요 하이퍼파라미터는

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

예비 실험을 통해 설정하였다. 또한 데이터 불균형을 고려하여 Logistic Regression과 Random Forest에 class_weight='balanced'를 적용하였다.

IV. 연구 결과

4.1 이진 분류 모델 성능 비교

세 이진 분류 모델의 테스트 성능 평가 결과는 <그림 1>, <표 1>과 같다.

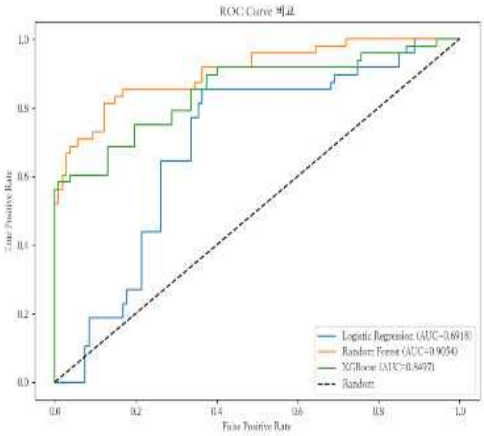


그림 1. 이진 분류 모델 테스트 성능 평가

표 1. 이진 분류 모델 테스트 성능 평가

Model	Logistic Regression	Random Forest	XGBoost
Accuracy	0.6645	0.8452	0.7419
Precision	0.4767	0.7143	0.5625
Recall	0.8542	0.8333	0.7500
F1-score	0.6119	0.7692	0.6429
ROC-AUC	0.6918	0.9054	0.8497

Random Forest가 F1-Score 0.7692, ROC-AUC 0.9054로 전체 분류 지표에서 가장 우수한 성능을 보여 최종 이진 분류 모델로 선정하였다. Logistic Regression은 Recall이 0.8542로 높아 발생케이스를 잘 탐지하지만, Precision이 0.4767로 낮아 오분류하는 비율이 높았다. 이는 선형 결정 경계의 한계로, 기상 변수 간 비선형 상호작용을 포착하지 못하는 것으로 보인다.

4.2 주요 예측 요인 분석

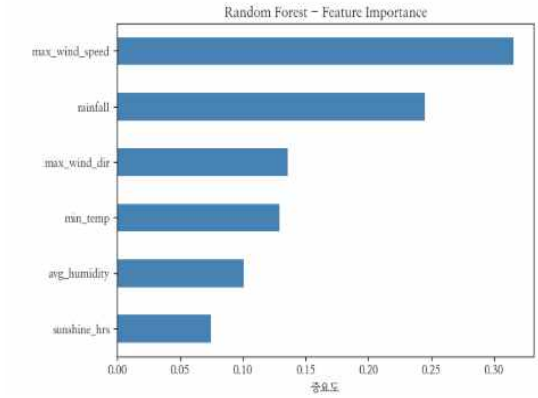


그림 2. 주요 예측 요인 분석

Random Forest 분류 모델의 주요 예측 요인 분석 결과, <그림 2>와 같이 최대풍속과 강수량이 가장 높은 중요도를 나타냈다. 이는 구제역 바이러스의 환경적 특성과 관련된 기존 연구 결과와 일치하며, 모델의 예측 타당성을 뒷받침한다. 특히 강수량은 구제역 발생에 영향을 미치는 주요 기상 변수로 확인되었다.

4.3 확률 보정 적용

Niculescu-Mizil과 Caruana(2005)는 Random Forest가 확률을 과도하게 편향시키는 경향이 있음을 실증하였다[3]. 본 연구에서도 보정 전 확률 상위 5개가 모두 0.8942로 동일하게 출력되는 현상이 관찰되었다. 이는 특정 기상 패턴에 대해 트리 다수결 투표가 동일한 값으로 수렴하며, 실제 발생 위험의 미세한 차이를 반영하지 못했음을 의

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

미한다. 방역 의사결정 도구로서 신뢰 가능한 확률값을 제공하기 위해 확률 보정이 필요하다.

sklearn.calibration 모듈의 CalibratedClassifierCV를 사용하였으며, 학습 데이터가 상대적으로 적은 점을 고려하여 Sigmoid 방식을 선택하였다. 보정 전후 확률 비교 결과, <그림 3>에서와 같이 보정 전 상위 확률이 모두 0.8942로 수렴했던 것이 보정 후 0.8648로 재조정되었다. 보정 후 분류 성능은 Sigmoid 기반 확률 보정을 적용한 결과, Accuracy는 0.8323, F1-Score는 0.7292, ROC-AUC는 0.9054로 나타났다.

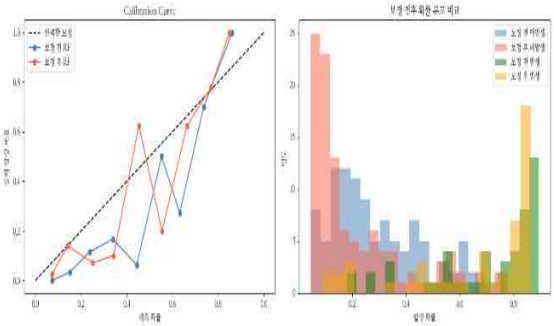


그림 3. 확률 보정 결과

V. 결론

본 연구는 국내 공공데이터를 활용하여 기상 조건 기반 구제역 발생 이진 분류 모델을 개발하고 확률 보정을 적용하였다. 국가가축방역통합시스템과 기상청 데이터를 직접 수집, 결합하고 미발생 데이터를 체계적으로 구축하여 이진 분류 데이터셋을 완성하였다. 세 이진 분류 모델을 비교한 결과 Random Forest가 F1-Score 0.7692, ROC-AUC 0.9054로 최적 모델로 선정되었다. 확률 보정 적용 후 Accuracy 0.8323, F1-Score 0.7292로 분류 성능이 나타났으며, 위험 등급별 실제 발생 비율(낮음 7.9%, 중간 28.6%, 높음 86.8%)을 통해 실증적 타당성을 확인하였다.

본 연구의 모델을 활용하면 선제적 방역 체계 수립이 가능하다. 기상청의 관측 및 예보 데이터를 일별로 입력하면 전국 지역별 구제역 발생 위험을 실시간으로 이진 분류하고 위험 등급을 산출할 수 있다. '높음'등급 지역에 사전 경보를 발령하고 농장 점검을 강화함으로써 초기 확산을 차단할 수 있다. 또한 이진 분류 모델에 확률 보정을 결합하여 단순한 발생/미발생 판정을 넘어 신뢰 가능한 확률값을 제공한다. '낮음'등급의 경우 실제 발생 비율이 7.9%에 불과하므로 방역 자원의 불필요한 투입을 줄이고, '높음'등급에 집중 배분함으로써 자원 효율성을 극대화할 수 있다.

본 연구는 다음과 같은 한계점을 가진다. 첫째, 기상 데이터만을 입력 변수로 사용하여 가축 이동, 농장 밀도, 백신 접종률 등 주요 방역 요인이 반영되지 않았다. 향후 이러한 변수를 추가할 경우 분류 정확도가 크게 향상될 것으로 기대된다. 둘째, 기상 관측소 단위 분석으로 읍면동 수준의 세밀한 위험 분류가 불가능하다.

참고문헌

[1] Anna Munsey, "Spatial distribution and risk factors for foot and mouth disease virus in Uganda: Opportunities for strategic surveillance", Prev Vet Med, Nov 1:171:104766, 2019

[2] Geofrey Kapalaga, "A unified Foot and Mouth Disease dataset for Uganda: evaluating machine learning predictive performance degradation under varying distributions", Frontiers in Artificial Intelligence, Vol 7, Article 1446368, 2024

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

[3] Alexandru Niculescu-Mizil, “Predicting Good Probabilities With Supervised Learning, Department Of Computer Science, Ithaca NY 14853, 2005

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

공개 EDA 데이터 기반 머신러닝과 U-Net을 활용한 반도체 라우팅 혼잡 Hotspot 예측

김영범[O], 백수진[*]
[O]한신대학교 AI시스템반도체학과
[*]한신대학교 AI시스템반도체학과
e-mail: myload4545@hs.ac.kr[O], croso79@hs.ac.kr[*]

Routing Congestion Hotspot Prediction in Semiconductor Physical Design Using Machine Learning and U-Net with Open EDA Data

Youngbeom Kim[O], Sujin Baek[*]
[O]Department of AI System Semiconductor Engineering, Hanshin University
[*]Department of AI System Semiconductor Engineering, Hanshin University

요 약

반도체 물리설계 과정에서 라우팅 혼잡은 배선 지연, timing closure 실패 및 설계 규칙 위반을 유발하는 주요 문제이다. 본 연구에서는 공개 EDA 데이터인 CircuitNet-N28 Routability Feature를 활용하여 라우팅 혼잡 Hotspot 예측을 수행하였다. 7개 물리설계 feature map으로부터 통계 기반 feature를 추출하여 Logistic Regression, Random Forest, XGBoost 모델을 비교하였으며, 원본 2D feature map 기반 U-Net segmentation 모델도 함께 적용하였다. 실험 결과 Random Forest는 Accuracy 0.9561, F1-score 0.8495, ROC-AUC 0.9858을 기록하였고, U-Net은 Accuracy 0.9169, F1-score 0.8450, IoU 0.7316을 달성하였다. 본 연구는 공개 EDA 데이터를 활용한 재현 가능한 라우팅 혼잡 예측 절차를 제시하였으며, 초기 물리설계 단계에서 혼잡 위험 영역 탐지 및 설계 수정 지원에 활용될 것으로 기대된다.

키워드: 공개 EDA 데이터, CircuitNet, 라우팅 혼잡, 머신러닝, U-Net, Hotspot 예측

I. 서론

최근 반도체 공정 미세화와 설계 복잡도 증가로 인해 물리설계 단계에서 routing congestion 문제가 중요해지고 있다. Routing congestion은 배선 지연, 설계 규칙 위반 및 timing closure 실패를

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역 혁신 중심 대학 지원 사업(경기RISE사업)의 연구 결과 입니다

유발하며, 설계 반복 증가의 원인이 된다. 따라서 라우팅 이후 문제가 발생하기 전에 초기 설계 단계에서 혼잡 가능 영역을 예측하는 것이 중요하다.

최근 EDA 분야에서는 머신러닝 기반 downstream prediction 연구가 활발히 진행되고 있다[1]. 그러나 산업용 설계 데이터는 공개가 제한되어 재현성과 비교 가능성이 낮다는 한계가 있다. 이를 해결하기 위해 공개 EDA 데이터셋인 CircuitNet이 제안되었으며, congestion prediction 등 다양한 prediction task를 지원한다[2]. 본 연구에서는 CircuitNet-N28 Routability Feature를 활용하여 라우팅 혼잡 Hotspot 예측 문제를 구성하였다.

본 연구는 공개 EDA 데이터를 기반으로 표 기반 머신러닝과 공간 정보 기반 U-Net segmentation 모델을 비교·분석하였다. 먼저 feature map의 통계값을 이용해 Logistic Regression, Random Forest, XGBoost를 비교하였고, 이후 원본 2D feature map을 활용한 U-Net 기반 Hotspot segmentation을 수행하였다. 본 연구의 주요 기여는 동일한 공개 EDA 데이터셋에서 설계 단위의 congestion 위험도 분류와 위치 단위의 Hotspot segmentation을 분리하여 구성하고, 두 접근의 역할과 활용 가능성을 정량적으로 비교한 데 있다.

II. 관련 연구

CircuitNet은 공개 EDA 데이터셋으로, 물리설계 feature와 label을 제공하여 ML 기반 EDA 연구의 재현성을 지원한다[2]. 데이터는 grid 기반 2D feature map 형태로 제공되므로, 표 기반 모델뿐 아니라 CNN 기반 모델 적용에도 적합하다.

Routing congestion prediction은 배선 혼잡 영역을 조기에 예측하기 위한 문제로, CNN 기반 연구들은 layout의 공간 정보를 활용할 경우 예측 성능이 향상될 수 있음을 보였다[3]. 표 기반 예측 모델로는 Random Forest[5]와 XGBoost[6]가 널리 활용되며, U-Net은 encoder-decoder 구조와 skip connection을 통해 위치 정보를 유지한 segmentation이 가능하므로 layout 내 congestion Hotspot 탐지에 적합하다[4]. 또한 U-Net 변형 모델을 routability prediction에 적용한 연구도 보고되었다[7]. 본 연구는 공개 데이터 기반으로 표 기반 머신러닝과 U-Net segmentation을 함께 비교·분석하는 데 목적이 있다.

III. 제안방법

3.1 데이터 구성 및 검증

본 연구는 CircuitNet-N28 Routability Feature 데이터를 사용하였다. 입력 feature 폴더와 label 폴더에서 동일한 sample ID가 존재하는지 확인한 결과, 공통 sample 수는 10,242개였다. 분석에 사용한 입력은 cell_density, macro_region, RUDY, RUDY_long, RUDY_short, RUDY_pin, RUDY_pin_long의 7개 2D map이며, label 생성을 위해 congestion_GR_horizontal_overflow와 congestion_GR_vertical_overflow를 사용하였다.

표 1. 사용한 입력 feature와 label 구성

구분	Feature	설명
입력	cell_density	각 grid/tile의 standard cell 밀도 분포
입력	macro_region	macro 또는 blockage 영역 정보를 나타내는 이진 map

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역 혁신 중심 대학 지원 사업(경기RISE사업)의 연구 결과 입니다

입력	RUDY	routing utilization distribution 기반 전체 배선 수요 추정 map
입력	RUDY_long / RUDY_short	net 길이 특성에 따른 routing demand feature
입력	RUDY_pin / RUDY_pin_long	pin 기반 routing demand 및 long-range pin demand feature
Label	horizontal / vertical overflow	global routing overflow를 나타내는 Hotspot label 생성용 map

원본 데이터의 각 sample은 서로 다른 레이아웃 크기를 가질 수 있다. 전처리된 sample-level dataset의 height와 width는 최소 222에서 최대 313×315 수준까지 분포하였으며, U-Net 실험에서는 최대 크기를 16의 배수로 맞추기 위해 320×320 크기로 zero padding을 적용하였다. 단, padding 영역은 valid mask를 사용하여 loss와 metric 계산에서 제외하였다.

3.2 Sample-level classification 문제 정의

머신러닝 baseline에서는 원본 2D map을 직접 입력하지 않고, 각 sample의 feature map을 통계값으로 요약하였다. 각 입력 feature에 대해 mean, standard deviation, maximum, 95 percentile(p95), non-zero ratio를 계산하였으며, 7개 feature에서 총 35개의 입력 변수를 구성하였다.

정답 label은 congestion severity를 기준으로 생성하였다. 먼저 horizontal overflow와 vertical overflow를 합산하여 congestion score를 계산하고, 각 sample에서 상위 5% tile의 평균값을 congestion severity로 정의하였다. 이후 전체 sample에서 congestion severity 상위 15%를 high congestion으로, 나머지를 low congestion으로 라벨링하였다. 이때 high congestion 경계값은 0.4338였으며, 최종 class 분포는 low congestion 8,705개, high congestion 1,537개로 나타났다.

모델 학습 시 label 생성에 직접 사용된 congestion_mean, congestion_max, congestion_p95, congestion_nonzero_ratio, congestion_severity 및 설계 크기 변수(height, width, tile_count)는 입력에서 제외하였다. 이는 정답 유도 변수에 의한 label leakage를 방지하고, 실제 routability feature만으로 예측하도록 하기 위함이다.

3.3 Pixel-level segmentation 문제 정의

U-Net 실험에서는 각 sample을 이미지형 데이터로 보고, 7개의 feature map을 channel 방향으로 쌓아 7채널 2D 입력을 구성하였다. 정답 mask는 horizontal overflow와 vertical overflow를 합산한 뒤, 합산 값이 0보다 큰 위치를 Hotspot으로 정의하였다. 즉, 모델은 각 위치가 routing congestion Hotspot인지 아닌지를 예측하는 binary segmentation 문제를 수행한다.

전체 sample은 train/validation/test set으로 약 70/15/15 비율로 분리하였다. 결과적으로 U-Net 실험에서는 train 7,169개, validation 1,536개, test 1,537개 sample을 사용하였다. Validation set은 threshold tuning과 best model 선택에 사용하였고, test set은 최종 성능 평가에만 사용하였다.

3.4 분석 절차 및 모델 구성

3.4.1 전체 분석 절차

본 연구의 분석 절차는 데이터 확인, sample-level dataset 생성, 머신러닝 baseline 학습, U-

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역 혁신 중심 대학 지원 사업(경기RISE사업)의 연구 결과 입니다

Net segmentation 학습, threshold tuning 및 결과 해석의 순서로 구성하였다. 먼저 raw feature와 label 폴더에서 공통 sample을 확인하고, 각 feature map의 shape와 값 범위를 점검하였다. 이후 두 가지 목적의 dataset을 별도로 구성하였다.

첫 번째 dataset은 설계 sample 전체의 혼잡 여부를 판단하기 위한 tabular dataset이다. 이는 빠른 위험도 screening과 feature importance 해석에 적합하다. 두 번째 dataset은 2D feature map을 그대로 입력하는 segmentation dataset이다. 이는 레이아웃 내 Hotspot 위치를 직접 예측하고 시각화하기 위한 목적이다.

3.4.2 머신러닝 baseline 구성

Sample-level classification에는 Logistic Regression, Random Forest, XGBoost를 사용하였다. 결측치는 median imputation으로 처리하였고, Logistic Regression에는 StandardScaler를 적용하였다. Class imbalance를 고려하여 Logistic Regression과 Random Forest에는 class_weight를 적용하였고, XGBoost에는 train set의 음성/양성 비율을 반영한 scale_pos_weight를 적용하였다.

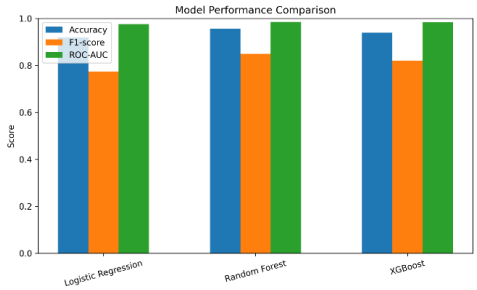
Random Forest는 n_estimators=300으로 설정하고, XGBoost는 n_estimators=300, max_depth=4, learning_rate=0.05, subsample=0.8, colsample_bytree=0.8로 설정하였다. 모든 모델은 동일한 train/test split에서 평가하였으며, Accuracy, Precision, Recall, F1-score, ROC-AUC를 사용하였다. 또한 Random Forest와 XGBoost의 feature importance를 분석하여 혼잡 예측에 중요한 물리설계 feature를 확인하였다.

3.4.3 U-Net segmentation 구성

U-Net은 encoder와 decoder가 대칭적으로 구성된 segmentation 모델이다. Encoder는 convolution과 max pooling을 통해 국소 패턴과 문맥 정보를 추출하고, decoder는 transposed convolution과 skip connection을 통해 위치 정보를 복원한다. 본 연구에서는 입력 채널을 7개, 출력 채널을 1개로 설정하여 Hotspot probability map을 출력하도록 구성하였다.

각 feature map은 sample 단위로 log 변환 및 표준화를 적용하였다. macro_region은 이진 mask 성격이 강하므로 별도 표준화 없이 사용하였다. 크기가 다른 sample은 320×320으로 padding하되, valid mask를 함께 사용하여 padding 영역이 loss와 평가 지표에 포함되지 않도록 하였다.

손실 함수는 weighted binary cross entropy(BCE) loss와 Dice loss를 결합하여 사용하였다. Hotspot pixel은 전체 layout 영역 중 일부에만 나타나므로 class imbalance 문제가 존재한다. 이에 positive pixel에 가중치를 부여하여 모델이 Hotspot 영역을 과소 예측하지 않도록 하였다. 모델 출력 probability map은 threshold를 적용하여 binary mask로 변환하였다.



※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역 혁신 중심 대학 지원 사업(경기RISE사업)의 연구 결과 입니다

그림 1. Sample-level 머신러닝 모델 성능 비교

3.5 실험 결과 및 분석

표 2. Sample-level 머신러닝 baseline 성능 비교

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Log.Reg.	0.9185	0.6620	0.9316	0.7740	0.9761
RF	0.9561	0.8729	0.8274	0.8495	0.9858
XGBoost	0.9390	0.7333	0.9316	0.8207	0.9850

<표 2>와 같이 Random Forest는 Accuracy 0.9561, Precision 0.8729, F1-score 0.8495, ROC-AUC 0.9858로 가장 균형 잡힌 성능을 보였다. Logistic Regression은 Recall이 0.9316으로 높았지만 Precision이 0.6620으로 낮아 high congestion을 넓게 예측하는 경향을 보였다. XGBoost 역시 Recall은 높았으나 Random Forest보다 Precision과 F1-score가 낮았다.

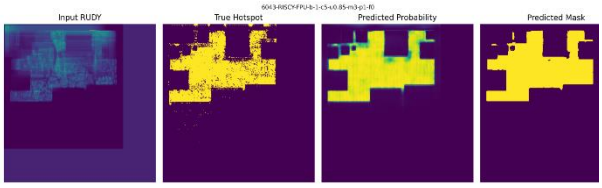


그림 2. U-Net 기반 routing congestion Hotspot 예측 예시 1(threshold=0.6)

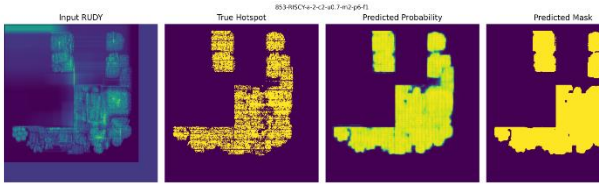


그림 3. U-Net 기반 routing congestion Hotspot 예측 예시 2(threshold=0.6)

U-Net은 validation set에서 선택한 threshold 0.6 기준으로 test set에서 Accuracy 0.9169, Precision 0.7721, Recall 0.9330, F1-score 0.8450, IoU 0.7316을 달성하였다. <그림 2>와 <그림 3>은 test sample에 대한 U-Net 예측 결과의 예시이다. Input RUDY map에서 routing demand가 높은 영역은 True Hotspot과 상당 부분 대응되며, Predicted Probability에서도 해당 영역의 확률이 높게 나타났다. 최종 Predicted Mask는 True Hotspot의 큰 구조를 잘 포착하였으나, 세부적인 점 형태의 noise까지 모두 따라가기보다는 주요 congestion region을 부드럽게 예측하는 경향을 보였다.

머신러닝 baseline과 U-Net은 경쟁 관계가 아니라 서로 다른 의사결정을 지원하는 상호보완적 접근이다. Random Forest는 빠른 sample-level screening과 feature importance 해석에 적합하고, U-Net은 layout 내 Hotspot 후보 영역을 시각적으로 제시하여 수정 후보를 탐지하는 데 활용될 수 있다. 따라서 본 연구는 설계 단위 위험도 선별과 위치 단위 localization을 동시에 지원하는 dual-level 분석 흐름을 제시한다.

IV. 결론

본 연구는 공개 EDA 물리설계 데이터인 CircuitNet-N28 Routability Feature를 활용하여 반도체 라우팅 혼잡 Hotspot 예측 문제를 정의하고, 머신러닝 기반 sample-level classification과 U-Net 기반 pixel-level segmentation을 수행하였다. 머신러닝 baseline에서는 Random Forest가 Accuracy 0.9561, F1-score 0.8495, ROC-AUC 0.9858로 가장 우수한 성능을 보였으며, feature importance 분석을 통해 RUDY 계열 feature와 cell density feature가 congestion 예측에 중요함을 확인하였다.

U-Net segmentation 실험에서는 7채널 2D feature map을 입력으로 사용하고, horizontal/vertical overflow를 합산하여 Hotspot mask를 생성하였다. Validation set threshold

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역 혁신 중심 대학 지원 사업(경기RISE사업)의 연구 결과 입니다

search 결과 threshold 0.6이 가장 적합하였고, 최종 test set에서 Accuracy 0.9169, Precision 0.7721, Recall 0.9330, F1-score 0.8450, IoU 0.7316을 달성하였다. 특히 높은 recall은 모델이 실제 Hotspot 영역을 놓치지 않고 탐지할 수 있음을 보여준다.

본 연구의 기대효과는 다음과 같다. 첫째, routing 이전 또는 초기 물리설계 단계에서 설계 sample의 congestion 위험도를 빠르게 판단하는 보조 지표로 활용될 수 있다. 둘째, U-Net의 예측 mask는 설계자가 레이아웃 내 문제 영역을 직관적으로 확인하고 수정 후보 영역을 선정하는 데 활용될 수 있다. 셋째, Random Forest는 screening에, U-Net은 Hotspot localization에 적합하여 두 접근은 상호보완적으로 활용될 수 있다.

향후 연구에서는 세 가지 방향의 개선을 수행할 예정이다. 첫째, false positive를 줄이기 위해 focal loss, Tversky loss, boundary-aware loss 등 다양한 segmentation loss를 비교할 것이다. 둘째, design-level split 또는 cross-design validation을 통해 서로 다른 설계 계열과 공정 노드에서 일반화 성능을 더 엄밀히 검증할 것이다. 셋째, 단순 threshold 기반 mask 생성 외에 morphology 기반 후처리, uncertainty estimation, attention U-Net 또는 transformer 기반 구조를 적용하여 예측 mask의 정밀도를 높일 계획이다.

다만 본 연구는 CircuitNet-N28 데이터와 sample 단위 분할에 기반한 실험이므로, 다른 공정 노드나 실제 산업 설계 데이터에서도 동일한 성능이 유지되는지는 추가 검증이 필요하다. 또한 Hotspot label을 overflow가 0보다 큰 위치로 정의했기 때문에 미세한 점 형태의 overflow도 positive로 포함된다. 이로 인해 U-Net은 세부적인 점 단위 경계보다 주요 혼잡 영역을 넓게 포착하는 경향을 보였으며, 향후에는 boundary-aware loss나 후처리 기법을 적용하여 mask의 정밀도를 개선할 필요가 있다.

또한 본 연구는 결과의 재현성을 높이기 위해 label 생성에 직접 사용된 congestion 관련 통계량을 머신러닝 입력 변수에서 제외하고, validation set에서 선택한 threshold를 test set 평가에 그대로 적용하였다. 이는 모델 선택과 최종 평가를 분리하여 성능이 과대평가되는 문제를 줄이기 위한 절차이다.

참고문헌

[1] G. Huang, J. Hu, Y. He, J. Liu, M. Ma, Z. Shen, J. Wu, Y. Xu, H. Zhang, K. Zhong, X. Ning, Y. Ma, H. Yang, B. Yu, H. Yang, and Y. Wang, "Machine Learning for Electronic Design Automation: A Survey," ACM Transactions on Design Automation of Electronic Systems, Vol. 26, No. 5, Article 40, 2021.

[2] Z. Chai, Y. Zhao, Y. Lin, W. Liu, R. Wang, and R. Huang, "CircuitNet: An Open-Source Dataset for Machine Learning Applications in Electronic Design Automation (EDA)," arXiv:2208.01040, 2022.

[3] Z. Xie, Y.-H. Huang, G.-Q. Fang, H. Ren, S.-Y. Fang, Y. Chen, and J. Hu, "RouteNet: Routability Prediction for Mixed-Size Designs Using Convolutional Neural Network," IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2018.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234-241, 2015.

[5] L. Breiman, "Random Forests," Machine Learning, Vol. 45, pp. 5-32, 2001.

[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역 혁신 중심 대학 지원 사업(경기RISE사업)의 연구 결과 입니다

Knowledge Discovery and Data Mining, pp. 785–794, 2016.
[7] H. Li, Y. Huo, Y. Wang, X. Yang, M. Hao, and X. Wang, "A Lightweight Inception Boosted U-Net Neural Network for Routability Prediction," arXiv:2402.10937, 2024.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역 혁신 중심 대학 지원 사업(경기RISE사업)의 연구 결과 입니다

청년 체크카드 소비 패턴과 금융상품 선호도 및 고객 이탈 분석

저자*, 진성민
한신대학교 빅데이터융합학과
e-mail:no2aqua@hs.ac.kr

An Analysis of Debit Card Spending Patterns, Financial Product Preferences, and Customer Churn Among Young Adults

Author*, Sungmin Jin
Dept. of Big Data Convergence, Hanshin University

요 약

본 연구는 청년층의 체크카드 소비 패턴과 금융상품 선호도 및 고객 이탈 간의 관계를 분석하기 위한 연구 방법을 제안한다. 선행연구에서는 대학생의 카드 사용 행태와 재무관리 행동, 청년층의 금융기관 선택 요인, 금융상품 이용 고객의 이탈 요인 등이 개별적으로 연구되어 왔다. 본 연구는 이러한 연구들을 바탕으로 청년층의 소비 패턴, 금융상품 선호도 및 고객 이탈 간의 연관성을 종합적으로 분석하고자 한다. 이를 통해 금융기관의 청년 고객 대상 맞춤형 상품 개발과 고객 유지 전략 수립에 활용 가능한 시사점을 제공하고자 한다.

키워드: 청년, 체크카드 소비, 금융상품 선호도, 고객 이탈, 금융소비자 행동

I. 서론

최근 청년층의 금융 활동이 증가함에 따라 체크카드 소비 데이터와 금융상품 이용 행태에 대한 관심이 높아지고 있다. 특히 청년층은 체크카드를 주요 결제수단으로 활용하고 있으며, 정부와 금융기관은 청년도약계좌, 청년주택드림청약통장 등 다양한 금융상품을 제공하고 있다. 그러나 금융상품 이용 과정에서 상품 해지나 타 금융 기관으로의 이동과 같은 고객 이탈 현상도 지속적으로 발생하고 있다. 따라서 청년층의 소비패턴이 금융상품 선호도와 고객 이탈에 어떠한 영향을 미치는지 분석할 필요가 있다. 본 연

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

구는 관련 선행연구를 바탕으로 청년층의 체크카드 소비 패턴과 금융상품 선호도 및 고객 이탈 간의 관계를 분석하기 위한 연구방법을 제안하고자 한다.

II. 관련 연구

2.1 국내 동향

윤하정,신민경,정순희(2012)는 대학생 소비자의 신용카드 및 체크카드 사용 실태와 재무관리행태를 분석하였다. 연구 결과 카드 사용 습관은 소비 행동 및 재무관리 수준과 밀접한 관련이 있으며, 체크카드 사용은 상대적으로 계획적인 소비 성향과 관련이 있는 것으로 나타났다. 이는 청년층의 소비 패턴이 금융 행동에 영향을 미칠 수 있음을 시사한다.

공병엽(2025)은 20대 청년층을 대상으로 금융기관 선택에 영향을 미치는 요인을 분석하였다. 연구 결과 서비스 편의성, 예·적금 상품 경쟁력, 보안 및 정보보호 등이 금융기관에 대한 신뢰와 지속 이용 의도에 영향을 미치는 것으로 나타났다. 이는 청년층이 금융상품 선택 시 다양한 요소를 고려함을 보여준다.

나광택 외(2020)는 증권사 고객 데이터를 활용하여 고객 이탈 예측 모델을 제안하였으며, 고객 거래 패턴과 자산 변 화가 이탈에 영향을 미친다고 분석하였다.

III. 제안방법

제안방법으로 청년층의 체크카드 소비 패턴과 금융상품 선호도 및 고객 이탈 간의 관계를 분석하기 위한 연구 방법을 제안한다. 먼저 체크카드 소비 데이터를 식비, 교통비, 쇼핑, 문화여가 등의 항목으로 구분하여 소비 패턴을 분석한다. 이후 청년도약계좌, 청년적금. 청약통장 등 청년 대상 금융상품의 이용 여부와 선호도를 조사한다. 또한 금융상품 해지 경험, 서비스 만족도, 혜택 수준 등을 분석하여 고객 이탈 요인을 파악한다. 이를 바탕으로 소비 패턴이 금융상품 선호도에 미치는 영향과 금융상품 선호도가 고객 이탈에 미치는 영향을 분석하고자 한다.

IV. 결론

본 연구는 청년층의 체크카드 소비 패턴과 금융상품 선호도 및 고객 이탈 간의 관계를 분석하기 위한 연구 방법을 제안하였다. 선행 연구 분석 결과 소비 패턴, 금융상품 선택, 고객 이탈은 서로 밀접한 관련이 있는 것으로 나타났고, 향후 본 연구를 통해 청년층의 금융 행동 특성을 파악하고 금융기관의 맞춤형 상품 개발 및 고객 유지 전략 수립에 활용할 수 있을 것으로 기대된다.

참고문헌

[1] 나광택, 이진영, 김은찬 and 이효찬. (2020). 증권 금융상품 거래 고객의 이탈 예측 및 원인 추론. 한국빅데이터학회 학회지, 5(2), 215-229.

[2] 공병엽. "금융기관 선택에 대한 인식 연구: 20대 청년층을 대상으로." 대한경영학회지 38.1 (2025): 65-95.

[3]윤하정(HaJungYoon),신민경(MinKyungShin),and정순희(SoonHee Jung). "대학생 소비자의 신용카드 및 체크카드 사용실태와 재무관리행태에 대한 연구." Financial Planning Review 5.3 (2012): 59-88.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

미국프로야구 MLB 투수 부상 위험 예측 모델 연구

김민준, 김유신
한신대학교 빅데이터융합학과
e-mail: lucas06@hs.ac.kr, yoosin25@hs.ac.kr

A Study on Predicting MLB Pitcher Injury Risk

Minjun Kim, Yoosin Kim
Dept. of Big Data Convergence, Hanshin University

요 약

본 연구는 MLB 투수의 Statcast 피치 트래킹 데이터를 활용하여 부상 위험을 사전에 예측하는 머신러닝 기반 모델을 제안한다. 투수 부상, 특히 토미 존 수술을 요하는 팔꿈치 내측 측부인대 손상은 MLB에서 지속적으로 증가하고 있으며 구단에 막대한 재정적 손실을 초래한다. 본 연구에서는 Baseball Savant의 Statcast 데이터를 기반으로 구속, 회전수, 릴리스 포인트, 구종 배합 변화 등 부상 전조 신호로 작용할 수 있는 변수들을 체계적으로 정의하고, XGBoost 기반 이진 분류 모델을 구축한다. SHAP 분석을 통해 부상 예측에 기여하는 핵심 변수 간의 상호작용을 해석하고, 실제 부상자 명단 등재 이력과 비교하여 성능을 검증한다. 본 연구는 데이터 기반의 투수 부상 예방 관리 체계 구축에 기여할 것으로 기대한다.

키워드: MLB 투수 부상 예측, Statcast, XGBoost, SHAP, 토미 존 수술

I. 서론

프로야구에서 투수 부상은 팀 전력에 치명적인 영향을 미치는 핵심 변수이다. MLB에서는 매 시즌 수백 명의 투수가 어깨·팔꿈치 부상으로 부상자 명단에 등재되며 이로 인한 구단의 재정적 손실과 전력 공백은 상당한 수준에 달한다. 특히 UCL 손상으로 인한 TJS는 평균 약 20.5개월의 회복 기간을 요하는 중증 부상으로 수술을 받은 투수 1인당 평균 약 190만 달러의 재정적 손실이 발생하는 것으로 추정된다. 또한 2023년 MLB의 피치 클락(pitch clock) 도입 이후 부상 양상의 변화가 관찰되고 있어, 부상 위험 요인에 대한 지속적인 모니터링의 필요성이 더욱 강조되고 있다.

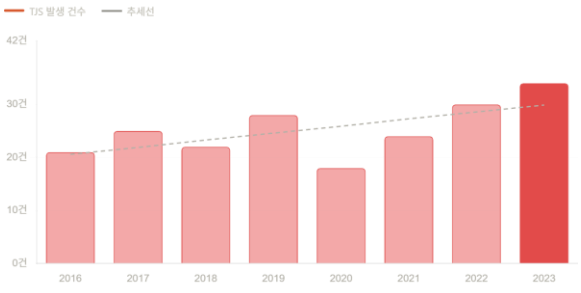
2015년 Statcast 시스템 도입 이후 MLB는 투구별 구속, 회전수, 릴리스 포인트, 투구 궤적 등 초정밀 피치 트래킹 데이터를 실시간으로 수집하고 있다. 이러한 데이터는 투수의 신체 이상 징후를 조기에 감지할 수 있는 중요한 단서를 포함하며 특히 구속의 점진적 감소나 릴리스 포인트의 하강은 UCL 손상의 전조 신호로 해석될 수 있다.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

그러나 기존 연구의 상당수는 다음 시즌 부상을 예측하는 시즌 단위 분석에 집중되어 있어 시즌 중 실시간 데이터를 활용한 조기 예측 모델에 관한 연구는 상대적으로 부족한 실정이다. 이에 본 연구는 Statcast 기반 피치 트래킹 데이터와 XGBoost 모델을 결합하여 투수 부상 위험을 사전에 예측하고 SHAP 분석을 통해 부상에 영향을 미치는 핵심 변수 간의 상호작용을 해석하고자 한다.

연도별 MLB 투수 팔꿈치 인대 손상(TJS) 건수 추이

출처: Kang et al. (2025), Journal of Big Data 참고 재구성 / 2016–2023



※ 2020년은 COVID-19로 인한 단축 시즌(60경기)으로 절대적 건수 감소. 피치 클락 도입 첫 해인 2023년에 역대 최고 수준 기록.

II. 관련 연구

2.1 국내 동향 피치 트래킹 데이터 기반 부상 예측 연구

Oeding et al은 2017년부터 2022년까지의 MLB 투수 데이터를 활용하여 XGBoost 기반 머신러닝 모델로 다음 시즌 어깨·팔꿈치 부상을 예측하였다. 65개 이상의 후보 변수 중 SHAP 분석을 통해 구속, 슬라이더 사용 비율, 직구 회전수, 직구 수평 무브먼트가 부상 예측의 핵심 변수임을 규명하였으며, 모델은 84%의 예측 정확도를 달성하였다. 특히 고속 슬라이더를 높은 비율로 구사하는 투수일수록 부상 위험이 유의미하게 높아지는 변수 간 상호작용이 두드러졌다.

2.2 피치 클락 도입과 부상 발생 변화

MLB는 2023년 경기 시간 단축을 위해 피치 클락을 도입하였다. 관련 연구에 따르면 피치 클락 도입 이후 전반적인 부상 건수는 감소하였으나 굴곡근 건·전완부 부상은 유의미하게 증가한 것으로 나타났다. 이는 투구 간격 단축으로 인한 특정 부위의 누적 피로 가중 가능성을 시사하며 부상 위험 변수 설계 시 규정 변화를 반영할 필요성을 보여준다.

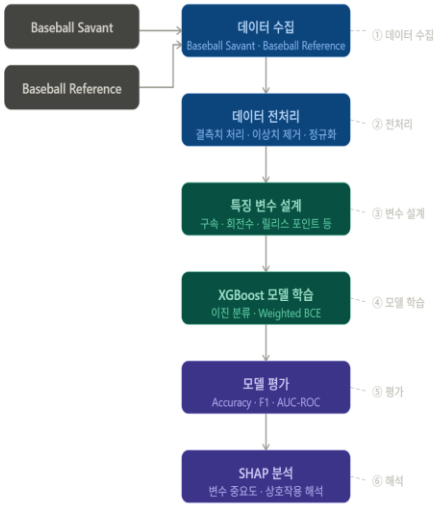
2.3 딥러닝 기반 TJS 위험 예측 연구

Kang et al. 은 2016년부터 2023년까지의 MLB 투구 데이터를 활용하여 딥러닝 프레임워크로 TJS 위험을 예측하였다. 분류 모델은 부상 발생 최대 100일 전에 F1-score 0.73으로 위험을 감지하였으며 회귀 모델은 수술까지 남은 일수를 R^2=0.79의 정확도로 추정하였다. SHAP 분석 결과 직구 릴리스 포인트 하강이 UCL 손상 가속의 핵심 역학적 지표로 확인되었다. 본 연구는 Kang et al.의 접근법을 참고하되, 딥러닝 대신 해석 가능성이 높은 XGBoost와 SHAP 기반 변수 상호작용 분석에 초점을 맞추어 실용적 적용 가능성을 높이고자 한다.

III. 제안방법

3.1 데이터 수집 및 변수 정의

본 연구에서는 Baseball Savant에서 2017년부터 2022년까지 MLB 선발 투수의 투구별 데이터를 수집한다. Oeding et al.의 데이터 구성을 참고하여 피치 트래킹 지표(구속, 회전수, 릴리스 포인트, 구종, 수평·수직 무브먼트), 투구량 지표(주간·월간 누적 투구 수), 인구통계 정보(나이, 투구 손잡이)를 포함한다. 부상 레이블은 어깨·팔꿈치 관련 IL 등재 이력을 기준으로 정의하며, 피치 클락 도입 전후의 부상 양상 변화를 고려하여 분석 구간을 2017~2022 시즌으로 제한한다.



3.2 특징 변수 설계

Oeding et al. 및 Kang et al.의 선행 연구를

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

바탕으로 다음과 같은 변수를 설계한다. 첫째, 구속 변화율로 직구 기준 최근 3경기 평균 구속과 시즌 초기 평균 구속의 차이를 산출한다. 둘째, 회전수 변화율로 직구 회전수의 주간 변화 추이를 반영한다. 셋째, 릴리스 포인트 수직 변화로 Kang et al.이 핵심 지표로 확인한 릴리스 포인트 하강 여부를 변수화한다. 넷째, 슬라이더 사용 비율로 Oeding et al.이 고위험 변수로 규명한 슬라이더 구사 빈도를 반영한다. 다섯째, 직구 수평 무브먼트로 팔의 컨디션 변화와 연관된 궤적 변화를 포착한다.

3.3 모델 구축 및 평가

XGBoost 이진 분류 모델을 구축하여 다음 30일 이내 IL 등재 여부를 예측한다. Oeding et al.의 방법론을 참고하여 SHAP 분석으로 변수 중요도 및 변수 간 상호작용 효과를 해석한다. 클래스 불균형 문제는 Weighted Binary Cross-Entropy를 적용하여 보완하며, Kang et al.이 제안한 방식과 동일하게 SMOTE 대비 성능 우위를 검증한다. 모델 성능은 정확도, 재현율, F1-Score, AUC-ROC를 기준으로 평가한다.

IV. 결론

본 연구는 MLB 투수의 Statcast 피치 트래킹 데이터를 활용하여 부상 위험을 사전에 예측하는 XGBoost 기반 모델을 제안하였다. Oeding et al.이 규명한 구속·슬라이더 비율·회전수 등의 핵심 변수와 Kang et al.이 확인한 릴리스 포인트 하강 지표를 통합적으로 반영하고, SHAP 분석을 통해 변수 간 상호작용 효과까지 해석함으로써 기존 연구의 한계를 보완하고자 하였다. 또한 피치 클락 도입 이후 부상 양상의 변화를 고려하여 분석 구간을 설정함으로써 연구의 현실적 타당성을 높였다.

본 연구의 결과는 구단 트레이너 및 코칭스태프가 투수의 신체 이상 징후를 조기에 감지하고 등판 스케줄을 조정하는 데 실질적인 의사결정 지원 도구로 활용될 수 있을 것이다. 향후 연구에서는 Kang et al.의 딥러닝 접근법을 결합하여 시즌 중 실시간 예측 시스템으로 확장하고, 피치 클락 도입 이후 데이터를 포함한 최신 분석으로 모델을 고도화 할 계획이다.

참고문헌

[1] J. F. Oeding, A. M. Boos, J. R. Kalk, D. Sorenson, F. M. Verhooven, G. Moatshe, and C. L. Camp, "Pitch-Tracking Metrics as a Predictor of Future Shoulder and Elbow Injuries in Major League Baseball Pitchers: A Machine-Learning and Game-Theory Based Analysis," Orthopaedic Journal of Sports Medicine, Vol. 12, No. 8, Aug. 2024.

[2] (저자명 확인 필요), "The Major League Baseball Pitch Clock: One Year Position Player Injury Analysis," The Physician and Sportsmedicine, Vol. 53, No. 2, pp. 147-151, Nov. 2024.

[3] B. Kang, M. Park, A. P. del Pobil, and E. Park, "Data-driven approaches for predicting Tommy John Surgery risk in major league baseball pitchers," Journal of Big Data, Vol. 12, No. 87, Apr. 2025.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

e스포츠 아시안 게임 국가대표 선출 지표 분석

김연승, 김유신
한신대학교 빅데이터융합학과
e-mail: kys0604@hs.ac.kr , yoosin25@hs.ac.kr

Indicators for the national team of e-sport players

Yeonseung Kim, Yoosin Kim
Dept. of Big Data Convergence, Hanshin University

요 약

최근 2026 아이치 나고야 아시안 게임의 e 스포츠 종목인 league of leagends(리그 오브 레전드)에서의 국가대표 명단에 따른 팬들의 의문이 많다. 차출에 이유와 기준을 명확하게 제시하였는데도 많은 이들이 예상한 선수와 달리 타 선수가 뽑히는 것에 대한 의문이 많았다. 본 연구는 이러한 상황 속에서 진실을 알아보고자 직접 기준표를 보며 명단을 다시 작성해본다. 기준표와 최근 선수들의 기량 등등을 고려하여 직접 점수를 산정하는 방식으로 선수들의 통계와 지표 분석을 한다. 하지만 저자의 생각만이 들어가게 된다면 객관적이 판단이 불허해지기 때문에 타국의 매체의 당일 경기 평점 등을 모두 고려하여 작성한다.

키워드: eSports, League of legends,국가대표, 차출 기준표, 지표 분석

I. 서론

최근 2026 아이치 나고야 아시안 게임의 e스포츠 종목인 리그 오브 레전드(league of legend)에서의 국가대표 명단에 따른 팬들의 의문이 많다. 한국e스포츠협회(KeSPA)가 명시한 표대로 제대로 선발이 되었는지, 그렇다면 왜 이 선수는 선발이 안되고 다른 선수가 선발이 되었는지 등의 말들이 다수 존재한다. KeSPA 외에도 경기력향상위원회, 대한체육회 또한 이러한 가담하였다.[1]

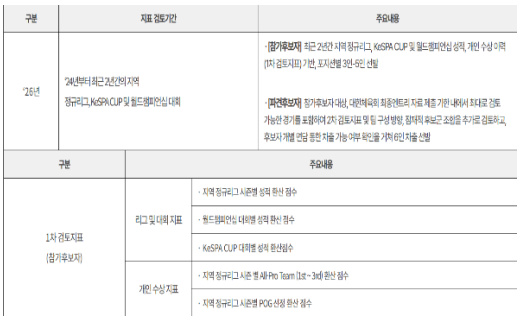
그렇다면 왜 이러한 게임 국가대표 선발에 이렇게나 많은 의문과 반발을 가지는가. 한국에 리그 오브 레전드(League of legends, 이하 LOL)은 2011년에 한국에 정식 출시된 이후로 수많은 게이머들에게 큰 인기를 받았고 지금까지도 꾸준히 가장 많은 인기를 보유한 게임이다. 그리고 또한 e스포츠 산업 또한 2015년 773억 원에서 연평균 10% 이상씩 성장하여 2022년에는 1514 억 원으로 2배

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

이상이 증가하는 모습 또한 보여졌다. 이러한 e스포츠는 젊은 인구층에게 많은 지지와 응원을 받고 있고 2018 자카르타 팔렘방 아시안 게임 시범 종목으로 채택되었던 LOL은 2022 항저우 아시안 게임부터 정식 종목으로 채택되어 지금 2026 아이치 나고야 아시안 게임까지 정식 종목으로 채택되어 지금까지 이어져 오고있다. 이러한 e스포츠의 인기로 국내외에 수많은 팬들이 생기고 그러한 팬들로 인한 여러가지의 논쟁은 지금까지도 끊이지를 않는다. [2] 이번 2026 아시안 게임 또한 이러한 팬들의 반발이 조금은 심한편이다. 국가대표 명단을 보며 “이 선수 말고 저 선수가 뽑히는게 맞는가” 등 여러 반응을 보이고 있다. 이러한 이유로 저자는 이 연구에서 직접 KeSPA가 제시한 선발 기준표를 보며 해당 사항에 맞는 선수들을 분석하여 선발 기준표의 정확성을 알아보고자 한다.

2.1 KeSPA의 선발 기준표

KeSPA는 아래와 같은 지표를 통하여 선수들을 차출 하였는데 그림에 포함된 내용으로 1차 검토지표를 확인하여 특정 선수들을 간추리고, 2차 검토지표로 명단을 확정짓는 방식이다.



구분	지표 검토기간	주요내용
'26년	'24년부터 최근 2년간의 지역 정규리그, KeSPA CUP 및 월드챔피언십 대회	· [참가후보자] 최근 2년간 지역 정규리그, KeSPA CUP 및 월드챔피언십 성적, 개인 수상 이력 (1차 검토지표) 기반, 포지션별 30인-6인 선발 · [파견후보자] 참가후보자 대상, 대한체육회 최종면트리 지표 제출 기한 내에서 최대로 검토 가능한 경기를 포함하여 2차 검토지표 및 팀 구성 방향, 참가팀 후보군 조합을 추가로 검토하고, 후보자 개별 면담 통한 차출 가능 이부 확인을 거쳐 6인 차출 선발

구분		주요내용
1차 검토지표 (참가후보자)	리그 및 대회 지표	· 지역 정규리그 시즌별 성적 환산 점수
		· 월드챔피언십 대회별 성적 환산 점수
		· KeSPA CUP 대회별 성적 환산점수
	개인 수상 지표	· 지역 정규리그 시즌 별 All-Pro Team (1st ~ 3rd) 환산 점수
		· 지역 정규리그 시즌별 POG 선정 환산 점수

그림 1. 세부 선발기준 및 경과 (1차 검토지표)[3]

1차 검토지표를 통해 이정도의 명단으로 간추릴 수 있다.

탑	Kiin	Zeus	Doran	Kingen	Perfect
정글	Canyon	Oner	Peanut	Cuzz	
미드	Faker	Chovy	Zeka	Bdd	Showmaker
원딜	Gumayusi	Viper	Ruler	Aiming	Peyz
서폿	Keria	Lehends	Delight	Peter	

III. 제안방법

3.1 차출 개요

선발 기준표를 토대로 점수를 산정하였을때 현 아시안게임 명단 “Zeus, Canyon, Faker, Zeka, Gumayusi, Keria” 로 구성된 현재 명단에는 더 이상에 이견을 제출할 필요가 없다고 판단한다. 그림 1에 명시된 내용 중 “개인 수상 지표”를 보았을때 “Kiin” 선수 차출에 대해 의문을 표할수 있으나, “Zeus” 선수 또한 2024년 월드챔피언십 우승 등을 포함한 국제전 기량 등이나 국내에서 보여주었던 활약과 최근 선수의 기량까지 모두 고려 하였을때 “Kiin” 선수보다 근소우위라고 판단한다. 이 외에도 “Canyon” 선수와 “Oner” 선수의 경우에는 최근에 보여줬던 개인 선수의 기량이나 최근 선수의 기량을 판단하였을때 “Canyon” 선수 또한 근소우위에 있다고 판단된다.

IV. 결론

본 논문에서는 각종 통계들과 자료들을 찾아보면서 직접 국가대표팀 차출에 대한 의구심을 가지고 스스로 산정하여 타 선수들의 미선발 이유를 찾아보았다. KeSPA는 모든 산정 방식을 모두 명시하였기에 그만큼 투명하게 산정하여서 또 다른 의견을 제출할수 없을 정도로 깔끔하였다.

참고문헌

[1] 이새하, 최은경. (2024). e스포츠는 공정한가?: 리그오브레전드의 공정을 위한 장치를 중심으로. 한국게임학회 논문지, 24(5), 27-37. 10.7583/JKGS.2024.24.5.27
[2] 박은수. (2023). 경기 결과가 팬의 감정에 미치는 향 : e스포츠 경기 하이라이트 영상 댓글 데이터를 중심으로 [석사학위논문, 성균관대학교].
[3]KeSPA / 제20회 아이치-나고야 하계아시아경기대회 e스포츠 국가대표 파견후보자 선발 공지 / 기관 공지사항 / 2026.05.15 / 2026.05.30 열람

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

제조 설비 시계열 데이터를 활용한 이상 탐지 분석

최민영*, 김유신
한신대학교 빅데이터융합학과
e-mail: cmy061126@hs.ac.kr, yoosin25@hs.ac.kr

Anomaly Detection Using Time Series Data from Manufacturing Equipment

Minyoung Choi, Yoosin Kim
Dept. of Big Data Convergence, Hanshin University

요 약

최근 스마트 제조 환경의 발전으로 인해 제조 설비에서 다양한 시계열 데이터가 지속적으로 생성되고 있다. 이러한 데이터에서 이상치를 탐지하는 것은 설비 고장을 사전에 예방하고 생산 효율성을 향상시키는 데 중요한 역할을 한다. 본 연구에서는 제조 설비에서 수집되는 시계열 데이터를 활용하여 이상치를 탐지하고 데이터 분석을 수행하는 방법론을 제시한다. 본 연구를 통해 시계열 데이터 기반 이상치 탐지 기법의 특성과 적용 가능성을 비교 분석하였으며, 제조 공정에서 데이터 분석의 중요성을 확인할 수 있었다. 향후 연구에서는 딥러닝 기반 기법을 포함한 보다 고도화된 분석 방법을 적용함으로써 이상치 탐지의 정확도를 더욱 향상시킬 수 있을 것으로 기대된다.

키워드: 머신러닝 기반 기법, 통계 기반 기법, 시계열 데이터, 제조설비, 고장 예측

1.서론

최근 스마트 제조 환경과 IoT 기술의 발전으로 인해 제조 설비에서는 다양한 센서기반 시계열 데이터가 지속적으로 생산되고 있다. 이러한 데이터는 설비의 상태를 실시간으로 반영할 수 있어 생산 공정의 모니터링과 품질 관리에 중요한 역할을 한다. 특히 설비의 이상 상태를 조기에 탐지하는 것은 고장 예방과 유지보수 비용 절감 측면에서 매우 중요하다. 제조 시계열 데이터의 이상치 탐지는 주로 통계 기반 방법과 머신러닝 기반 방법을 중심으로 연구되고 잇다. 통계 기반 방법은 평균과 표준편차를 활용하여 이상 여부를 판단하는 방식으로 구현이 간단하지만, 복잡한 데이터 패턴을 반영하는 데 한계가 있다, 반면 머신러닝 기반 방법은 isolation Forest, AutoEncoder 등 다양한 기법을 통해 통해 정교한 이상 탐지가 가능하지만, 데이터 의존성과 계산 비용이 증가하는 특징이 있다. 최근에는 제조 설비 데이터를 활용한 다양한 이상 탐지 연구가 진행되고 있으며, 준지도학습 기반 방법이나 생성적 딥러닝 모델을 활용한 접근 방식이 제안되고 있다.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

또한 여러 모델을 비교 분석하여 산업 환경에 적합한 방법을 찾고자 하는 연구도 활발히 이루어지고 있다. 따라서 본 연구에서는 제조 설비 시계열 데이터를 활용하여 통계 기반 방법과 머신러닝 기반 방법을 비교하고, 각 방법의 장단점을 분석하고자 한다. 이를 통해 실제 산업 환경에서 적용 가능한 효과적인 이상치 탐지 방법론을 제시하는 것을 목표로 한다.

II. 관련 연구

제조 설비 시계열 데이터를 활용한 이상치 탐지 연구는 일반적으로 데이터 수집, 전 처리, 이상치 탐지 모델 적용의 순서로 이루어진다. 초기 단계에서는 센서로부터 온도, 진동, 압력 등의 시계열 데이터를 수집하며, 데이터의 노이즈 제거 및 정규화를 통해 분석 가능한 형태로 변환한다. 이후 이상치 탐지 단계에서는 통계 기반 방법 또는 머신러닝 기반 모델을 적용하여 정상 데이터와 비정상 데이터를 구분한다. 통계 기반 접근에서는 평균과 표준편차를 활용한 관리 기준을 설정하여 기준 범위를 벗어나는 데이터를 이상치로 판단한다. 머신러닝 기반 접근에서는 학습된 모델을 통해 데이터의 패턴을 학습하고, 기존 분포와 다른 데이터를 이상치로 탐지한다. 최근 연구에서는 단일 모델 기반 접근의 한계를 보완하기 위해 다양한 이상치 탐지 방법을 비교하거나 결합하는 방식이 활용되고 있다. 또한 탐지 결과의 신뢰성을 높이기 위해 여러 모델의 결과를 교차 검증하거나 앙상블 형태로 활용하는 연구도 진행되고 있다. 본 연구에서는 이러한 기존 연구 흐름을 바탕으로, 제조 설비 시계열 데이터에 대해 통계 기반 방법과 머신러닝 기반 방법을 비교하여 각 접근 방식의 특징과 한계를 분석하고자 한다.

III. 제안방법

본 연구에서는 제조 설비로부터 수집된 시계열 데이터를 활용하여 이상치를 탐지하고 데이터 분석을 수행하기 위한 방법론을 제안한다. 제안 방법은 크게 데이터 수집 및 전처리 단계, 이상치 탐지 단계, 그리고 결과 분석 단계로 구성된다. 먼저 데이터 수집 및 전처리 단계에서는 제조 설비에서 발생하는 온도, 진동, 압력 등의 센서 데이터를 시계열 형태로 수집한다. 이후 결측값 처리 및 이상값 제거를 수행하고, 데이터의 스케일 차이를 보정하기 위해 정규화를 적용하여 분석에 적합한 형태로 변환한다. 다음으로 이상치 탐지 단계에서는 두 가지 접근 방법을 적용한다. 첫째, 통계 기반 방법으로 평균과 표준편차를 활용한 Z-score 방식을 적용하여 기준 범위를 벗어나는 데이터를 이상치로 판단한다. 둘째, 머신러닝 기반 방법으로 Isolation Forest 또는 One-Class SVM과 같은 비지도 학습 모델을 활용하여 데이터의 분포와 패턴을 기반으로 이상치를 탐지한다. 두 방법은 동일한 데이터에 적용하여 탐지 결과를 비교한다. 마지막으로 결과 분석 단계에서는 각 방법의 이상치 탐지 결과를 비교하여 탐지 민감도와 안정성을 평가한다. 이를 통해 통계 기반 방법과 머신러닝 기반 방법의 장단점을 분석하고, 제조 설비 데이터 특성에 적합한 이상치 탐지 접근 방식을 도출한다.

IV. 결론

본 연구에서는 제조 설비 시계열 데이터를 활용하여 이상치를 탐지하고 데이터 분석을 수행하기 위한 방법론을 제안하였다. 이를 위해 통계 기반 방법과 머신러닝 기반 방법을 동일한 데이터에 적용하여 각 접근 방식의 특징과 성능 차이를 비교하였다. 통계 기반 방법은 평균과 표준편차를 활용하여 이상치를 직관적으로 탐지할 수 있다는 장점이 있으나, 복잡한 데이터 패턴을 반영하는 데 한계가 있었다. 반면 머신러닝 기반 방법은 데이터의 분포와 구조를 학습하여 보다 정밀한 이상치 탐지가 가능하였지만, 상대적으로 높은 계싼 비용과 데이터 의존성을 갖는 특징이 있었다. 이러한 비교를 통해 각 방법은 서로 다른 장단점을 가지며, 제조 설비 데이터의 특성과 목적에 따라 적절한 방법을 선택하는 것이 중요함을 확인하였다. 따라서 단일 방법보다는 데이터 환경에 맞는 다양한 접근 방식을 고려하는 것이 효과적임을 알 수 있다. 향후 연구에서는 딥러닝 기반 모델을 포함한 보다 다양한 이상치 탐지 기법을 적용하고, 실시간 데이터 처리 환경에서의 적용 가능성을 검토할 필요가 있다.

참고문헌

[1] 김태현, 진교홍. (2024). 제조 시계열 데이터 이상 레이블 생성 및 준지도학습 기반 이상탐지. 한국정보통신학회논문지.
[2] 이재호, 조성훈, 유동현. (2023). 생성적 딥러닝 모형 기반 시계열 데이터의 이상치 탐지 비교 연구. 한국데이터정보과학회지.
[3] 박상민, 전채욱, 박주환, 옥승호. (2024). 시계열 데이터 이상치 탐지 모델 비교 및 디지털 트윈 활용. 한국정보기술학회논문지.
[4] 이준희, 권동우, 지영민. (2024). 전력 시계열 데이터를 이용한 비지도 설비 이상 패턴 탐지 방법. 한국통신학회논문지.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

가축 분뇨 퇴비의 부숙도 지수 예측을 위한 머신러닝 기반 회귀모델 개발 및 평가

강희준*(발표자), 강희용*, 김승완O
*(주)로봇팜
O(주)그린맥스
e-mail:koko1578@gmail.com*, rf7986@naver.com*, kswsamson@jbnu.ac.krO

Development and Evaluation of a Machine Learning-Based Regression Model for Predicting the Maturity Index of Livestock Manure Compost

HeeJun Kang*, HeeYong Kang*, SeungWan KimO
* Robotfarm Co., Ltd
O Greenmax Co., Ltd

요 약

본 연구는 가축분뇨 퇴비발효 시료(발효일수0-120일)에 대해 발효일수와 11개의 퇴비 상태변수(온도·수분·pH·C/N비·암모니아·질산염·총질소·총탄소·전기전도도·유기물)를 입력으로 사용하여 퇴비 부숙도 지수를 예측하는 11개의 머신러닝 회귀모델을 비교·평가하였다. 5-Layer 교차검증과 독립 테스트셋(20%) 평가결과, Extra Trees Regressor가 가장 우수한 성능(Test R^2 = 0.9137, RMSE = 4.87, MAE = 3.38)을 보였으며, Histogram Gradient Boosting(R^2 = 0.861)과 SVR-RBF(R^2 = 0.842)가 뒤를 이었다. 변수 중요도 분석(impurity-based 및 permutation importance) 결과, C/N 비(21.3%)·발효일수(15.8%)·수분함량(15.2%)·질산염(9.5%)이 부숙도를 결정하는 핵심 인자로 나타났으며, 이는 호기성 퇴비화의 생화학적 메커니즘과 일치한다. 본 모델은 실시간 센서 데이터를 활용한 퇴비 부숙도 자동 모니터링 시스템으로 활용 가능하다.

키워드: 가축분뇨 퇴비, 부숙도 지수, 머신러닝, Extra Trees, 변수중요도, C/N 비, 수분함량

I. 서론

가축분뇨는 한국을 비롯한 축산 선진국에서 가장 큰 규모의 농업부산물로, 적절히 처리되지 않을 경우 악취·온실가스·수질오염 등의 환경문제를 야기한다. 호기성 퇴비화(aerobic composting)는 가축분뇨를 안정적이고 위생적이며 작물에 안전한 유기질비료로 전환하는 가장 보편적인 자원화 기술이다. 그러나 미부숙(immature) 상태의 퇴비를 농지에 시용하면 식물독성(phytotoxicity), 질소고정화, 종자 발아 저해, 미생물 활성에 의한 작물

※ "이 성과는 정부(중소벤처기업부)의 재원으로 중소기업기술정보진흥원의 2024년도 중소기업기술혁신개발사업의 지원을 받아 수행된 연구임(과제 고유번호: S-3429612)"

뿌리 손상 등이 발생할 수 있어 부숙도(maturity)의 정량적 평가가 필수적이다.

전통적으로 부숙도는 종자발아지수(Germination Index, GI), C/N 비, 부식화정도(humification degree), T-value 등의 다중 지표를 조합하여 판정한다. 미국 캘리포니아 퇴비 품질협회(California Compost Quality Council) 기준에 따르면 C/N 비가 25 이하인 경우에 한해 추가 부숙도 등급을 부여하며, GI 80% 이상이 안정 퇴비의 기준이다. 그러나 이러한 지표는 측정에 수일~수주가 소요되고, 종자발아시험 등은 표준화가 어렵다는 한계를 갖는다.

이러한 한계를 극복하기 위해 최근에는 머신러닝 기반의 부숙도 예측모델이 활발히 연구되고 있다. Bioresource Technology(2024)에 게재된 연구에서는 6종의 머신러닝 알고리즘(RF, ET, XGBoost 등)을 활용하여 가축분뇨퇴비의 GI를 예측하였으며, Random Forest와 Extra Trees가 $R^2 > 0.9$의 우수한 성능을 보였다. Waste Management(2025)의 트리기반 모델 비교연구에서는 AdaBoost가 GI 예측에서 $R^2 = 0.9720$을달성했으며, Gini 지수와 SHAP 값을 통해 주요변수를 분석하였다.

본 연구는 국내 가축분뇨 퇴비 발효데이터 452개 시료를 대상으로, 발효일수와 발효 상태변수 만으로 부숙도 지수를 예측하는 종합회귀모델 비교 프레임워크를 제안한다. 구체적인 연구목적은 다음과 같다. 첫째, 선형모델, 트리 기반 앙상블, SVM, 신경망 등 11개 회귀 알고리즘의 성능을 동일조건에서 비교한다. 둘째, 5-Layer 교차검증과 독립 테스트셋 평가를 통해 일반화 성능과 안정성을 함께 평가한다. 셋째, Impurity-based 중요도 평가와 permutation importance를 활용하여 부숙도 결정의 핵심변수를 식별한다. 넷째, 잔차분석을 통해 모델의 편향성과 예측 신뢰구간을 진단한다.

II. 재료 및 방법

2.1 데이터 셋 개요

분석에 사용된 데이터셋은 가축분뇨퇴비의 호기성 발효과정에서 측정된 452개 시험데이터로 구성되며, 발효일수는 0일부터 120일에 걸쳐 36개의 시점이 포함된다. 입력변수(11개)는 발효일수(Day, 0-120일), 온도(℃), 수분함량(%), pH, C/N 비, 암모니아(mg/kg), 질산염(mg/kg), 총질소(%), 총탄소(%), 전기전도도(ms/cm), 유기물(%)이며, 타겟변수는 퇴비 부숙도 지수(0-100점)이다. T-value와 종자발아 지수(GI)는 부숙도 지수의 계산 구성요소이므로 데이터 누출을 방지하기 위해 입력에서 제외하였다. 결측치는 존재하지 않았으며, 모든 변수는 수치형이다. 표1은 분석에 사용된 변수의 기술통계량을 요약한 것이다.

표 1. 입력변수와 타겟변수의 기술통계 요약(n = 452)

변수	평균	표준편차	최솟값	중앙값	최댓값
Day (일)	14.03	12.15	0	12	120
Temperature (℃)	34.17	14.16	9.52	30.46	68.80
Moisture (%)	49.11	12.28	11.01	48.58	79.22
pH	7.39	0.85	4.28	7.39	9.37
C/N Ratio	19.70	6.57	4.63	18.43	39.72
Ammonia (mg/kg)	1,268.81	1,884.64	3.70	710.44	13,407.00
Nitrate (mg/kg)	460.85	840.50	0.00	247.89	6,048.23
Total N (%)	1.75	0.76	0.16	1.84	4.07
Total C (%)	31.47	11.56	1.13	32.67	58.83
EC (ms/cm)	2.74	1.94	0.69	2.00	11.14
Organic Material (%)	64.76	13.53	20.57	62.83	97.07
Maturity Index (target)	50.91	17.49	0.00	51.43	100.00

2.2 탐색적 데이터 분석(EDA)

각변수의 시간적 변화 양상을 파악하기위해 발효일수별 평균 및 표준편차를 계산하였고, 변수 간 상관관계를 Pearson 및 Spearman 계수로 분석하였다. 그림1(Figure 1)은 발효진행에 따른 주요 변수의 시간적 변화를 평균과 표준편차 띠로 나타

※ "이 성과는 정부(중소벤처기업부)의 재원으로 중소기업기술정보진흥원의 2024년도 중소기업기술혁신개발사업의 지원을 받아 수행된 연구임(과제 고유번호: S-3429612)"

낸 결과이다

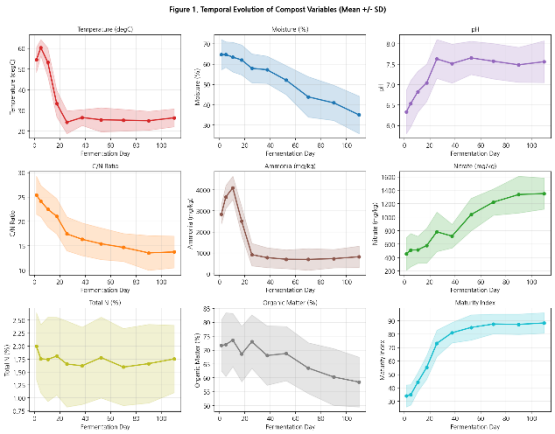


그림 1. 퇴비변수의 시간적 변화 특성(평균± 표준편차)

2.3 머신러닝 모델

부숙도 지수를 회귀 예측하기 위해 다음 11개의 알고리즘을 비교하였다. 즉, (1) Linear Regression, (2) Ridge Regression(α=1.0), (3) Lasso Regression(α=0.1), (4) k-Nearest Neighbors(k=5), (5) SVR-RBF(C=10), (6) Decision Tree(max_depth=10), (7) Random Forest(n=500), (8) Extra Trees(n=500), (9) Gradient Boosting(n=400, lr=0.05), (10) Histogram Gradient Boosting(max_iter=500), (11) MLP Neural Network(hidden=(64,32))이다. 이 중 거리·스케일에 민감한 선형·k-NN·SVR·MLP 모델에는 Standard Scaler를 파이프라인에 포함하여 입력변수의 평균을 0, 분산을 1로 표준화 하였다.

2.4 학습절차 및 성능 평가

전체 데이터 452개를 무작위로 학습셋(80%, n = 361)과 테스트셋(20%, n = 91)으로 분할하였다(random_state = 42). 학습셋에 대해 5-Layer 교차검증(KFold, shuffle = True)을 수행하였고, 최종모델은 학습셋 전체에 적용한 뒤 미사용 테스트셋에서 평가하였다. 성능지표는 결정계수(R^2), 평균 제곱근 오차(RMSE), 평균절대오차(MAE)를 사용하였다. 최우수 모델(Extra Trees)에 대해 후보 하이퍼 파라미터(n_estimators ∈ {300, 500}, min_samples_leaf ∈ {1, 2}, max_features ∈ {'sqrt', 0.5})를 5-겹CV로 비교하여 최종모델을 선정하였다.

III. 결과분석

3.1. 11개 모델의 성능 비교

11개 회귀모델의 5-Layer 교차검증 결과와 독립 테스트셋 성능을 표2에 정리하였다. 트리 기반 앙상블모델은 R^2 = 0.84–0.90 범위에서 일관되게 우수한 성능을 보였으며, Extra Trees가 가장 높은 CV R^2(0.879)와 Test R^2(0.900)을 동시에 달성하였다.

표 2. 11개 머신러닝 회귀모델의 성능 비교

순위	모델	CV R^2 (±std)	CV RMSE	Test R^2	Test RMSE	Test MAE
1	Extra Trees	0.879 ± 0.026	6.07	0.9002	5.24	3.63
2	Hist Gradient Boosting	0.838 ± 0.017	7.02	0.8606	6.19	4.34
3	SVR (RBF)	0.812 ± 0.020	7.57	0.8422	6.59	4.15
4	Gradient Boosting	0.863 ± 0.035	6.45	0.8381	6.67	4.08
5	Random Forest	0.851 ± 0.045	6.70	0.8366	6.70	4.28
6	k-NN (k=5)	0.823 ± 0.045	7.32	0.8336	6.76	4.79
7	MLP (Neural Network)	0.682 ± 0.043	9.83	0.8133	7.17	5.18
8	Decision Tree	0.743 ± 0.064	8.82	0.7987	7.44	5.21
9	Ridge Regression	0.751 ± 0.041	8.69	0.7600	8.12	6.21
10	Lasso Regression	0.753 ± 0.039	8.65	0.7599	8.13	6.20
11	Linear Regression	0.750 ± 0.042	8.69	0.7598	8.13	6.22

※ "이 성과는 정부(중소벤처기업부)의 재원으로 중소기업기술정보진흥원의 2024년도 중소기업기술혁신개발사업의 지원을 받아 수행된 연구임(과제 고유번호: S-3429612)"

3.2 하이퍼 파라미터 튜닝 결과

최우수 모델 Extra Trees에 대해 후보 하이퍼 파라미터 조합을 비교한 결과, n_estimators = 300, min_samples_leaf = 1, max_features = 0.5의 조합이 5-Layer CV R^2 = 0.8815로 최적인 것으로 나타났다. 표3은 튜닝된 최종 Extra Trees 모델의 학습셋, 교차검증, 테스트셋에서의 성능을 정리한 것이다.

표. 3. 튜닝된 최종 Extra Trees 모델의 성능 지표

지표	학습셋(n=361)	테스트셋(n=91)
R^2	1.0000	0.9137
RMSE	0.00	4.87
MAE	0.00	3.38

3.3 예측-실측 산점도 및 잔차분석

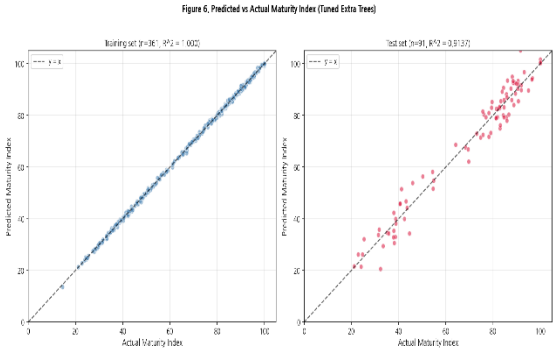


그림 2. 튜닝된 Extra Trees 모델의 예측값 vs 실측값 비교

그림2에서 점들이 대각선(y = x) 주위에 좁은 띠를 형성하고 있으며, 부숙도 지수의 전범위(0-100)에 걸쳐 균형있는 예측능력을 보여준다. 특히 부숙도가 매우 낮은 영역(0-20)과 매우 높은 영역(80-100)에서도 체계적 편향없이 예측이 이루어졌으며, 이는 트리 앙상블의 구간별 분할특성이 S-자 형태의 부숙 동역학을 효과적으로 포착한 결과로 해석된다

3.4 변수중요도 분석

그림 3에서 확인되듯 상위 3개 변수(C/N 비, 발효일수, 수분함량)가 전체permutation importance 합의 약 60%를 차지하며, 이는 호기성 퇴비화의 핵심 메커니즘과 부합한다. C/N 비는 미생물 활성을 직접적으로 제어하는 영양 비율이며, 수분함량은 호기성 미생물 활성에 필수적이고, 발효일수는 시간적 진행자체를 반영하는 강력한 단순지표다. 질산염(4위, 9.5%)은 부숙 후반부의 질산화를 반영하고, 암모니아(5위, 3.6%)는 미부숙 단계의 식물독성과 관련된 변수이다. pH·총탄소·전기전도도는1% 수준으로 매우 낮은 중요도를 보였는데, 이는 이들 변수가 다른변수(C/N, 유기물, 암모니아등)와 강한 공선성을 갖기 때문에 단독 기여도가 흡수된 결과로 해석된다.

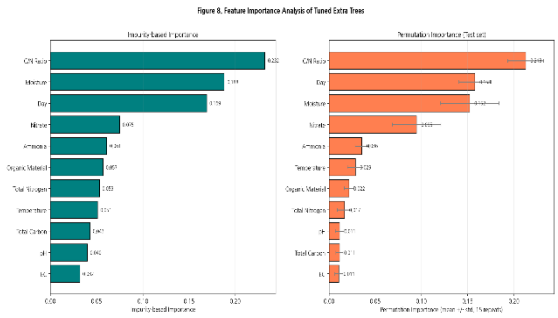


Figure 8. 튜닝된Extra Trees의변수중요도분석

(좌: Impurity-based, 우: Permutation importance)

IV. 결론

본 연구는 가축분뇨 퇴비 발효 데이터 452개 시료에 대해 11개의 머신러닝 회귀모델을 비교 평가하여, 튜닝된 Extra Trees Regressor가 테스트셋 R^2 = 0.9137, RMSE = 4.87, MAE = 3.38의 우수한 성능으로 부숙도 지수를 예측함을 보였다. 트리 기반 앙상블이 일관되게 선형모델을 초과하는 성능을 보였으며, 이는 부숙도와 발효상태 변수간의 본질적 비선형성을 시사한다. 변수 중요도 분석에서 C/N 비, 발효일수, 수분함량, 질산염이 부숙도 예측의 핵심 4대인자로 도출되었으며, 이는 호기성 퇴비화의 생화학 메커니즘 및 기존 문헌과 정합한다. 본 모델은 IoT 센서 기반 실시간 모니터링과 결합되어 농가 단위의 퇴비부숙도 자동 평가시스템 구축에 직접 활용될 수 있을 것으로 기대된다.

※ "이 성과는 정부(중소벤처기업부)의 재원으로 중소기업기술정보진흥원의 2024년도 중소기업기술혁신개발사업의 지원을 받아 수행된 연구임(과제 고유번호: S-3429612)"

중환자실 환자의 진정 · 진통 약물 처방 패턴을 활용한 장기 재원 예측 모델 개발

이승희*, 이지훈*, 배지원*, O
*건양대학교병원 의료데이터연구단
O건양대학교병원 의료데이터연구단
e-mail: {shleedynamics, 602608}@kyuh.ac.kr*, 602648@kyuh.ac.krO

Development of a Model for Predicting Length of Stay Using Prescription Patterns of Sedatives and Analgesics for ICU Patients

Lee Seunghee*, Lee Ji Hoon*, Bae Ji WonO
*, OKYMERA(KonYang MEdical data ReseArch group, Konyang University Hospital

요 약

본 논문은 K-MIMIC 데이터를 활용하여 48시간 이상 인공호흡기 치료를 받은 중환자실 환자의 진정제 및 진통제 처방 패턴을 분석하고, 이를 기반으로 환자의 장기 재원(7일 이상) 예측모델을 개발하고자 하였다. 2012년부터 2021년까지 수집된 ICU 환자 데이터를 대상으로 다양한 머신러닝 모델을 비교한 결과 CatBoost 모델이 AUROC 0.854 (95% CI: 0.836- 0.873), AUPRC 0.717, F1-score 0.802로 가장 우수한 성능을 보였으며, Dexmedetomidine 및 Remifentanil 처방 관련 변수와 opioid 및 benzodiazepine 사용 부담이 주요 예측 인자로 확인되었다. 본 연구는 중환자실 환자의 진정 · 진통 약물 사용 패턴이 환자 예후 예측에 유의한 정보를 제공할 수 있음을 제시하며, 향후 데이터 기반 중환자 치료 의사결정 지원 시스템 개발에 활용될 수 있을 것으로 기대된다.

키워드: K-MIMIC, ICU, Long Ventilation, Prescription, Machine Learning, Pattern Analysis

I. 서론

중환자실(ICU)에서 인공호흡기 치료를 받는 환자의 적절한 진정(sedation) 및 진통(analgesia)관리는 환자 안전과 임상 예후에 중요한 요소이다. 과도한 진정은 섬망, 인공호흡기 의존 기간 증가 및 재원 기간 연장과 연관되며, 반대로 불충분한 진정 및 진통은 환자의 불안, 통증 및 인공호흡기 비동조를 유발할 수 있다. 또한, 최근 중환자 치료에서는 과거의 깊은 진정(Deep sedation) 중심 전략에서 벗어나 환자 맞춤형 진정 전략과 Light

※This research was supported by a grant of the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (Grant Number: RS-2021-KH114109)

sedation 중심의 관리가 강조되고 있다.

본 연구는 이러한 임상적 흐름에 맞추어 K-MIMIC 데이터를 활용하여 48시간 이상 인공호흡기를 착용한 중환자실 환자의 진정제 및 진통제 처방 패턴을 분석하고, 이를 기반으로 환자 장기 재원을 예측할 수 있는 모델을 개발하고자 한다.

II. 관련 연구

2.1 국내 동향

국내에서는 인공호흡기 치료 환자의 진정 및 진통 관리와 관련하여 다기관 기반 연구와 임상 가이드라인 개발이 지속적으로 이루어지고 있다. 2021년 대한중환자실의학협회(KSCCM)에서는 중환자실 통증, 진정, 섬망 관리 가이드라인을 통해 환자 상태를 지속적으로 모니터링하며 가능한 얕은 진정(Light sedation)을 유지할 것을 권고하였다[1]. 또한, 국내 연구에서는 깊은 진정(Deep sedation)이 기계 환기 기간 연장과 관련될 수 있다고 보고하였다[2].

그러나 기존 연구는 주로 진정 깊이와 임상 결과 간의 연관성 분석에 초점을 두고 있으며, 실제 임상 현장에서 사용되는 진정제 및 진통제 처방 패턴을 기반으로 한 예측모델 연구는 아직 제한적인 실정이다. 따라서 본 연구는 K-MIMIC 데이터를 이용하여 48시간 이상 인공호흡기 치료를 받은 중환자실 환자의 진정제 및 진통제 처방 패턴을 분석하고, 이를 기반으로 예측모델을 개발하고자 한다.

III. 제안방법

3.1 데이터셋

본 연구에서는 건양대학교병원 중환자실(ICU)에서 2012.01.01 ~ 2021.12.31까지 총 10년간 수집된 K-MIMIC 데이터를 활용하였다. 연구 대상자는 ICU 입원 기간 중 48시간 이상 인공호흡기 치료를 받은 18세 이상 환자 중 진정제 및 진통제를 처방받은 환자 470명으로 정의하였다.

대상 환자의 인구학적 정보, 입원 정보, 검사 결과, 인공호흡기 관련 정보 및 약물 처방 데이터를 추출하였다. 주요 약물군은 중환자실에서 주로 사용되는 진정제(Dexmedetomidine, Midazolam, Lorazepam, Diazepam, Propofol)와 진통제(Remifentanil, Ketamine, Fentanyl, Morphine)로 구성하였다.

3.2 방법론

본 연구에서는 중환자실 환자의 진정제 및 진통제 처방 패턴을 기반으로 장기 재원(Prolonged length of stay) 여부를 예측하기 위한 머신러닝 모델을 구축하였다. 장기 재원은 선행 중환자실 연구에서 제시된 기준을 참고하여 중환자실 재원기간이 7일 이상인 경우로 정의하였다[3]. 모델로는 Logistic Regression, XGBoost, Random Forest, LightGBM, CatBoost를 사용하였으며, 클래스 불균형 문제를 보완하기 위해 None, SMOTE, ADASYN 샘플링 기법을 적용하였다. 모델 성능 평가는 AUROC, AUPRC, F1-score를 이용하여 수행하였다.

모델 학습에는 환자의 인구학적 정보, 검사 결과, 인공호흡기 관련 정보 및 진정제·진통제 처방 데이터를 활용하였으며, 약물의 종류, 처방 횟수, 투여 패턴 및 약물군별 burden 변수를 주요 feature로 구성하였다.

※This research was supported by a grant of the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (Grant Number: RS-2021-KH114109)

3.3 연구 결과

1) 약물별 처방 패턴 및 Los 연관성

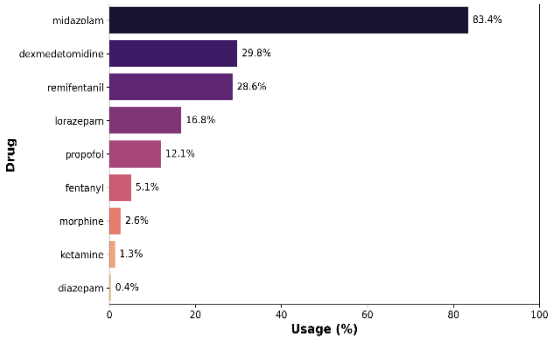


그림 1. Frequency of Sedative and Analgesic Use Across ICU Stays

Midazolam이 가장 많이 사용된 진정제(83.4%)였으며, 그 다음으로 Dexmedetomidine (29.8%)과 Remifentanil (28.6%)이 뒤를 이었다. 이러한 결과는 기계적 인공호흡을 받는 중환자실 환자들에게서 벤조디아제핀 기반 진정 요법이 주된 전략이었음을 시사한다.

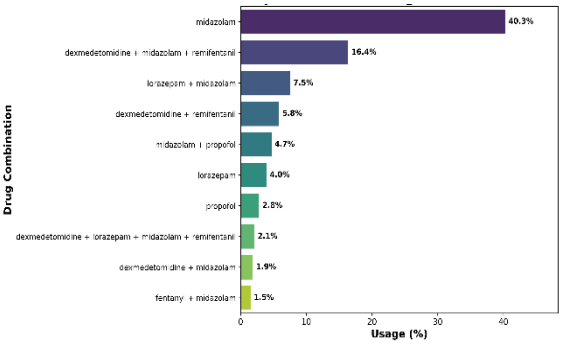


그림 2. Top 10 sedation-Analgesia Combination Patterns

Midazolam 단독 요법이 가장 흔한 진정 패턴(40.3%)이었으며, Dexmedetomidine + Midazolam + Remifentanil과 같은 다중 모드 병용 요법도 빈번하게 관찰되었다.

표 1. Association Between Sedation-Analgesia exposure and ICU LOS

Exposure	Yes n (%)	No n (%)	Median LOS Yes [IQR]	Median LOS No [IQR]	p-value
used_benzo	469 (88.3%)	62 (11.7%)	12.4 [6.0-22.3]	7.1 [4.6-12.6]	<0.001
used_dex	158 (29.8%)	373 (70.2%)	8.3 [4.7-17.8]	13 [6.0-23.5]	<0.001
used_propofol	64 (12.1%)	467 (87.9%)	14.8 [5.5-25.2]	11.2 [5.7-20.3]	0.387
used_opioid	184 (34.7%)	347 (65.3%)	9.2 [4.7-18.0]	12.8 [6.0-23.1]	0.003
multidrug	132 (24.9%)	399 (75.1%)	10.5 [4.6-21.1]	11.8 [6.0-21.1]	0.158

벤조디아제핀 노출은 중환자실 입원 기간 연장과 관련이 있었던 반면, Dexmedetomidine 노출은 입원 기간 단축과 관련이 있었다. 이러한 결과는 진정 전략이 기계적 인공호흡을 받는 중환자실 환자의 임상 결과에 영향을 미칠 수 있음을 시사한다.

2) 모델 결과

표 2. Model-specific sampling and cross-validation results

Sampler	Model	AUROC (95% CI)	AUPRC	F1-score
None	CatBoost	0.717 (0.836-0.873)	0.854	0.802
SMOTE	RandomForest	0.717 (0.836-0.864)	0.850	0.666
ADASYN	RandomForest	0.717 (0.836-0.860)	0.848	0.701

총 3가지 샘플링 기법(None, SMOTE, ADASYN)에 대해 다양한 머신러닝 모델을 비교한 결과, 샘플링을 적용하지 않은 CatBoost 모델이 가장 우수한 성능을 보였다. 해당 모델은 AUROC 0.854 (95% CI: 0.836-0.873), AUPRC 0.717, F1-score 0.802를 기록하여 다른 모델 대비 전반적으로 높은 예측 성능을 나타냈다.

※This research was supported by a grant of the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (Grant Number: RS-2021-KH114109)

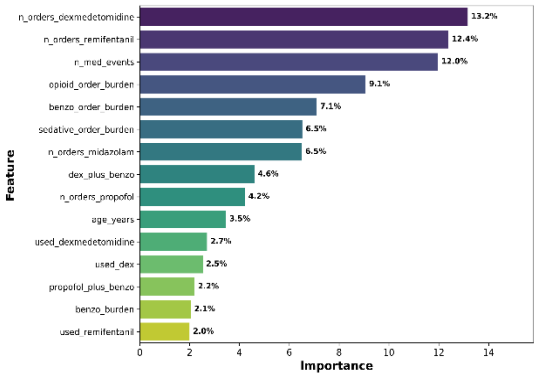


그림 3. CatBoost Feature Importance Top 15

Feature importance 분석 결과, Dexmedetomidine 처방 횟수(n_orders_dexmedetomidine)와 Remifentanil 처방 횟수(n_orders_remifentanil)가 가장 높은 중요도를 보였으며, 전체 약물 이벤트 수(n_med_events), opioid 처방 부담(opioid_order_burden), benzodiazepine 관련 변수 등이 주요 예측 인자로 확인되었다. 이는 중환자실 환자의 진정제 및 진통제 사용 패턴이 임상 결과 예측에 중요한 영향을 미칠 수 있음을 시사한다.

IV. 결론

본 연구에서는 K-MIMIC 데이터를 활용하여 48시간 이상 인공호흡기 치료를 받은 중환자실 환자의 진정제 및 진통제 처방 패턴 기반 7일 이상 재원 예측모델을 개발하였다. 다양한 머신러닝 모델 중 CatBoost 모델이 가장 우수한 예측 성능을 보였으며, Dexmedetomidine 및 Remifentanil 처방 관련 변수와 opioid 및 benzodiazepine 사용 부담이 주요 예측 인자로 확인되었다. 본 연구 결과는 중환자실 환자의 진정·진통 약물 사용 패턴이 임상 결과 예측에 유의한 정보를 제공할 수 있음을 시사하며, 향후 데이터 기반 중환자 치료 의사결정 지원 시스템 개발에 활용될 수 있을 것으로 기대된다. 다만, 본 연구는 단일기관 데이터를 기반으로 수행되었으므로, 모델의 일반화 가능성을 확인하기 위해 다기관 데이터를 활용한 추가 검증이 필요하다.

참고문헌

[1] Seo, Y., Lee, H. J., Ha, E. J., & Ha, T. S. (2022). 2021 KSCCM clinical practice guidelines for pain, agitation, delirium, immobility, and sleep disturbance in the intensive care unit. *Acute and critical care*, *37*(1), 1–25. https://doi.org/10.4266/acc.2022.00094

[2] Hyun, D. G., Ahn, J. H., Gil, H. Y., Nam, C. M., Yun, C., Lee, J. M., Kim, J. H., Lee, D. H., Kim, K. H., Kim, D. J., Lee, S. M., Ryu, H. G., Hong, S. K., Kim, J. B., Choi, E. Y., Baek, J., Kim, J., Kim, E. J., Park, T. Y., Kim, J. H., … Lim, C. M. (2023). The Profile of Early Sedation Depth and Clinical Outcomes of Mechanically Ventilated Patients in Korea. *Journal of Korean medical science*, *38*(19), e141. https://doi.org/10.3346/jkms.2023.38.e141

[3] Lefering R, Waydhas C, TraumaRegister DGU. Prediction of prolonged length of stay on the intensive care unit in severely injured patients—a registry-based multivariable analysis. Frontiers in Medicine. 2024;11:1358205.

※This research was supported by a grant of the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (Grant Number: RS-2021-KH114109)

터치스크린과 가상 교실: 감염병 대유행이 견인한 영유아 발달 치료의 디지털 패러다임 쉬프트

유영의*, 장윤호*, 조보경O, 김미량*
*순천향대학교 유아교육과, *청운대학교 IR성과관리센터, O순천향대학교 휴먼헬스케어융합연구소, *순천향대학교 스포츠응용산업학과
e-mail:{조보경} sch7307@sch.ac.krO

Touchscreens and Virtual Classrooms: The Digital Paradigm Shift in Early Childhood Developmental Therapy Driven by the Pandemic

Yoo, YungEui*, Jang, YunHo*, Cho, BoKyoungO, Kim Mi-Ryang*
*Early Childhood Education, Soonchunhyang University, *Institutional Research & Performance Management Center, Chungwoon University, OHuman Healthcare Convergence Institute, Soonchunhyang University, *Dept. of Sport Industry Studies, Soonchunhyang University,

요 약

본 연구는 코로나19 팬데믹 전후 국내 영유아 발달지연 치료 연구의 변화 양상을 분석하고, 디지털 기반 중재 연구의 확산 경향을 탐색하고자 하였다. 이를 위해 RISS 데이터베이스를 활용하여 2020년을 기준으로 팬데믹 이전(2015~2019년, 133편)과 이후(2020~2024년, 93편)의 국내 학술지 게재 논문 총 226편을 수집하였다. 분석은 K-DST 기준과 지원 인프라를 결합한 7개 범주(언어, 인지, 사회정서, 대소근육, 자조, 부모지원, 교사지원)를 기반으로 수행하였으며, 87개 키워드 사전을 활용하여 빈도 분석과 카이제곱 독립성 검정을 실시하였다. 분석 결과, 팬데믹 전후 범주별 분포 차이는 통계적으로 유의하지 않은 것으로 나타났다(χ^2=5.952, p=.428). 이는 대분류 수준에서 연구 관심 영역의 전체적 구조가 비교적 안정적으로 유지되었음을 시사한다. 다만 범주별 비율 변화와 표준화 잔차를 검토한 결과, 언어 영역은 26.3%에서 17.2%로 감소하였으며(표준화 잔차 −1.613), 인지 및 교사지원 영역은 상대적으로 증가하는 경향이 확인되었다. 또한 일부 연구에서는 태블릿 기반 인지훈련, 비대면 놀이중재, 모바일 기반 부모지원 등 디지털 기술을 활용한 중재 사례가 확대되는 양상이 나타났다. 한편 자조 영역은 전체 연구의 1.8% 수준에 머물러 다른 영역에 비해 연구 축적이 제한적인 것으로 확인되었다. 본 연구는 팬데믹 전후 영유아 발달지연 치료 연구의 분포와 변화 양상을 체계적으로 제시하고, 향후 디지털 기반 중재 연구의 발전 방향과 자조 영역에 대한 후속 연구 필요성을 제안하였다는 점에서 의의를 가진다.

키워드: 발달지연, 코로나19 팬데믹, 디지털 치료, 비대면 중재, 가상 가이드라인

※ 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2025S1A5C2A02022632)

I. 서론

2020년 코로나19 팬데믹에 따른 보육기관 운영 제한과 사회적 거리두기는 영유아의 일상적 상호작용 환경과 발달 지원 체계에 다양한 변화를 초래하였다(최윤경 외, 2024; World Health Organization, 2020). 특히 마스크 착용, 대면 활동 축소, 기관 이용 제한 등은 영유아의 언어적·사회적 경험의 양상을 변화시켰으며, 일부 연구에서는 이러한 환경 변화와 아동 발달지표 간의 관련 가능성이 보고되었다(Lee et al., 2024). 이에 따라 아동 연구 및 치료 현장 역시 운영 방식과 중재 환경의 조정을 경험하였으며, 기존의 대면 중심 관찰 및 치료 접근에 대한 대안적 시도가 확대되었다(Shuffrey et al., 2022). 이러한 변화 속에서 디지털 기술을 활용한 원격 중재, 가상 플랫폼 기반 치료, 스마트 기기 활용 프로그램 등이 점차 연구와 실천 영역에서 주목받기 시작하였다. 다만 이러한 변화가 실제로 국내 영유아 발달지연 치료 연구의 관심 영역과 연구 주제 분포에 어떠한 변화를 가져왔는지에 대한 종합적 검토는 아직 충분히 이루어지지 않았다. 이에 본 연구는 국내 영유아 발달지연 관련 학술논문을 대상으로 팬데믹 이전과 이후의 연구 동향을 비교·분석하고, 발달 영역 및 지원 체계별 분포 변화와 디지털 기반 중재 연구의 등장 양상을 탐색하고자 한다. 이를 통해 향후 영유아 발달지원 정책과 연구 방향 설정을 위한 기초자료를 제공하고자 한다.

II. 연구 방법 및 분석

최종 선정된 국내 학술논문 226편을 대상으로 K-DST 범주 및 지원 인프라 체계에 근거하여 언어, 인지, 사회정서, 대소근육, 자조, 부모지원, 교사지원 영역으로 분류하였다. 이후 팬데믹 전후 시기별 연구 분포의 차이를 확인하기 위해 빈도 분석과 카이제곱 독립성 검정을 실시하였다.

분석 결과, 팬데믹 이전과 이후의 발달 영역 및 지원체계별 분포 차이는 통계적으로 유의하지 않은 것으로 나타났다($\chi^2 = 5.95$, $df = 6$, $p = .43$). 이는 대분류 수준에서 연구 관심 영역의 전체적 구성에는 유의한 변화가 확인되지 않았음을 의미한다. 다만 범주별 비율과 표준화 잔차를 함께 검토한 결과, 일부 영역에서는 상대적인 변화 양상이 관찰되었다. 언어 영역은 표준화 잔차 −1.613으로 감소 경향을 보였으며, 인지 영역(+1.535)과 교사지원 영역(+1.219)은 상대적으로 증가하는 양상이 확인되었다. 이러한 결과는 팬데믹 이후 일부 연구에서 디지털 기술 활용, 원격 중재, 지원체계 확장 등에 대한 관심이 확대되었을 가능성을 시사한다.

따라서 본 연구 결과는 팬데믹 이후 영유아 발달지연 연구 전반의 구조적 변화보다는, 특정 영역을 중심으로 한 연구 주제의 점진적 재배치와 관심 변화 가능성을 보여주는 탐색적 결과로 해석할 필요가 있다.

표1. 코로나19 팬데믹 전후 발달 영역 및 지원체계별 학술논문 동향 분포

발달 영역	이전 빈도	이전 비율	이후 빈도	이후 비율	전체 빈도	전체 비율	증감폭
언어	35	26.3%	16	17.2%	51	22.6%	-9.1%p
인지	6	4.5%	9	9.7%	15	6.6%	+5.2%p
사회정서	36	27.1%	24	25.8%	60	26.5%	-1.3%p
대소근육	15	11.3%	11	11.8%	26	11.5%	+0.5%p
자조	2	1.5%	2	2.2%	4	1.8%	+0.7%p
부모지원	15	11.3%	8	8.6%	23	10.2%	-2.7%p
교사지원	24	18.0%	23	24.7%	47	20.8%	+6.7%p
합계	**133**	**100.0%**	**93**	**100.0%**	**226**	**100.0%**	**0.0%p**

III. 연구 결과 및 디지털 패러다임 분포

※ 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2025S1A5C2A02022632)

팬데믹 이후 영유아 발달지연 연구에서는 일부 영역을 중심으로 디지털 기술 활용과 비대면 중재 관련 연구가 증가하는 양상이 관찰되었다. 다만 이러한 변화는 통계적으로 유의한 구조 변화로 해석하기보다는, 연구 관심의 상대적 이동과 새로운 중재 방식에 대한 탐색 확대의 관점에서 이해할 필요가 있다.

영역별 변화를 살펴보면, 언어 영역은 전체 비율 기준 가장 큰 감소폭(−9.1%p)을 보였으며, 일부 연구에서는 기존 대면 중심 접근을 보완하기 위한 모바일 애플리케이션 기반 부모 교육 및 원격 지원 방식이 제안되었다. 인지 영역(+5.2%p)에서는 디지털 기반 인지훈련 프로그램과 비대면 적용 가능성에 대한 연구가 증가하는 경향을 보였다. 또한 사회정서 영역에서는 태블릿 기반 놀이중재, 비대면 부모-아동 상호작용 프로그램 등 다양한 접근이 확인되었으며, 교사지원 영역(+6.7%p)은 조기발견과 기관 기반 지원체계에 대한 관심 확대와 함께 상대적 증가 양상을 나타냈다. 이러한 결과는 팬데믹 이후 일부 연구에서 디지털 기술을 활용한 중재 환경과 지원 체계에 대한 관심이 확대되었을 가능성을 시사한다. 한편 자조 영역은 전체 연구의 1.8% 수준으로 다른 영역에 비해 상대적으로 낮은 연구 비중을 보였다. 이는 자조 관련 중재와 지원 전략에 대한 후속 연구가 필요함을 보여주는 결과로 해석될 수 있다.

참고문헌

[1] 최윤경, 조숙인, 박진아, 배윤진, 정윤경, 이기재, 김영민, "코로나 세대 (Covid-Generation) 아동의 발달 추적 연구 (Ⅰ)" 육아정책연구소 연구보고서, pp. 1-250, Dec. 2024.

[2] World Health Organization, "WHO Director-General's opening remarks at the media briefing on COVID-19" World Health Organization Speeches, Vol. 1, No. 1, pp. 1-3, Mar. 2020.

[3] Lee, K. S., Choi, Y. Y., Kim, Y. S., Kim, Y., Kim, M. H., & Lee, N., "Association between the COVID-19 pandemic and childhood development aged 30 to 36 months in South Korea, based on the National health screening program for infants and children database" BMC Public Health, Vol. 24, No. 1, pp. 989-1002, Apr. 2024.

[4] Shuffrey, L. C., Firestein, M. R., Kyle, M. H., Fields, A., Alcántara, C., Amso, D., ... & Dumitriu, D., "Association of birth during the COVID-19 pandemic with neurodevelopmental status at 6 months in infants with and without in utero exposure to maternal SARS-CoV-2 infection" JAMA Pediatrics, Vol. 176, No. 6, pp. 215563-215572, June 2022.

[5] 최은영, 김은영, 김태우, 윤정환, 이영애, 박희현, 최윤경, "코로나 19 이후 영유아 발달 및 학습 격차 해소 방안 연구 (Ⅱ)" 육아정책연구소 연구보고서, pp. 1-180, Nov. 2023.

※ 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2025S1A5C2A02022632)

모바일 브라우저 YOLO26 객체 탐지 파이프라인의 WebGPU-WASM 지연 분해 및 병목 분석

박찬빈[O], 하동호[*], 이호웅[**]
[O, *, **]호서대학교 컴퓨터공학과
coldbeen8954@gmail.com[O], hadoo6487@gmail.com[*], always14@hoseo.edu[**]

Frame-Level Latency Decomposition and Bottleneck Analysis of a WebGPU-WASM YOLO26 Object Detection Pipeline in Mobile Browsers

Chanbin Park[O], Dongho Ha[*], Howoong Lee[**]
[O, *, **]Dept. of Computer , Hoseo University

요 약

모바일 SoC와 GPU 성능 향상, WebGPU 지원 확대는 브라우저 기반 온디바이스 객체 탐지를 가능하게 하고 있다. 그러나 실시간 성능은 모델 추론뿐 아니라 카메라 입력, 전처리, 후처리, 렌더링을 포함한 전체 파이프라인 지연에 의해 결정된다. 본 연구는 동일한 YOLO26 파이프라인에서 ONNX Runtime Web의 WebGPU EP와 단일 스레드 WASM EP를 비교하고, Capture-Preprocess-Inference-Postprocess-Render 단계별 p95 지연을 프레임 단위로 분해하였다. Galaxy S25 Ultra와 iPhone 17에서 YOLO26n/s/m을 대상으로 실험했으며, warm-up 제외 후 WebGPU 29,100프레임, WASM 11,320프레임을 분석하였다. WebGPU 경로에서는 Inference p95가 E2E의 74.7~92.6%를 차지해 지배적 병목으로 나타났고, Postprocess와 Render p95는 각각 0.02ms, 1.4ms 이하로 부담이 작았다. 단일 스레드 WASM 기준선의 E2E p95는 WebGPU보다 3.2~15.3배 높았으며, iPhone 17에서는 긴 추론이 Capture로 기록한 프레임 콜백 스케줄링 지연에도 영향을 주었다. Galaxy WebGPU에서는 Preprocess가 iPhone 대비 3.5~4.4배 높아 2차 병목 후보로 확인되었다. p95 E2E 지연으로부터 환산한 처리율은 WebGPU YOLO26n 7.3~10.4fps, YOLO26m 3.1~3.9fps였다. 본 연구는 최적화 제안보다 파이프라인 지연을 재현 가능하게 계측·비교할 수 있는 분석 프레임과 WebGPU-WASM 정량 벤치마크를 제시한다.

키워드: WebGPU, WASM, YOLO26, Mobile Browser, Latency Decomposition

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행되었음"(2025-0-00040)

I. 서론

모바일 기기의 System on Chip(SoC) 및 Graphics Processing Unit(GPU) 성능 향상과 WebGPU 지원 확대는 웹 브라우저에서 YOLO 기반 온디바이스(on-device) 객체 탐지를 실행할 수 있는 환경을 만들고 있다[1]. 온디바이스 방식은 영상 프레임을 서버로 전송하지 않아 프라이버시 위험과 네트워크 의존성을 줄이고, 사용자 단말로 연산을 분산해 서버 측 GPU 및 트래픽 비용을 완화할 수 있다[2].

이러한 장점에도 불구하고 브라우저 기반 객체 탐지가 실시간 응용으로 이어지려면 모델 추론뿐 아니라 카메라 입력, 전처리, 후처리, 렌더링이 결합된 전체 지연이 허용 범위 안에 있어야 한다. 특히 End-to-End(E2E) 지연만으로는 어느 단계가 병목인지 구분하기 어렵고, 실행 경로, 기기, 모델 크기에 따라 지연 구조가 달라질 수 있다.

이에 본 연구는 YOLO26 객체 탐지 파이프라인을 Capture, Preprocess, Inference, Postprocess, Render 단계로 분해하고, ONNX Runtime Web의 WebGPU Execution Provider(EP)와 WebAssembly(WASM) EP를 동일 조건에서 비교한다.

II. 관련 연구

2.1 브라우저 온디바이스 추론 실행 경로

브라우저에서 딥러닝 모델을 실행하는 경로는 WebGL 기반 GPU 우회 구현, WASM 기반 CPU 추론, WebGPU 기반 GPU 가속 추론, 그리고 일부 환경에서 제공되는 WebNN 기반 네이티브 가속 경로 등으로 구분할 수 있으나 본 연구는 이 중 ONNX Runtime Web에서 비교 가능한 WebGPU EP와 WASM EP에 초점을 둔다. 초기 브라우저 환경에서는 WebGL이 대표적인 GPU 활용 수단이었으나, 그래픽 렌더링 중심 설계로 인해 추론 연산을 텍스처·셰이더로 우회해야 하는 한계가 있었다[3]. WebGPU는 compute shader 기반 GPGPU를 웹 표준화 API 형태로 제공하며, ONNX Runtime Web(ORT Web) 등의 런타임은 WebGPU EP를 통해 ONNX 그래프 연산을 GPU 커널로 실행한다[4]. 딥러닝 추론 관점에서 WebGPU의 장점은 연산을 GPU에 배치해 CPU 부하를 줄이고 병렬 처리를 활용할 수 있다는 점이다. 다만 실제 종단 지연에는 GPU 커널 실행뿐 아니라 전처리, CPU-GPU 간 데이터 전송, 커맨드 제출·동기화 비용이 함께 포함되며, 모바일 브라우저에서는 기기 성능, 메모리 대역폭, 발열에 따른 클럭 저하 등이 추가로 작용한다[5][6].

WASM EP는 ORT Web에서 제공하는 CPU 기반 실행 경로로, GPU를 사용하지 않고 브라우저의 CPU 실행 환경에서 모델 연산을 처리한다. 따라서 WebGPU EP에 비해 추론 지연이 크게 나타날 수 있지만, 주요 브라우저에서 안정적으로 동작한다는 장점이 있다. 본 연구에서는 WASM EP를 WebGPU 적용 효과를 비교하기 위한 CPU 기준선으로 사용하였다. 다만 모바일 환경에서는 SIMD 지원 여부, 스레드 사용 가능 여부, 메모리 제약 등에 따라 WASM EP의 성능이 달라질 수 있다[7].

2.2 YOLO26

YOLO(You Only Look Once)는 한 번의 신경망 순전파로 객체의 위치와 클래스를 동시에 예측하는 1-stage 객체 탐지 계열로, 2-stage 탐지기 대비 구조가 단순하고 지연이 낮아 실시간 영상 처리에 널리 사용된다. 초기 YOLO는 작은 객체 탐지와 위치 정밀도에서 한계가 있었으나, 이후 버전들은 다중 스케일 특징 융합, 개선된 라벨 할당과 손실 함수, 앵커 기반/앵커 프리 예측 전략을 도입해 정확도와 속도의 균형을 지속적으로 개선해 왔다.

Ultralytics 계열의 YOLOv8 이후 세대는 탐지 외에도 분할, 포즈, 분류, OBB를 통합 지원하는 방향으로 확장되었으며, 다양한 태스크 지원은 배포 목적에 따라 모델 선택과 후처리 구성을 더 신중하게 요구한다[8].

YOLO26은 이러한 흐름에서 엣지 및 저지연 배포를 강조한 모델로, n·s·m·l·x 변형을 제공해 정확도-지연 절충을 선택할 수 있게 한다. 주요 변화는 NMS-free end-to-end 구조를 통해 후처리 부담을 줄이고, DFL 제거로 export 및 하드웨어 호환성을 단순화하며, ProgLoss·STAL로 작은 객체 탐지 성능을 보완한 점이다[9]. 또한 학습 단계에서 MuSGD 기반 안정화 전략을 제시해 수렴 효율을 높이는 방향을 취한다[10].

III. 실험 환경 및 방법

모든 측정은 본 연구에서 구현한 동일한 웹 벤치마크 애플리케이션으로 수행하였으며, 실행 경로(WebGPU / WASM)와 모델·기기만을 변경하여 비교 가능한 조건을 유지하였다.

표1. 실험 단말 및 브라우저 구성

구분	iPhone 17	Galaxy S25 Ultra
OS	iOS 26.5	Android 16
브라우저	Chrome 148.0.7778.166	Samsung Internet 30.0.0.6
SoC	Apple A19	Qualcomm Snapdragon 8 Elite for Galaxy
RAM	8 GB	12 GB
메인 카메라	4,800만 화소	2억 화소

카메라 입력은 후면 광각 카메라로 획득하였고, 요청 해상도는 1280x720, 초당 처리 프레임 수는 30fps(frame per second)로 설정하였다. 실제 수신 해상도는 브라우저와 기기에 따라 달라질 수 있으나, 전처리에서는 640x640 letterbox 입력으로 통일하였다. iOS의 Chrome은 WebKit 기반이므로 iPhone 17 결과는 Chrome 앱과 iOS WebKit 구현의 영향을 함께 반영한다.

추론 런타임은 ONNX Runtime Web 1.26.0을 사용하였다. 실행 경로는 WebGPU EP와 WASM EP로 구분했으며, 각 실험 세션에서는 해당 경로만 활성화하였다. WASM은 모바일 메모리 및 스레딩 제약을 고려해 단일 스레드로 실행하였다. 모델은 YOLO26n, YOLO26s, YOLO26m의 ONNX 변환본(입력 640x640, FP32)을 사용했고, 객체 신뢰도 임계값은 0.25로 고정하였다. 카메라 접근과 런타임 실행 조건을 일관되게 유지하기 위해 HTTPS와 cross-origin isolation(COOP/COEP)을 적용한 환경에서 벤치마크를 수행하였다.

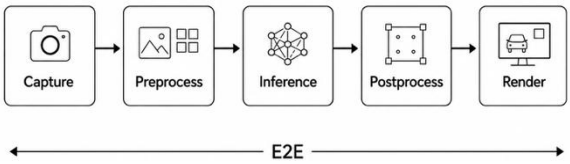


그림 1. 실험 파이프라인

Capture는 브라우저 비디오 프레임의 presentation timestamp부터 애플리케이션 처리 시작까지의 프레임 스케줄링 지연을 의미한다. Preprocess는 letterbox 리사이즈, 픽셀 읽기, NCHW 정규화를 포함하며, Inference는 ONNX 추론 완료까지의 시간으로 WebGPU 경로에서는 GPU 커널 실행, 데이터 전송, 동기화 비용을 포함한다. Postprocess는 출력 디코딩, confidence filtering, 좌표 역변환을 포함하고, Render는 Canvas 2D 오버레이 합성 완료까지의 시간으로 정의하였다. E2E는 프레임 presentation timestamp부터 Render 완료까지의 전체 지연이다.

실험은 이전 프레임의 추론이 완료된 뒤 다음 프레임을 처리하는 백프레셔 조건에서 수행하였다. 이 조건에서 E2E는 한 프레임을 끝까지 처리하는 데 걸린 시간이다. 본문과 표의 fps는 카메라

입력에 대한 초당 처리 프레임 수나 실제 디스플레이 갱신률이 아니라, p95 E2E 지연을 1000/E2E(ms)로 나누어 얻은 지연 환산 처리율이다. 단계별 지연은 브라우저 고해상도 타임스탬프 API로 기록하고, 각 세션의 처음 30프레임은 warm-up으로 제외하였다. 병목 판별의 1차 지표는 p95로 하였고 평균은 보조 지표로만 사용하였다. 단, 단계별 p95와 E2E p95는 각각 별도의 분포에서 계산되므로, 단계별 p95 값을 더한 결과가 E2E p95와 반드시 일치하지는 않는다. 또한 Capture는 센서 또는 ISP 수준의 카메라 획득 시간이 아니라, 비디오 프레임의 presentation timestamp 이후 애플리케이션 콜백이 실행되기까지의 브라우저 프레임 스케줄링 지연을 의미한다.

표2. warm-up 프레임 제외 전 데이터 수집량

실행경로	YOLO26n	YOLO26s	YOLO26m
WebGPU	1,000 × 5	1,000 × 5	1,000 × 5
WASM	1,000 × 5	300 × 2	300 × 1

WASM 경로는 모델 크기가 커질수록 세션 시간이 과도하게 길어지고 발열 영향을 통제하기 어려워 프레임 수집량을 줄였다. 또한 세션 종료 후 2~3분간 기기를 유휴 상태로 두어 thermal throttling 영향을 줄였다. warm-up 제외 후 분석 프레임 수는 WebGPU 29,100프레임, WASM 11,320프레임이다. 원시 측정 CSV와 집계 스크립트는 공개 저장소(https://github.com/ColdBeenInthePark/mobile-browser-yolo26-benchmark)에 포함하였다.

IV. 실험 결과

4.1 실행 경로별 E2E 지연 비교

표3. E2E p95(ms) 및 실효 처리율

모델	실행경로	Galaxy S25		iPhone 17	
		Ultra			
		E2E p95	fps	E2E p95	fps
yolo26n	WebGPU	137.3	7.3	96.1	10.4
	WASM	433.7	2.3	353.9	2.8
yolo26s	WebGPU	156.4	6.4	114.4	8.7
	WASM	1,339.2	0.7	1,147.0	0.9
yolo26m	WebGPU	326.8	3.1	254.0	3.9
	WASM	4,633.1	0.2	3,886.8	0.3

WebGPU 경로에서는 두 기기 모두 모델 크기가 커질수록 E2E p95가 증가하였다. p95 E2E 지연으로부터 환산한 처리율은 YOLO26n 7.3~10.4fps, YOLO26s 6.4~8.7fps, YOLO26m 3.1~3.9fps로 나타났다. WASM 경로의 E2E p95는 동일 조건의 WebGPU보다 3.2~15.3배 높았으며, 모델이 커질수록 격차가 확대되었다. 이는 단일 스레드 CPU 기반 WASM 기준선이 모델 연산량 증가에 더 민감함을 보여준다.

4.2 WebGPU 경로 단계별 지연

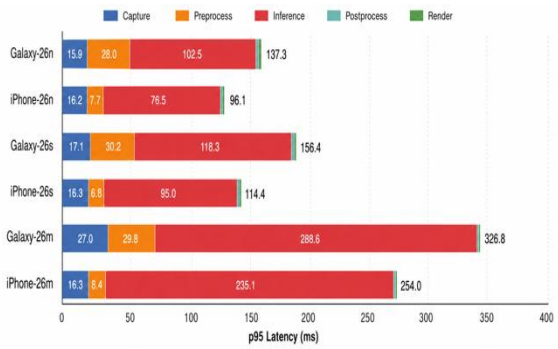


그림 2. WebGPU 경로 단계별 p95 지연

E2E 차이의 원인을 보기 위해 WebGPU 경로를 단계별로 분해하면, 모든 모델과 기기에서 Inference p95가 가장 컸다. 모델 크기가 커질수록 Inference p95와 E2E p95가 함께 증가하므로, WebGPU의 E2E 증가는 주로 GPU 가속 추론 구간에 의해 설명된다. Preprocess는 Galaxy S25 Ultra에서 28.0~30.2ms로 iPhone 17의 6.8~8.4ms보다 높게 나타나, 2D 캔버스 기반 리사이즈와 픽셀 버퍼 처리 비용이 플랫폼별로 다를 수 있음을 시사한다. 반면

Postprocess p95는 0.02ms 이하, Render p95는 1.4ms 이하로, NMS-free YOLO26 구조에서는 후처리와 오버레이 렌더링이 병목으로 나타나지 않았다.

4.3 WASM 경로 단계별 지연

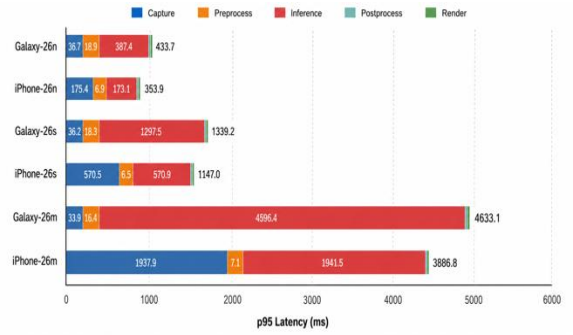


그림 3. WASM 경로 단계별 p95 지연

WASM 경로에서도 추론 지연은 크게 증가했지만, 그 지연이 파이프라인에 드러나는 방식은 기기별로 달랐다. Galaxy S25 Ultra에서는 Inference p95가 E2E 대부분을 차지해 CPU 추론이 지배적 병목이었다. 반면 iPhone 17에서는 Capture p95와 Inference p95가 유사했으며, YOLO26m 조건에서도 각각 1,937.9ms와 1,941.5ms로 거의 같았다. 이는 Capture가 순수 카메라 획득 시간이 아니라 긴 단일 스레드 WASM 추론과 백프레셔 조건이 반영된 브라우저 프레임 콜백 스케줄링 지연임을 보여준다. Preprocess, Postprocess, Render는 WASM에서도 전체 병목 구조에 미치는 영향이 제한적이었다.

V. 결론

본 연구는 모바일 브라우저 YOLO26 객체 탐지 파이프라인을 단계별로 분해하고, ONNX Runtime Web의 WebGPU EP와 단일 스레드 WASM EP를 동일한 백프레셔 조건에서 비교하였다. Galaxy S25 Ultra와 iPhone 17에서 YOLO26n/s/m을 측정한 결과, WebGPU 경로는 단일 스레드 WASM 기준선 대비 명확한 성능 이점을 보였지만 p95 지연 환산 처리율 기준으로 30fps 실시간 처리에는 도달하지 못했다.

전체 결과에서 핵심 병목은 추론 단계였다. WebGPU 경로에서는 모든 조건에서 Inference p95가 가장 컸고, 모델 크기 증가에 따라 E2E p95도 함께 증가하였다. Galaxy S25 Ultra에서는 Preprocess가 2차 병목 후보로 관찰되었으며, Postprocess와 Render는 두 실행 경로 모두에서 미미했다. WASM 경로는 모델 크기가 커질수록 WebGPU와의 격차가 확대되었고, iPhone 17에서는 긴 추론 시간이 Capture로 기록한 프레임 콜백 스케줄링 지연에도 반영되었다.

향후 연구에서는 GPU timestamp 기반의 추론 내부 분해와 다양한 모델 변형에 대한 확장 실험을 통해 추론 병목을 더 세밀하게 분석할 필요가 있다. 특히 FPS 향상을 위해서는 제1병목으로 확인된 추론 단계의 최적화가 핵심이며, 이를 위해 모델 경량화, 실행 경로 최적화, 데이터 전송 및 동기화 비용 최소화에 대한 후속 연구가 필요하다.

참고문헌

[1] Microsoft, “Using the WebGPU execution provider,” ONNX Runtime Documentation. Accessed: Jun. 5, 2026. [Online]. Available: https://onnxruntime.ai/docs/tutorials/web/ep-webgpu.html

[2] X. Wang, Z. Tang, J. Guo, T. Meng, C. Wang, T. Wang, and W. Jia, “Empowering edge intelligence: A comprehensive survey on on-device AI models,” ACM Computing Surveys, vol. 57, no. 9, pp. 1–39, 2025, doi: 10.1145/3724420.

[3] S. Sengupta, N. Wu, M. Varvello, K. Jana, S. Chen, and B. Han, “From WebGL to WebGPU: A reality check of browser-based GPU acceleration,” in

Proc. 2025 ACM Internet Measurement Conf. (IMC '25), Madison, WI, USA, Oct. 2025, pp. 1018–1024, doi: 10.1145/3730567.3764504.

[4] F. Jia, S. Jiang, T. Cao, W. Cui, T. Xia, X. Cao, Y. Li, Q. Wang, D. Zhang, J. Ren, Y. Liu, L. Qiu, and M. Yang, "Empowering in-browser deep learning inference on edge devices with just-in-time kernel optimizations," in Proc. 22nd Annu. Int. Conf. Mobile Syst., Appl. Services (MobiSys '24), Minato-ku, Tokyo, Japan, Jun. 2024, pp. 438–450, doi: 10.1145/3643832.3661892.

[5] D. Tian, H. Shen, and Y. Ma, "Poster: Parallelizing DNN inference in mobile web browsers on heterogeneous hardware," in Proc. 20th Annu. Int. Conf. Mobile Syst., Appl. Services (MobiSys '22), Portland, OR, USA, Jun.–Jul. 2022, pp. 519–520, doi: 10.1145/3498361.3538763.

[6] Q. Wang, S. Jiang, Z. Chen, X. Cao, Y. Li, A. Li, Y. Ma, T. Cao, and X. Liu, "Anatomizing deep learning inference in web browsers," ACM Trans. Softw. Eng. Methodol., vol. 34, no. 2, Art. no. 47, pp. 1–43, 2025, doi: 10.1145/3688843.

[7] A. P. Placitelli, S. Hochscheid, F. Maffucci, A. F. Rees, G. Treglia, and A. Staiano, "Towards decentralized ecological monitoring: A local-first web-based AI platform for sea turtle detection," Ecological Informatics, vol. 93, Art. no. 103569, 2026, doi: 10.1016/j.ecoinf.2025.103569.

[8] Ultralytics, "Ultralytics YOLO26," Ultralytics Docs. Accessed: Jun. 5, 2026. [Online]. Available: https://docs.ultralytics.com/models/yolo26/

[9] J. Chen and X. Ran, "Deep learning with edge computing: A review," Proceedings of the IEEE, vol. 107, no. 8, pp. 1655–1674, Aug. 2019, doi: 10.1109/JPROC.2019.2921977.

[10] G. Jocher, J. Qiu, M. Liu, S. Lyu, F. C. Akyon, and M. E. Kalfaoglu, "Ultralytics YOLO26: Unified real-time end-to-end vision models," arXiv:2606.03748, 2026, doi: 10.48550/arXiv.2606.03748.

개인정보보호 및 상호운용성 향상을 위한 데이터 메시 기반 의료데이터 공유 아키텍처 구현

김현겸[O], 조현태[*], 이호웅[**]
[O, *, **]호서대학교 컴퓨터공학과
rlagusrua7@gmail.com[O], whgusxo2055@gmail.com[*], always14@hoseo.edu[**]

Implementation of Privacy-Preserving Data Mesh-Based Healthcare Data Sharing Architecture

Hyeongyeom Kim[O], Hyeontae Cho[*], Howoon Lee[**]
[O, *, **] Dept. of Computer , Hoseo University

요 약

기존 의료 데이터 공유 방식은 중앙집중형 구조로 인해 데이터 소유권 문제, 운영 병목, 개인정보 노출 위험 등의 한계를 가진다. 또한 기관별로 상이한 EMR/EHR 시스템과 데이터 구조는 다기관 데이터 활용을 어렵게 만든다. 본 논문은 여러 선행연구에서 제안된 데이터 공유 및 상호운용성 아키텍처를 기반으로, 각 연구의 주요 개념을 통합·확장하여 개인정보 보호와 상호운용성을 동시에 지원하기 위한 데이터 메시(Data Mesh) 기반 의료 데이터 공유 아키텍처를 구현했다. 구현 아키텍처는 각 기관이 데이터의 소유권과 관리 책임을 유지하는 데이터 메시 원칙을 적용하며, 기관 내부에서는 로컬 CDM(Common Data Model)을 이용해 데이터를 표준화하고 기관 간 데이터 교환에는 FHIR(Fast Healthcare Interoperability Resources)를 활용한다. 데이터는 Raw 영역에 적재된 후 CDM 변환 과정을 거치며, De-ID 게이트를 통해 가명처리, 일반화, 토큰화 등의 비식별화가 수행된다. 이후 검증을 통과한 데이터만 데이터 제품(Data Product)으로 등록되어 카탈로그를 통해 공유된다. 또한 데이터 접근 시 요청자의 권한과 사용 목적에 따라 동적 마스킹을 적용하고, 감사 로그(Audit Log)와 데이터 라인리지(Lineage)를 통해 추적성과 책임성을 보장한다. 이를 통해 의료기관은 데이터 자율성을 유지하면서도 안전하고 확장 가능한 데이터 공유 환경을 구축할 수 있다. 본 연구는 데이터 메시, CDM, FHIR를 통합한 참조 아키텍처를 구현함으로써 의료 데이터 공유의 효율성과 신뢰성을 향상시킬 수 있는 방안을 제공한다.

키워드: 데이터 메시, CDM, FHIR

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행되었음"(2025-0-00040)

I. 서론

데이터 기반 의료 응용은 진단과 예후 모델의 성능을 높이며 의료 품질 향상에 기여하고 있으나, 보안과 프라이버시, 서비스 품질(QoS) 요구가 채택을 지속해서 제약한다. 중앙집중형 수집과 활용은 민감 정보의 외부 노출 위험과 운영상의 한계를 초래하고, 실시간 처리와 지연 최소화 요구는 분석을 분산·현장화하는 흐름을 강화한다[1]. 이러한 배경에서 기관 간 데이터 공유는 상호운용성과 프라이버시를 동시에 만족하는 체계적 접근이 필요하다.

본 논문은 여러 선행연구[2, 3]에서 제안된 데이터 공유 및 상호운용성 아키텍처를 기반으로, 각 연구의 주요 개념을 통합·확장하여 병원 단위의 자율성을 보전하고 프라이버시를 만족하면서 병원과 병원 간, 병원과 기업 간 상호운용성을 극대화하기 위해 데이터 메시 원칙 위에 내부 CDM 표준과 외부 FHIR 인터페이스를 결합한 아키텍처를 구현한다. 제안 시스템은 데이터 메시의 네 가지 원칙을 토대로, 병원 내부에서는 CDM으로 표준화하고, De-ID 게이트에서 비식별화를 수행하며, 셀프서브 플랫폼에서 이 CDM을 FHIR 리소스로 노출하여 병원과 병원, 병원과 기관 간 데이터 교환을 FHIR 기반 외부 인터페이스로 단순화한다. 각 기관은 하나의 도메인 팀으로 간주되며 자체 데이터 레이크를 운영한다. 원천 데이터는 로컬 CDM으로 정제·적재되고, 셀프서브 데이터 플랫폼이 비식별화된 데이터를 일관된 FHIR 리소스로 노출한다. 다른 기관은 FHIR API로 필요한 데이터를 안전하게 조회하거나, 수신한 FHIR을 자 기관의 로컬 CDM으로 가공해 자체 레이크에 적재한다. 플랫폼은 ETL 배치, 메쉬/데이터 카탈로그, CI/CD 게이트를 제공하고, 연합 컴퓨테이셔널 거버넌스는 표준·보안·변경 관리를 Policy as Code로 일관되게 집행한다. 이로써 기관별 내부 표준을 유지하되 외부 교환면에서는 단일 계약을 따르게 하여 상호운용성과 확장성을 동시에 달성한다.

II. 관련 연구

2.1 CDM

다기관 공동연구에는 비용과 시간의 부담이 크고 개인정보 공유의 법적 제약도 많다. 이를 완화하기 위해 각국 의료기관은 CDM을 도입하여 데이터 구조와 어휘를 표준화하고 분산연구 네트워크에서 분석 결과만 공유하는 방식을 확산하고 있다[4, 5]. Sentinel, PCORnet, OMOP이 대표적이며 특히 OMOP CDM은 기관별로 다른 EMR(Electronic Medical Record)/EHR(Electronic Health Record) 데이터를 표준 어휘와 열네 개 임상 테이블로 정규화하여 종단 연구와 재현 가능한 분석을 지원하는 모델로 평가된다[4]. CDM을 활용하면 의료기관 데이터를 표준화된 스키마와 용어 체계를 갖춘 공통된 데이터베이스 모델로 변환하고, 재사용 가능한 매개변수화 질의 도구를 사용해 의료기관 밖으로 어떠한 개인정보도 노출하지 않으면서 동시 다기관 분석을 수행할 수 있다[5, 9].

2.2 FHIR

FHIR는 HL7 v2의 단순함과 v3와 CDA(Clinical Document Architecture)의 의미 일관성을 잇고 웹 표준을 채택하여 구현이 쉽고 상호운용성이 높은 차세대 교환 표준이다[10, 12]. 환자와 관찰 같은 모듈식 리소스를 조합해 모델을 구성하며 프로파일과 확장을 통해 현장의 맥락을 반영하되 상호운용성을 유지하도록 설계되었다[10, 11]. 사양은 자유롭게 활용할 수 있고 구현 라이브러리와 예제가 풍부하여 인터페이스를 빠르게 구축하기 쉽다[10].

연구 영역에서 FHIR는 데이터 수집, 표준화, 모집, 동의 관리 등 다양한 단계에 적용되고 OMOP같은 다른 리소스 내용의 변동 가능성, 보안 및 법적 이슈와 FHIR 서버 운영 필요성같은 한계도 존재한다[9]. FHIR는 임상과 연구를 잇는 인터페이스를

단순화하고 CDM 및 표준 용어와 결합하여 연구 데이터의 상호운용성과 재사용성을 높이는 기반으로 기능한다[11].

2.3 데이터 메시

데이터 메시는 기존의 중앙 집중형 데이터 레이크와 데이터 웨어하우스에서 나타나는 확장성과 민첩성 저하 그리고 중앙팀 병목을 해소하기 위해 제안된 도메인 지향 분산 아키텍처이다. 데이터의 제품화, 도메인 소유권, 연합 컴퓨테이셔널 거버넌스, 셀프서브 데이터 플랫폼을 핵심 원칙으로 한다[13, 14].

데이터 제품은 데이터·메타데이터, 수집·처리· 제공·거버넌스 코드, 그리고 이를 위한 인프라를 포함하는 최소 단위로 정의된다. 제품은 발견 가능성, 주소 가능성, 신뢰성, 자체 기술성, 상호운용성, 보안을 갖추어야 한다[13, 14]. 도메인 팀은 자신들의 데이터를 가장 잘 아는 주체이므로 제품의 구축과 운영, 규정 준수를 책임진다. 중앙의 연합 컴퓨테이셔널 거버넌스는 제품 간 상호운용 표준, 품질 정의 방법론, 공통 용어집을 마련하고 정책을 코드로 집행하여 확장성을 확보한다[13]. 셀프서브 데이터 플랫폼은 다양한 저장소, 파이프라인 오케스트레이션, 메타데이터 카탈로그, 연합 질의, 모니터링, 보안, 접근 관리, 정책 집행 같은 공통 역량을 제공하여 도메인별 중복 노력을 줄이고 수명주기 관리를 지원한다[13, 14].

데이터 메시는 소스와 소비자가 늘어날 때의 확장성 향상, 중앙 병목 제거에 따른 민첩성 제고, 원천에 가까운 책임 배분으로 인한 품질 향상을 기대할 수 있다. 또한 발견성과 거버넌스가 좋아지고 데이터 리드타임이 짧아지는 장점도 있다. 반면 조직 변화관리의 어려움, 데이터 복제 및 관리 비용 증가, 기술 부채, 초기 복잡성 같은 과제가 존재한다[13]. 대규모이면서 데이터가 다양한 조직이나 민첩성과 거버넌스 요구가 높은 조직에 특히 적합하며 도메인 지향 조직 구조와 성숙한 거버넌스, 플랫폼 역량이 정착의 전제 조건으로 제시된다[13, 14].

III. 시스템 개요

본 연구는 병원 단위의 자율성을 보전하고 프라이버시를 만족하면서 병원과 병원 간, 병원과 기업 간 상호운용성을 극대화하기 위해, 데이터 메시의 네 가지 핵심 원칙 위에 CDM 내부 표준과 FHIR 외부 인터페이스를 결합한 아키텍처를 구현한다. 각 기관은 하나의 도메인 팀이며 자체 데이터 레이크를 보유한다. 병원 내부에서는 원천 데이터를 정제하여 로컬 CDM으로 표준화하고, De-ID 게이트가 이를 비식별화하며, 셀프서브 데이터 플랫폼은 이 CDM을 FHIR 리소스로 일관되게 노출한다. 다른 병원 또는 기업은 FHIR API를 통해 필요한 데이터를 안전하게 조회하거나, 수신한 FHIR를 자 기관의 로컬 CDM으로 가공해 자체 데이터 레이크에 적재한다. 플랫폼은 ETL 배치 처리와 메시/데이터 카탈로그, CI/CD 게이트를 제공하고, 연합 컴퓨테이셔널 거버넌스는 메시 표준과 보안 정책, 변경 관리를 Policy-as-Code로 집행한다. 외부로 노출되는 모든 데이터는 정책 기반 비식별화 계층을 반드시 통과하며, 업로드 시 정적 비식별화, FHIR API·Bulk Export 시 동적 마스킹을 적용한다. 연구용, 기업용 등의 비식별화 프로파일은 연합 컴퓨테이셔널 거버넌스가 Policy-as-Code로 관리한다. 본 시스템은 운영 절차를 업로드와 다운로드 두 흐름으로 정형화하여, 동일한 정책·검증·인터페이스로 처리되도록 설계한다.

IV. 시스템 구조

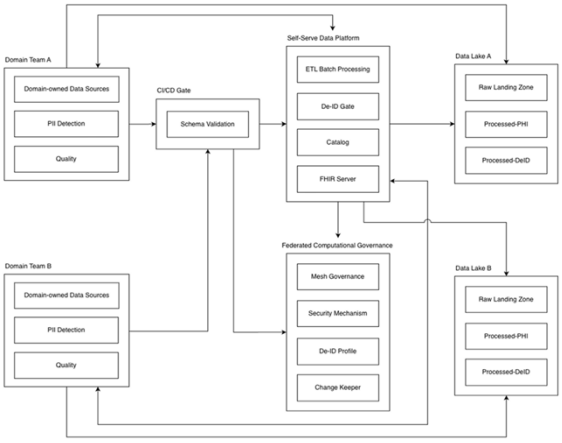


그림 1. 시스템 구성도

도메인 팀 A는 원천 데이터를 데이터 레이크 A의 Raw 영역에 적재한 뒤 PII 탐지를 수행하고, 내부 ETL로 로컬 CDM 스키마에 맞춰 변환하여 Processed 영역에 저장한다. 변환 산출물은 CI/CD 게이트에서 스키마 호환성, 데이터 계약 준수, 프라이버시 규정 준수를 자동 점검받고, 통과하면 메시/데이터 카탈로그에 제품 메타데이터가 등록된다. 셀프서브 데이터 플랫폼은 등록된 CDM 데이터를 FHIR 리소스와 엔드포인트로 노출하며, FHIR 서버가 FHIR Subscription을 지원하는 경우 변경 알림 구독을 설정하고, 미지원인 경우 Bulk Export와 증분 파라미터로 동기화를 수행한다. 도메인 팀 B는 카탈로그에서 해당 제품을 검색해 접근 승인을 받고, FHIR API로 직접 조회하거나 수신한 FHIR를 도메인 팀 B의 CDM 스키마에 매핑해 데이터 레이크 B로 적재한다. 전 구간에서 연합 컴퓨테이셔널 거버넌스가 표준과 보안을 강제하며, 인증·권한·암호화·감사 로깅이 공통 보안 메커니즘으로 적용되고, Change Keeper가 버전과 데이터 구조 변화를 기록해 호환성 위험을 알린다. 데이터 레이크 Processed Zone은 Processed-PHI와 Processed-DeID로 분리한다. 업로드 파이프라인에는 De-ID 게이트를 포함시켜 토큰화/가명처리, 날짜 시프트, 일반화/버킷화, 자유텍스트 PII 레드액션 등을 정적으로 적용하고, 결과를 Processed-DeID에 적재한다. FHIR 서버는 요청자의 권한과 데이터 사용 목적에 따라 동적 마스킹 규칙을 적용한다.

V. 구현 결과

표 1. 구현 환경

항목	소프트웨어
서버	Spring Boot 3.2.4
데이터베이스	PostgresSQL 15
빌드 도구	Gradle 8.5
컨테이너	Docker 29.5.2
웹 인터페이스	HTML, CSS, JavaScript

구현은 표 1에 나열된 소프트웨어, 프레임워크 등을 사용하여 구현하였고, 셀프서브 데이터 플랫폼과 각 기관별 데이터레이크에 해당하는 컨테이너들을 생성하여 격리 실행하였다.

구현 결과, PostgresSQL 데이터베이스에는 raw_record와 deid_record 두 개의 주요 테이블이 생성되었다. raw_record에는 원천 EMR 데이터가 저장되었고, deid_record에는 비식별화된 데이터가 저장되었다. 이를 통해 원천 데이터와 공유용 데이터를 분리 관리할 수 있음을 확인하였다.

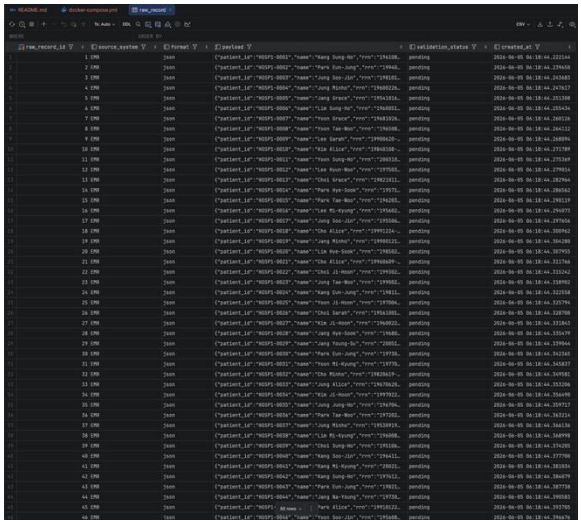


그림 2. raw_record 테이블

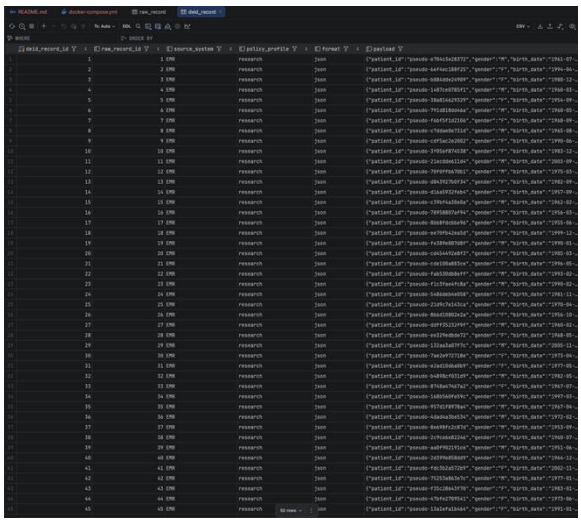


그림 3. deid_record 테이블

웹 인터페이스에서는 데이터 제품 등록, 카탈로그 조회, 제품 상세 확인, 접근 요청, 소유자 승인, 승인 후 다운로드까지의 기능이 정상적으로 동작하였다. 특히 hospital1 사용자는 데이터 제품을 등록하고 승인 권한을 수행하며, hospital2 사용자는 카탈로그를 통해 제품을 발견하고 접근 요청을 수행하는 방식으로 도메인 간 데이터 공유 시나리오를 구현하였다.

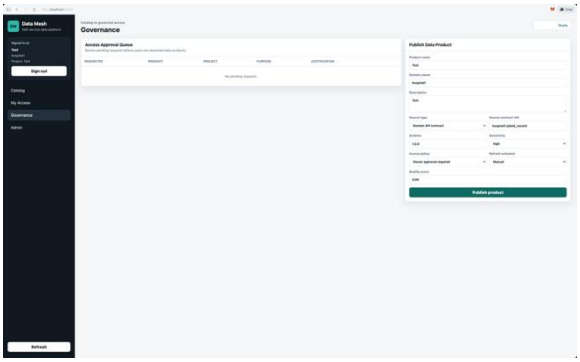


그림 4. 데이터 제품 등록

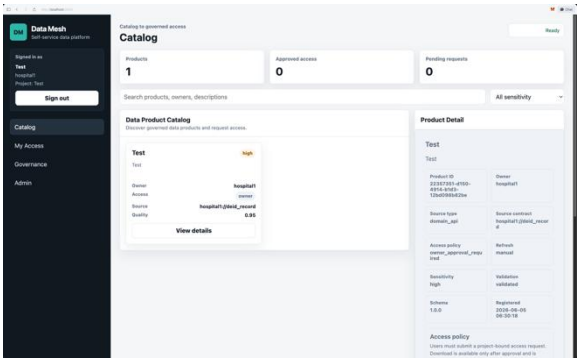


그림 5. 카탈로그 조회

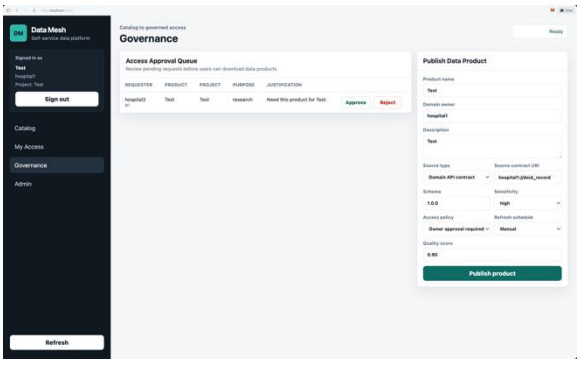


그림 6. 접근 요청 및 소유자 승인

VI. 결론

본 논문은 여러 선행 연구에서 제안된 데이터 공유 및 상호운용성 아키텍처를 기반으로, 각 연구의 주요 개념을 통합·확장하여 개인정보보호 및 상호운용성 향상을 위한 데이터 메시 기반 의료 데이터 공유 아키텍처를 실제 프로토타입으로 구현하였다. 구현 시스템은 원천 데이터 저장,

비식별 데이터 생성, 데이터 제품 등록, 카탈로그 기반 탐색, 접근 요청, 소유자 승인, 승인 후 다운로드의 전 과정을 지원한다. 이를 통해 의료 데이터 공유 환경에서 데이터 소유권을 기관에 유지하면서도, 외부 소비자가 통제된 절차를 통해 데이터를 활용할 수 있음을 확인하였다.

향후 연구에서는 FHIR API 연동, 세분화된 행·열 단위 접근 제어, 동적 마스킹, 감사 로그 시각화, 대용량 Bulk Export 성능 평가를 추가하여 시스템의 실용성과 확장성을 검증할 예정이다.

참고문헌

[1] Rauniyar, A., Hagos, D. H., Jha, D., Håkegård, J. E., Bagci, U., Rawat, D. B., & Vlassov, V. (2023). Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. IEEE Internet of Things Journal, 11(5), 7374-7398.

[2] Machado, I., Costa, C., & Santos, M. Y. (2021). Data-driven information systems: the data mesh paradigm shift.

[3] Kim, H., Han, M., & Lee, H. (2025). Privacy-Preserving Data Mesh-Based Healthcare Data Sharing Architecture. 한국차세대컴퓨팅학회 학술대회, 330-333.

[2] 이승민 (2024). Research on Data Modeling and CDM-based Analysis for Effective Primary and Secondary Use of Health Data [Doctoral Dissertation, 서울대학교]. https://www.riss.kr/link?id=T17109854.

[3] 부유경, 송시현, 박지환, & 노미정 (2024). 단국대학교병원 EHR 기반 MOA CDM 구축을 위한 용어 매핑 사례와 시사점. Korea journal of hospital management, 29(1), 1-18.

[4] Chae-Eun Jeong, Yunhee Kang, & Young B. Park (2020). Automation System for Sharing CDM Data. JOURNAL OF PLATFORM TECHNOLOGY, 8(3), 3-9.

[5] Lenert, L. A., Ilatovskiy, A. V., Agnew, J., Rudisill, P., Jacobs, J., Weatherston, D., & Deans, K. R., Jr (2021). Automated production of research data marts from a canonical fast healthcare interoperability resource data repository: applications to COVID-19 research. _Journal of the American Medical Informatics Association : JAMIA_, _28_(8), 1605-1611. https://doi.org/10.1093/jamia/ocab108.

[6] Choi, W., Yang, Y. S., Chang, D. J., Chung, Y. W., Kim, H., Ko, S. J., Yoo, S., Oh, J. S., Kang, D. Y., Yang, H. J., & Choi, I. Y. (2021). Association between the use of allopurinol and risk of increased thyroid-stimulating hormone level. _Scientific reports_, _11_(1), 20305. https://doi.org/10.1038/s41598-021-98954-1.

[7] Son, N., Kim, B., Chung, S., & Han, S. (2019). Korean Pharmacovigilance System Based on EHR-CDM. _Studies in health technology and informatics_, _264_, 1592-1593. https://doi.org/10.3233/SHTI190550.

[8] FHIR foundation. (n.d.). Introducing HL7 FHIR. https://build.fhir.org/fhir-glossy.pdf.

[9] Vorisek, C. N., Lehne, M., Klopfenstein, S. A. I., Mayer, P. J., Bartschke, A., Haese, T., & Thun, S. (2022). Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review. _JMIR medical informatics_, _10_(7), e35724. https://doi.org/10.2196/35724.

[10] Bender, D., & Sartipi, K. (2013). HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. _Proceedings of the 26th IEEE International Symposium on Computer-

Based Medical Systems_, 326-331.

[11] Goedegebuure, A., Kumara, I., Driessen, S., Van Den Heuvel, W. J., Monsieur, G., Tamburri, D. A., & Nucci, D. D. (2024). Data mesh: a systematic gray literature review. ACM Computing Surveys, 57(1), 1-36.

[12] Machado, I. A., Costa, C., & Santos, M. Y. (2022). Data mesh: concepts and principles of a paradigm shift in data architectures. Procedia Computer Science, 196, 263-271.

이중 시점 융합 기반의 저조도 객체 분류

Muhammad Uzair*(발표자), 김수균O
*, O제주대학교 컴퓨터공학과
e-mail: uzairmuhammad@office.jejunu.ac.kr*, kimsk@jejunu.ac.krO

Low-Light Object Classification via Dual-View Fusion

Muhammad Uzair*, Soo-kyun KimO
*, ODept. of Computer Engineering, Jeju National University

요 약

저조도 환경은 질감, 경계, 색 정보를 약화시켜, 객체 분류를 어렵게 한다. 본 논문은 저조도 객체 분류를 위한 검증 기반 이중 관점 구조 모델을 연구한다. 각 이미지에 대해 원본 저조도 입력과 Zero-DCE로 향상된 이미지를 두 가지 관점으로 인식하고, 동일한 계층화된 학습, 검증, 테스트 분할을 사용하여 각 관점에 대해 ConvNeXt-Tiny 분류기를 독립적으로 학습한다. 추론 단계에서는 두 클래스 확률 벡터를 결합하며, 융합된 가중치는 검증 단계에서 활용한다. ExDark 데이터세트에서 제안하는 방식은 83.89%의 정확도와 83.78%의 F1-Score를 달성하였으며, 이는 Zero-DCE 단일 관점 ConvNext-Tiny 기준 모델의 83.62% 정확도 및 83.45%를 개선한 결과로 볼 수 있다.

키워드: 저조도 이미지, 객체 분류, Zero-DCE, ExDark, ConvNext-Tiny

I. 서론

저조도 환경에서의 객체 분류는 가시성과 판별적 구조의 영향을 받으며, 객체가 이미지 내에 존재하더라도 분류기는 덜 신뢰할 수 있는 증거를 입력받게 된다. 이를 해결하기 위해 이미지 품질 개선 방법이 연구되며, 원본 이미지의 정보 훼손을 최소화하는 것이 중요하다. 이러한 관찰을 바탕으로 원본 관점과 Zero-DCE 관점을 동시에 융합하여 저조도 외형을 보존하면서 어두운 구조가 더 잘 드러나는 조명 보정 표현을 사용한다. 최종적인 픽셀 결정은 확률 수준에서 수행되며, 각 관점에 대해 ConvNext-Tiny로 학습하고 추론 시 softmax 방식을 통해 두 관점의 출력을 결합한다. 범용성있는 결과를 위해서 가중치를 고정하고 테스트 세트에서 최종 결정 규칙 조정을 방지한다.

II. 관련 연구

2.1 저조도 개선 연구

Zero-DCE는 쌍을 이루는 정상 조도 없이 이미지별 향상 곡선을 추정한다. Zero-DCE는 분류 손실을 통해 최적화되지 않으며, 조명 보정 관점이 저조도 품질 개선에 기여하는지 평가할 수 있다.

2.2 인식 모델

ResNet18 모델은 경량형 분류 모델로 사용되며,

본 연구는 교육부와 제주특별자치도의 지원을 받아 제주 RISE 센터를 통 해 Regional Innovation System and Education (RISE) 사업의 지원을 받아 수행된 연구임. (2026 RISE_2-2-11)

ConvNeXt-Tiny는 주요 백본으로 사용된다. 두 모델은 이미지 향상과 관점 예측을 융합하기 위해 ImageNet으로 초기화하고 ExDark에서 평가한다.

III. 제안방법

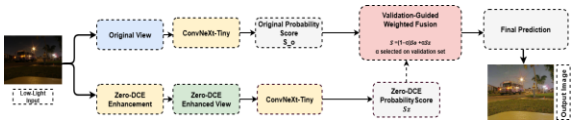


그림 1. 이중 관점 예측 융합 프레임워크

그림 1은 제안하는 이중 관점 예측 융합 프레임워크를 보여준다. 하나의 입력 이미지를 통하여 원본 저조도 이미지와 Zero-DCE 향상 이미지를 사용한다. 이를 통해 공통된 최종 결정을 위한 분류기의 확률 점수를 활용한다. 과적합을 줄일 수 있도록 크기 조정, 뒤집기, 색조 변형 등의 증강 방식을 적용한다. 원본 관점 분류기의 예측 확률 벡터 S_o 와 Zero-DCE 관점 분류기의 예측 확률 벡터 S_z를 이용하여 융합 점수를 다음과 같이 계산한다.

$$S = (1-\alpha)S_o + \alpha S_z$$

α 값으로 향상 관점의 기여도를 조절하며, 0부터 1까지 0.05 간격으로 탐색한다. 가중치 융합은 확률 추정 이후에 적용되며, 두 관점 분류기는 독립적으로 학습된다. 최종 결정 규칙은 특정 수준의 재학습을 필요로 하지 않으며, 단일 ConvNeXt-Tiny와 이중 융합 방식의 최종 확률 점수를 비교한다.

IV. 결과

사용하는 데이터세트는 12개의 객체 클래스로 구성된 7,363 장의 저조도 이미지로 구성된다. 전체의 15% 정도를 테스트 데이터로 활용하였으며, 평가 지표로는 정확도, 재현율, F1-Score를 사용한다. 표 1에서는 백본별 결과를 비교하며, 제안하는 모델이 F1-Score 기준으로 83.78%를 달성하여, 다른 방식에 비해 더 효과적임을 보인다.

표 1. ExDark 데이터세트에서의 실험 성능

Method	Accuracy (%)	Recall (%)	F1-score (%)
ResNet18 + Original	70.95	70.95	70.90
ResNet18 + Zero-DCE	73.39	73.39	73.52
ConvNeXt-Tiny + Original	82.35	82.35	82.25
ConvNeXt-Tiny + Zero-DCE	83.62	83.62	83.45
Proposed Dual-View Fusion	**83.89 ↑**	**83.89 ↑**	**83.78 ↑**

IV. 결론

데이터사이언스는~~ (신명조, 9) 본 논문은 저조도 객체 분류를 위한 검증 기반 이중 관점 융합 방법을 제안하며, 원본과 향상된 이미지를 상호 보완적인 인식 관점으로 활용한다. 또한 ExDark 데이터세트에서 83.89%의 정확도와 83.78%의 F1-Scroe를 달성하여 기존 방식보다 우수함을 보인다. 향후에는 동일한 융합 설계를 다른 데이터세트에 적용하여 범용성을 강화할 예정이다.

참고문헌

[1] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the Exclusively Dark dataset," Computer Vision and Image Understanding, Vol. 178, pp. 30-42, 2019.

[2] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," Proc. IEEE/CVF CVPR, pp. 1780-1789, 2020.

[3] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," Proc. IEEE/CVF CVPR, pp. 11976-11986, 2022.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE CVPR, pp. 770-778, 2016.

본 연구는 교육부와 제주특별자치도의 지원을 받아 제주 RISE 센터를 통 해 Regional Innovation System and Education (RISE) 사업의 지원을 받아 수행된 연구임. (2026 RISE_2-2-11)

발달지연 유아 운동중재 연구의 지식 진화 분석: 키워드 네트워크 기반 연구동향 탐색

김아름[O], 김미량[*]
[O]순천향대학교 휴먼헬스케어융합연구소
*순천향대학교 스포츠응용산업학과

e-mail:{김아름}kar6794@sch.ac.kr[O], {김미량}mlkim@sch.ac.kr[*]

Knowledge Evolution in Exercise Intervention Research for Young Children with Developmental Delay: A Keyword Network-Based Trend Analysis

Areum Kim[O] , Miryang Kim[*]
[O]Human Healthcare Convergence Institute, Soonchunhyang University,
[*]Dept. of Sport Industry Studies, Soonchunhyang University

요 약

기존 연구는 개별 중재 효과 검증에 초점을 두고 있어 연구주제와 성과영역의 변화 과정을 종합적으로 분석한 연구는 부족하였다. 이에 따라 본 연구는 발달지연 유아 대상 운동중재 연구의 지식구조 변화와 연구동향을 탐색하기 위하여 수행되었다. 국내 학술 데이터베이스(RISS, DBpia)를 통해 수집된 최종 64편의 연구를 대상으로 중재 및 성과 관련 키워드를 추출하고 정제하였다. 이후 연구 시기를 2000~2005년, 2006~2010년, 2011~2015년, 2016~2020년 및 2021~2025년으로 구분하여 시기별 핵심 키워드 변화와 지식 진화 과정을 분석하였다..

분석 결과, 초기 연구에서는 놀이와 운동 중심의 중재가 수행되었으며 인지, 행동, 정서 및 언어와 같은 기초 발달영역이 주요 성과로 나타났다. 2011~2015년에는 운동 및 심리운동 기반 연구가 증가하면서 행동, 인지, 언어뿐 아니라 균형, 협응 및 운동발달과 같은 신체적 성과가 확대되었다. 2016~2020년에는 감각통합, 신체활동 및 ICT 기반 중재가 등장하면서 중재유형이 다양화되었으며, 인지, 정서, 문제해결 및 적응행동과 같은 다차원적 발달성과가 주요 연구주제로 나타났다. 최근 연구에서는 놀이와 신체활동 기반 중재가 지속적으로 수행되는 가운데 인지, 언어 및 주의집중 관련 연구가 증가하는 경향을 보였다.

이러한 결과는 발달지연 유아 운동중재 연구가 초기의 놀이 중심 접근에서 운동 및 심리운동 중심 연구를 거쳐 감각통합, ICT 및 통합형 프로그램 기반 연구로 발전하였음을 보여준다. 또한 연구성과는 신체발달 중심에서 인지, 정서 및 적응을 포함하는 다차원적 발달 지원 체계로 확장되고 있는 것으로 나타났다.

※ 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임
(NRF-2025S1A5C2A02022632)

키워드: 발달지연, 운동중재, 지식진화, 키워드 네트워크 분석, 연구동향

I. 서론

발달지연 유아를 대상으로 한 운동중재 연구는 신체기능 향상뿐 아니라 인지 및 행동 발달 지원을 목적으로 지속적으로 확대되고 있다. 그러나 대부분의 선행연구는 개별 중재의 효과 검증에 초점을 두고 있어 연구주제와 핵심 개념이 시간에 따라 어떻게 변화해 왔는지에 대한 분석은 부족한 실정이다. 최근 데이터사이언스 기반 문헌분석 기법은 특정 연구영역의 지식구조와 연구동향을 체계적으로 파악하는 방법으로 활용되고 있다[1,2]. 따라서 본 연구는 발달지연 유아 운동중재 연구의 지식 진화 과정을 키워드 네트워크 관점에서 분석하고자 하였다.

II. 연구방법

2.1 자료수집 및 문헌 선정

본 연구는 국내 학술 데이터베이스인 RISS와 DBpia를 활용하여 발달지연 유아 운동중재 관련 연구를 수집하였다. 검색어는 '발달지연', '발달부진', '발달지체' 및 '유아'를 중심으로 구성하였으며, 운동중재와 관련된 국내 학술논문 및 학위논문을 대상으로 검색을 수행하였다. 문헌 선정은 PRISMA 2020 가이드라인을 참고하여 수행하였으며[3], 중복 문헌 제거 및 연구목적 적합성 검토를 거쳐 최종 64편의 연구를 선정하였다.

또한 연구동향 변화를 파악하기 위하여 연구 시기를 2000~2005년, 2006~2010년, 2011~2015년, 2016~2020년 및 2021~2025년으로 구분하여 분석하였다.

2.2 키워드 추출 및 분석

최종 선정된 64편의 문헌을 대상으로 중재유형과 발달성과 관련 키워드를 추출하였다. 이후 동일 의미를 갖는 키워드에 대해 통합 및 정제 과정을 수행하여 분석용 키워드 목록을 구축하였다.

키워드 네트워크 분석은 공출현(co-word) 분석과 사회연결망분석 이론에 기반하여 수행하였다[1,2,4,5]. 이를 통해 키워드 간 연결 특성과 연구주제의 변화 양상을 확인하였으며, 시기별 중재 키워드 및 성과 키워드의 출현 빈도를 분석하여 발달지연 유아 운동중재 연구의 지식 진화 과정을 탐색하였다.

III. 연구결과

3.1 발달지연 유아 운동중재 연구 시기별 핵심 중재 키워드 변화

표1. 시기별 핵심 중재 키워드 변화

연구 시기	연구 수	주요 중재 키워드
2000~2005	2	놀이, 운동
2006~2010	13	놀이, 프로그램, 운동, 감각통합, 훈련
2011~2015	15	운동, 심리운동, 놀이, 프로그램, 치료
2016~2020	27	프로그램, 운동 , 치료, 놀이, 감각통합, 신체활동, ICT
2021~2025	7	프로그램, 놀이, 신체활동, 운동

표 1에 제시된 바와 같이 발달지연 유아 운동중재 연구의 중재 키워드는 시기에 따라 변화하는 양상을 보였다. 초기(2000~2005)에는 놀이 중심 접근이 주를 이루었으며, 2006~2010년에는 감각통합 및 훈련 기반 중재가 등장하기 시작하였다. 이후 2011~2015년에는 운동과 심리운동 관련 연구가 크게 증가하면서 신체활동 기반 연구가 확대되었다. 2016~2020년에는 프로그램, 감각통합,

※ 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임
(NRF-2025S1A5C2A02022632)

신체활동 및 ICT 기반 중재가 등장하면서 중재유형이 다양화되었으며, 최근(2021~2025)에는 놀이와 신체활동을 기반으로 한 통합형 프로그램 연구가 지속적으로 수행되는 것으로 나타났다.

3.2 발달지연 유아 운동중재 연구 시기별 핵심 성과 키워드 변화

표2. 시기별 핵심 성과 키워드 변화

연구 시기	연구 수	주요 성과 키워드
2000~2005	2	인지, 행동, 정서, 언어, 사회적 상호작용
2006~2010	13	인지, 행동, 주의, 언어, 적응
2011~2015	15	행동, 인지, 언어, 주의, 균형, 협응, 운동발달
2016~2020	27	행동, 인지, 정서, 언어, 문제해결, 적응
2021~2025	7	인지, 언어, 정서, 주의

발달지연 유아 운동중재 연구의 성과 키워드 변화는 표 2에 제시하였다. 초기 단계인 2000~2005년에는 인지, 행동, 정서, 언어 및 사회적 상호작용과 같은 기초 발달영역이 주요 성과로 나타났다. 당시 연구는 운동중재가 유아의 전반적인 발달에 미치는 영향을 탐색하는 수준에서 수행된 것으로 확인되었다.

2006~2010년에는 인지와 행동이 핵심 성과로 지속적으로 나타났으며, 주의집중과 적응행동 관련 연구가 증가하기 시작하였다. 이는 발달지연 유아의 신체기능 향상뿐만 아니라 인지 및 행동조절 능력에 대한 관심이 확대되었음을 보여준다.

2011~2015년에는 행동, 인지, 언어가 주요 성과로 나타났으며, 균형, 협응, 운동발달과 같은 신체적 성과가 함께 증가하였다. 특히 운동 및 심리운동 중재 연구의 확대와 함께 신체기능 향상과 행동변화 간의 관계를 탐색하는 연구가 활발하게 수행된 것으로 나타났다.

2016~2020년에는 인지와 행동이 가장 높은 빈도로 나타났으며, 정서, 문제해결, 적응행동 관련 성과가 확대되는 경향을 보였다. 이는 운동중재 연구가 단순한 신체기능 향상을 넘어 인지조절, 사회적 적응 및 정서발달을 포함하는 다차원적 발달 성과에 관심을 갖기 시작했음을 시사한다.

최근 시기인 2021~2025년에는 인지가 가장 높은 빈도로 나타났으며, 언어, 정서, 주의집중 관련 연구가 지속적으로 수행되는 것으로 확인되었다. 특히 인지와 주의집중을 중심으로 한 연구가 증가하면서 발달지연 유아 운동중재 연구의 관심이 신체영역에서 인지·행동 영역으로 확장되고 있는 경향이 나타났다.

전반적으로 성과 키워드는 초기의 기초 발달 및 행동 중심 연구에서 인지, 적응, 문제해결, 정서조절을 포함하는 다차원적 발달성과 중심 연구로 진화한 것으로 나타났다. 이러한 변화는 발달지연 유아 운동중재 연구가 신체적 발달뿐 아니라 인지적·심리사회적 발달을 함께 고려하는 통합적 접근으로 발전하고 있음을 보여준다.

3.3 발달지연 유아 운동중재 연구의 지식 진화 과정

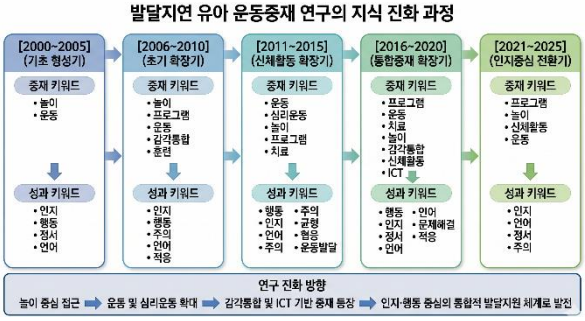


그림 1. 시기별 지식 진화 과정(2000~2025)

그림 1은 발달지연 유아 운동중재 연구의 지식 진화 과정을 시기별로 나타낸 것이다. 초기에는 놀이 중심 접근이 주를 이루었으나 이후 운동 및 심리운동 중심 연구가 확대되었으며, 최근에는 감각통합, ICT 및 통합형 프로그램을 포함하는 다차원적 중재체계로 발전하는 경향이 나타났다.

IV. 결론

본 연구는 발달지연 유아를 대상으로 한 운동중재 연구의 지식 진화 과정을 탐색하기 위하여 국내 연구 64편을 대상으로 키워드 네트워크 기반

※ 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2025S1A5C2A02022632)

연구동향 분석을 수행하였다. 분석 결과, 발달지연 유아 운동중재 연구는 시기에 따라 연구주제와 성과 영역이 점진적으로 확장되는 경향을 보였다.

중재 측면에서는 초기의 놀이 중심 접근에서 운동 및 심리운동 중심 연구로 발전하였으며, 이후 감각통합, 신체활동 및 ICT 기반 중재가 등장하면서 다양한 중재전략이 활용되는 양상을 나타냈다. 최근에는 놀이와 신체활동을 기반으로 한 통합형 프로그램 연구가 지속적으로 수행되는 것으로 확인되었다.

성과 측면에서는 초기의 행동, 정서, 언어와 같은 기초 발달영역 중심 연구에서 인지, 적응, 문제해결 및 주의집중을 포함하는 다차원적 발달성과 중심 연구로 확장되는 경향이 나타났다. 특히 최근 연구에서는 인지 및 주의집중과 같은 인지조절 영역이 주요 연구주제로 부각되고 있는 것으로 확인되었다.

이러한 결과는 발달지연 유아 운동중재 연구가 단순한 신체기능 향상 중심 연구에서 벗어나 인지, 행동 및 사회적 참여를 포함하는 통합적 발달지원 연구로 발전하고 있음을 보여준다. 또한 키워드 네트워크 분석은 특정 연구영역의 지식구조 변화와 연구동향을 탐색하는 데이터사이언스 기반 문헌분석 기법으로서 유용한 분석 도구로 활용될 수 있음을 확인하였다.

본 연구는 개별 중재의 효과성 검증에 집중되어 있던 기존 연구에서 나아가 연구영역 전체의 지식구조 변화와 발전 방향을 탐색하였다는 점에서 의의가 있다. 특히 중재유형과 성과영역의 변화 과정을 체계적으로 정리함으로써 향후 발달지연 유아 운동중재 연구의 방향 설정과 통합적 중재 프로그램 개발을 위한 기초자료를 제공할 수 있을 것으로 기대된다.

향후 연구에서는 국내 연구뿐 아니라 국제 문헌을 포함한 비교 분석을 수행하고, 시계열 네트워크 분석을 통해 연구주제 변화 과정을 보다 정교하게 규명할 필요가 있다.

참고문헌

[1] M. Callon, J. P. Courtial, W. A. Turner, and S. Bauin, “From Translations to Problematic Networks: An Introduction to Co-word Analysis,” Soc. Sci. Inf., Vol. 22, No. 2, pp. 191-235, 1983. doi:10.1177/053901883022002003

[2] Q. He, “Knowledge Discovery Through Co-word Analysis,” Libr. Trends, Vol. 48, No. 1, pp. 133-159, 1999.

[3] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow et al., “The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews,” BMJ, Vol. 372, Article n71, 2021. doi:10.1136/bmj.n71

[4] S. P. Borgatti and M. G. Everett, “Network Analysis of 2-Mode Data,” Social Networks, Vol. 19, No. 3, pp. 243-269, 1997. doi:10.1016/S0378-8733(96)00301-2

[5] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge, UK, 1994. doi:10.1017/CBO9780511815478

※ 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2025S1A5C2A02022632)

공공데이터를 활용한 경기도 지역별 카드 소비 특성 비교분석 및 시각화

이도윤*, 정미현(교신)**
*차의과학대학교 미래융합대학, **차의과학대학교 차오름교양대학
e-mail: rheedoy@gmail.com*, eduforest@cha.ac.kr**

Analysis and Visualization of Regional Spending Characteristics in Gyeonggi-do Using Public Data

Doyun Lee*, Mihyun Chung**
*College of Future Convergence, Cha University
**College of Chaoreum Liberal Arts, Cha University

요 약

최근 공공데이터 개방 확대와 데이터 분석 기술의 발전으로 데이터 기반 지역 분석 연구가 활발히 이루어지고 있다. 본 연구는 경기도 카드 소비 공공데이터와 인구 데이터를 활용하여 지역별 소비 규모와 소비 구조의 차이를 분석하였다. 분석 대상은 2025년 12월 기준 경기도 12개 시·군의 카드 소비 데이터이며, 지역 규모 차이를 고려하여 총소비금액과 1인당 소비액을 함께 비교하였다. 분석 결과, 총소비금액은 화성시·용인시·수원시에서 높게 나타났으며, 1인당 소비액은 광명시·의왕시·포천시가 상대적으로 높은 수준을 보였다. 업종별 소비구조 분석에서는 하남시의 소매·유통, 과천시의 음식·교육·여가, 수원시와 안양시의 의료·건강 분야 소비 비중이 상대적으로 높게 나타나는 등 지역별 소비 특성의 차이가 확인되었다. 본 연구는 공공데이터와 데이터 시각화를 활용하여 지역 소비 특성을 비교·분석하였다는 점에서 의의가 있다.

키워드: 공공데이터, 카드 소비 데이터, 지역 소비 구조, 데이터 시각화, 소비 패턴

I. 서 론

최근 공공데이터 개방 확대와 데이터 기반 행정의 중요성이 높아지면서 다양한 공공데이터가 정책 및 연구 분야에서 활용되고 있다[1]. 특히 공공데이터를 활용한 데이터 분석과 시각화 연구가 활발하게 이루어지고 있으며, 카드 소비 데이터를 활용하여 지역별 소비 구조와 경제 특성을 분석하려는 연구도 증가하고 있다[2]. 카드 소비 데이터는 실제 소비 활동을 기반으로 수집된다는 점에서 지역 간 소비 특성을 비교하고 지역 경제의 차이를 분석하는 데 활용 가치가 높으며, 지역의 경제 특성과 소비 구조를 파악할 수 있는 유용한 자료로 활용되고 있다[3]. 지역별 소비 구조의 차이를 파악하는 것은 지역 경제의 특성과 소비 환경을 이해하는 데 중요한 기초 자료가 되며, 지역별 소비 특성과 상권 구조의 차이를 비교·분석하는 데 유용하게 활용될 수 있다.

이에 본 연구에서는 경기도 카드 소비 공공데이터와

인구 데이터를 활용하여 지역별 소비 규모와 소비 구조의 차이를 분석하고자 하였다. 이를 위해 총소비금액과 1인당 소비액을 비교하고, 업종별 소비 비율을 분석하여 지역별 소비 특성을 탐색하였다.

II. 이론적 배경

지역 소비 구조는 지역의 경제활동과 소비 특성을 이해하는 데 중요한 지표로 활용될 수 있다. 송소연과 이윤석(2026)은 소비 데이터가 지역의 경제활동, 생활양식 및 상권 구조를 반영하는 유용한 자료이며, 도시의 복합적 특성을 고려할 때 지역 소비 구조를 이해하기 위해서는 단순한 소비 규모뿐만 아니라 업종별 소비 비중을 함께 고려할 필요가 있다고 보고하였다[3]. 또한 김현아 외 2인(2025)은 대규모 상업시설의 입지가 지역 상권과 소비 변화에 영향을 미칠 수 있음을 분석하였다[4]. 이는 지역별 소비 환경에 따라 소비가 집중되는 업종과 소비 형태가 달라질 수 있음을 시사한다.

최근에는 설문조사의 한계를 보완하고 실제 소비 활동을 반영하기 위해 카드 소비 데이터를 활용한 연구가 증가하고 있다. 카드 소비 데이터는 응답자의 기억이나 인식에 의존하는 설문조사와 달리 실제 소비 행위를 기반으로 수집된다는 점에서 소비 특성 분석에 유용하며, 선행연구에서도 신용카드 사용 데이터가 실제 소비 흐름을 반영하고 지역별 소비 특성과 소비 패턴을 분석하는 데 활용될 수 있음이 보고되고 있다[3].

한편 공공데이터와 데이터 시각화는 대량의 데이터를 효과적으로 분석하고 지역 간 차이를 직관적으로 이해하는 데 활용되고 있다. 특히 시각화는 데이터의 구조와 패턴, 변수 간 관계를 보다 명확하게 표현함으로써 분석 결과의 이해를 높이는 데 기여한다. 이에 본 연구에서는 경기도 카드 소비 공공데이터와 인구 데이터를 활용하여 지역별 소비 규모와 소비 구조를 비교하고, 시각화를 통해 지역 간 소비 특성의 차이를 분석하고자 하였다.

III. 분석 방법 및 결과

본 연구에서는 공공데이터포털에서 제공하는 2025년 12월 경기도 카드 소비 데이터와 경기통계의 시군별 인구 데이터를 활용하였다. 분석 대상은 과천시, 광명시, 시흥시, 수원시, 안산시, 안양시, 용인시, 의왕시, 포천시, 하남시, 화성시, 이천시 등 12개 지역이다. 지역별 소비 구조를 비교하기 위하여 총소비금액, 1인당 소비액 및 업종별 소비 비율을 산출하였으며, Excel과 Power BI를 활용하여 데이터 전처리와 시각화 분석을 수행하였다.

분석 결과, 총소비금액은 화성시, 용인시, 수원시에서 높게 나타났으며, 과천시와 포천시는 상대적으로 낮은 수준을 보였다. 반면 1인당 소비액은 광명시가 가장 높게 나타났고, 의왕시와 포천시도 비교적 높은 수준을 보였다. 이는 소비 규모와 소비 수준이 반드시 동일한 방향으로 나타나지 않음을 보여준다.

업종별 소비구조 분석 결과, 전체적으로 소매·유통과 음식 업종의 비중이 높게 나타났다. 지역별로는 과천시가 음식·학문·교육·여가·오락 분야, 하남시가 소매·유통 분야, 수원시와 안양시가 의료·건강 분야에서 상대적으로 높은 소비 비중을 보였다. 광명시는 미디어·통신, 포천시와 이천시는 생활서비스, 시흥시는 공연·전시 분야의 소비 비중이 상대적으로 높게 나타났다. 이러한 결과는 지역별 소비 구조와 소비 특성에 차이가 존재함을 보여주며, 지역의 상권 구조와 생활환경, 소비 여건 등이 소비 패턴 형성에 영향을 미칠 수 있음을 시사한다.

IV. 결론 및 향후 연구 계획

본 연구에서는 경기도 지역별 카드 소비 데이터와 인구 데이터를 활용하여 지역별 소비 규모와 소비 구조의 차이를 비교·분석한 결과, 지역별 소비 규모와 소비 수준은 서로 다른 양상을 보였으며, 업종별 소비구조 또한 지역에 따라 차이가 나타났다. 이는 지역 소비 특성을 이해하기 위해 소비 규모와 소비 구조를 종합적으로 고려할 필요가 있음을 시사한다. 이러한 결과는 공공데이터와 데이터 시각화를 활용한 지역 소비 특성 분석의 활용 가능성을 보여주며, 총소비금액과 1인당 소비액을 함께 고려함으로써 지역 소비 특성을 보다 다각적으로 파악할 수 있음을 확인하였다.

다만 본 연구는 2025년 12월의 카드 소비 데이터만을 활용하였으며, 카드 소비 데이터만으로는 지역 내 소비 활동 전체를 충분히 설명하는 데 한계가 있다. 향후 연구에서는 분석 기간을 확대하고 유동인구, 소득 수준, 연령 구조 등 다양한 지역 특성 변수를 함께 고려함으로써 지역 소비 구조를 보다 심층적으로 분석할 필요가 있다.

참고문헌

[1] 정보통신정책연구원, "공공데이터 개방 및 활용 현황 분석", 2023.

[2] 행정안전부, "공공부문 데이터 분석·활용 우수사례집", 2024.

[3] 송소연, 이윤석, "농촌 지역의 관광소비 유형화 연구: 신용카드 데이터 기반 군집분석", 관광진흥연구, Vol. 14, No. 1,

pp. 1-29, 2026.
[4] 김현아, 서진형, 조춘한, "복합쇼핑몰 출점이 지역 상권과 소비에 미치는 영향: 수원 사례를 중심으로", 유통연구, Vol. 30, No. 4, pp. 69-91, 2025.

생활 습관 및 건강 지표 기반 당뇨 위험도 예측 모델 비교 분석

이성철[O] 백수진*
[O]한신대학교 AI시스템반도체학
*한신대학교 AI시스템반도체학
e-mail:sungchul945@naver.com[O],croso79@hs.ac.kr

Comparative Analysis of Diabetes Risk Prediction Models Based on Lifestyle and Health Indicators

Sungchul Lee[O] Sujin Back*
[O]Department of AI System Semiconductor Engineering, Hanshin University
*Department of AI System Semiconductor Engineering, Hanshin University

요 약

당뇨병은 환자 수가 지속적으로 증가하고 있는 대표적인 만성질환으로, 조기 예측과 예방이 중요하다. 본 연구에서는 생활 습관 및 건강 지표 데이터를 활용하여 머신러닝 기반 당뇨병 예측 모델의 성능을 비교 분석하였다. BRFSS2015 데이터셋을 사용하였으며, 실제 의료 데이터 환경을 반영하기 위해 클래스 불균형 상태를 유지한 채 Logistic Regression, Decision Tree, Random Forest 모델을 비교하였다. 실험 결과 Logistic Regression과 Random Forest는 Accuracy 0.86을 보였으나 Recall은 각각 0.16과 0.18로 낮게 나타났다. 반면 Decision Tree는 Recall 0.33으로 가장 높았지만 과적합 문제가 확인되었다. 또한 Feature Importance 분석 결과 BMI가 가장 중요한 변수로 나타나 비만도가 당뇨병 예측에 큰 영향을 미치는 것을 확인하였다. 이를 통해 클래스 불균형 의료 데이터에서는 Accuracy뿐만 아니라 Recall을 함께 고려한 성능 평가가 중요함을 확인하였다. 본 연구 결과는 머신러닝 기반 당뇨병 조기 예측 및 예방 관리 시스템 개발에 활용될 수 있을 것으로 기대된다.

키워드: Diabetes Prediction, Machine Learning, Class Imbalance, Recall, Feature Importance, BMI

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

I. 서론

현대 사회로 들어서며 식습관 변화와 고열량 음식 섭취 증가 등의 영향으로 당뇨병환자 수가 지속적으로 증가하고 있다[1]. 당뇨병은 다양한 합병증을 유발할 수 있기 때문에 조기 예측과 예방이 중요하다. 최근에는 의료 데이터와 머신러닝 기술을 활용하여 당뇨병 위험도를 예측하려는 연구가 활발히 진행되고 있다[2][3]. 기존 연구에서는 클래스 불균형 문제를 해결하기 위해 SMOTE와 같은 sampling 기법을 적용하거나, Feature Selection 기법을 활용하여 BMI 등의 주요 건강 지표를 선택하는 연구가 진행되었다[2][3]. 또한 Random Forest, XGBoost 등의 머신러닝 모델을 활용하여 높은 예측 성능을 얻고자 하는 연구도 수행되었다[2]. 그러나 기존 연구들은 데이터 균형 조정을 기반으로 성능 향상에 집중한 경우가 많았으며, 실제 의료 데이터 환경에서 발생하는 클래스 불균형 문제를 충분히 반영하지 못한 한계가 있다. 이에 따라 본 연구에서는 생활 습관 및 건강 지표 데이터를 활용하여 당뇨병 발병 요인을 분석하고, 머신러닝 기반 모델을 통해 당뇨병 위험군을 분류하고자 하였다. 따라서 실제 의료 데이터의 클래스 불균형 상태를 유지한 채 Logistic Regression, Decision Tree, Random Forest 모델의 특성과 한계를 비교 분석하였다. 또한 Feature Importance 분석을 통해 당뇨병 예측에 영향을 미치는 주요 건강 요인을 확인하고자 하였다.

II. 관련 연구

2.1 머신러닝 기반 당뇨병 예측 연구

당뇨병 예측 연구에 활용되는 대표적인 머신러닝 모델로는 Logistic Regression, Decision Tree, Random Forest 등이 있다. 각 모델은 서로 다른 특징과 장단점을 가지고 있으며, 의료 데이터 분석 및 질병 예측 분야에서 널리 활용되고 있다.

Logistic Regression은 대표적인 이진 분류 모델로, 입력 데이터와 결과 사이의 선형 관계를 기반으로 분류를 수행한다. 또한 각 feature가 결과에 미치는 영향을 비교적 쉽게 확인할 수 있다는 장점이 있어 의료 데이터 분석 분야에서 널리 활용되고 있다. Decision Tree는 조건 분기를 기반으로 데이터를 분류하는 모델로, 분류 과정을 직관적으로 해석할 수 있다는 장점이 있다. 반면 학습 데이터에 과도하게 적합될 경우 과적합 문제가 발생할 수 있다는 한계가 존재한다. Random Forest는 여러 개의 Decision Tree를 결합한 앙상블 모델로, 각 트리의 결과를 종합하여 최종 결과를 예측한다. 이를 통해 Decision Tree의 과적합 문제를 완화하고 보다 안정적인 예측 성능을 얻을 수 있어 다양한 질병 예측 연구에 활용되고 있다.

또한 당뇨병 예측의 정확도를 높이기 위해 Feature Selection 기법과 다양한 머신러닝 모델을 활용한 연구가 수행되었다. 이를 통해 중요한 변수만을 선택하여 예측 성능을 향상시키고자 하였으며, 의료 데이터의 클래스 불균형 문제를 해결하기 위해 데이터 균형 조정을 적용한 연구도 진행되었다. 그러나 실제 의료 데이터는 클래스 불균형 특성을 가지는 경우가 많으며, 이러한 환경에서는 소수 클래스에 해당하는 질환자를 정확하게 분류하는 것이 중요하다. 단순 accuracy만을 기준으로 모델 성능을 평가할 경우 정상 데이터의 비율이 높아 대부분을 정상으로 예측하는 문제가 발생할 수 있다. 따라서 최근 연구에서는 accuracy뿐만 아니라 recall, precision, F1-score 등의 다양한 평가지표를 함께

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

활용하여 모델의 성능을 평가하고 있다.

2.2 기존 연구의 한계 및 본 연구 방향

기존 연구에서는 sampling 기법을 활용하여 데이터 균형을 조정하거나, Feature Selection 기법을 통해 주요 feature를 선택하여 모델의 정확도를 향상시키고자 하였다. 그러나 이러한 방법들은 실제 의료 데이터 환경에서 발생하는 클래스 불균형 특성을 충분히 반영하지 못할 가능성이 있다. 따라서 본 연구는 데이터의 클래스 불균형 상태를 유지한 채 Logistic Regression, Decision Tree, Random Forest 모델을 비교 분석하였다. 이를 통해 각 모델의 과적합 문제와 Random Forest의 안정성을 확인하고, Feature Importance 분석을 통해 당뇨병 예측에 중요한 건강 지표를 분석하고자 하였다.

III. 제안방법

3.1 데이터셋 구성

본 연구에서는 BRFSS2015 데이터셋을 활용하였을 때 미국 질병통제예방센터(CDC)에서 수행한 건강 설문조사 데이터를 기반으로 구성되어 있으며,<표1>과 같이 구성되어 있다[4]. 특히 해당 데이터셋은 실제 의료 데이터와 유사하게 정상인 데이터의 비율이 당뇨 환자 데이터보다 훨씬 높은 클래스 불균형 특성을 가지고 있다. 이러한 특성은 머신러닝 모델이 정상 데이터에 편향된 예측을 수행하게 만들 수 있으며, 실제 당뇨 환자를 정확하게 탐지하는 데 한계가 발생할 수 있다. 따라서 본 연구에서는 실제 의료 환경을 반영하기 위해 별도의 데이터 균형 조정을 수행하지 않고 클래스 불균형 상태를 유지한 채 모델 학습 및 성능 분석을 진행하였다.

표1. 주요 Feature

Feature	Description
BMI	체질량지수
Age	연령
HighBP	고혈압 여부
Fruits	과일 섭취 여부

3.2 데이터 전처리

본 연구에서는 우선 데이터를 Feature와 Target으로 분리하였다. Feature에는 BMI, Age 등 건강 및 생활 습관 관련 데이터를 사용하였으며, Target에는 당뇨 여부를 사용하였다. 또한 데이터 전처리 과정에서 결측치는 제거한 후 모델 학습을 진행하였다. 이후 머신러닝 모델 학습 및 성능 평가를 위해 데이터를 8:2 비율로 학습용 데이터와 테스트용 데이터로 분리하였다.

데이터 분리 과정에서는 클래스 불균형으로 인해 특정 데이터셋에 당뇨 환자 데이터가 부족하거나 과도하게 포함되는 문제를 방지하기 위해 stratify=target 옵션을 적용하였다. 이를 통해 학습용 데이터와 테스트용 데이터 모두 실제 데이터와 유사한 당뇨 및 정상인 비율을 유지하도록 하였다. 기존 연구에서는 클래스 불균형 문제를 해결하기 위해 oversampling, undersampling 등의 방법을 활용하여 데이터 균형을 조정하는 경우가 많다. 그러나 본 연구에서는 실제 의료 데이터 환경을 반영하기 위해 별도의 데이터 균형 조정을 수행하지 않았으며, 클래스 불균형 상태를 유지한 채 모델 학습을 진행하였다. 이를 통해 실제 의료 데이터 환경에서 각 모델의 특성과 한계를 분석하고자 하였다.

3.3 머신러닝 모델 구성

본 연구에서는 Logistic Regression(LR), Decision Tree(DT), Random Forest(RF) 모델을 사

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

용하여 당뇨병 예측 성능을 비교하였다. 각 모델의 accuracy와 recall 성능을 비교하여 클래스 불균형 환경에서의 예측 성능을 분석하였다. 특히 의료 데이터 환경에서는 실제 당뇨 환자를 정확히 탐지하는 것이 중요하기 때문에 recall 성능을 중점적으로 분석하였다. 또한 Feature Importance 분석을 통해 당뇨병 예측에 영향을 미치는 주요 건강 지표를 확인하고자 하였다.

3.4 실험 결과 및 분석

머신러닝 모델 성능 비교 결과 <표2>와 같이 Logistic Regression과 Random Forest는 높은 accuracy를 보였으나 recall 성능은 낮게 나타났다. 반면 Decision Tree는 세 모델 중 가장 높은 recall 성능을 보였지만, train accuracy(TrAcc)와 test accuracy(TeAcc)의 차이가 크게 나타나 과적합 문제가 발생한 것으로 확인되었다. 본 연구에서는 이러한 과적합 문제를 완화하기 위해 Random Forest 모델을 적용하였다. Random Forest는 여러 개의 Decision Tree를 결합한 앙상블 모델로, 보다 안정적인 성능을 보였다. 그러나 Logistic Regression과 유사하게 accuracy는 높게 나타난 반면 recall 성능은 상대적으로 낮게 나타났다. 이는 실제 의료 데이터에서 정상인 데이터의 비율이 당뇨 환자 데이터보다 훨씬 높은 클래스 불균형 문제의 영향을 받은 결과로 해석할 수 있다.

표 2. 머신러닝 모델 성능 비교

	TrAcc	TeAcc	Recall	Preci.	F1
LR	0.86	0.86	0.16	0.52	0.24
DT	0.99	0.79	0.33	0.30	0.31
RF	0.99	0.86	0.18	0.49	0.26

<그림1>은 Logistic Regression 모델의 Confusion Matrix를 나타낸다. 좌측 상단(42622)은 실제 정상인을 정상인으로 예측한 경우(TN)를 의미하며, 우측 상단(1045)은 실제 정상인을 당뇨 환자로 잘못 예측한 경우(FP)를 의미한다. 좌측 하단(5950)은 실제 당뇨 환자를 정상인으로 잘못 예측한 경우(FN)를 나타내며, 우측 하단(1119)은 실제 당뇨 환자를 당뇨 환자로 정확하게 예측한 경우(TP)를 의미한다. 특히 실제 당뇨 환자 7069명 중 5950명을 정상인으로 분류하고 1119명만을 정확하게 탐지한 것으로 나타났다. 이는 Accuracy는 높게 나타났지만 Recall 성능이 낮게 측정된 원인을 보여준다. 따라서 클래스 불균형이 존재하는 의료 데이터에서는 Accuracy뿐만 아니라 실제 환자를 얼마나 탐지할 수 있는지를 나타내는 Recall 지표를 함께 고려해야 함을 확인할 수 있다.

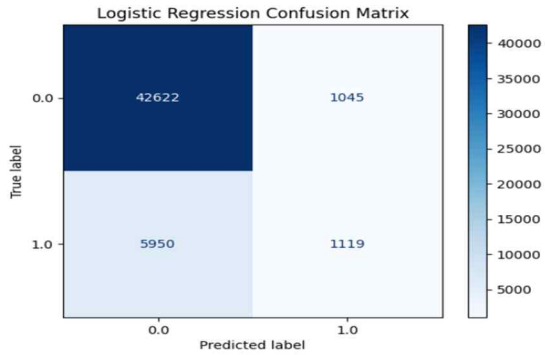


그림 1. LR 모델의 Confusion Matrix

따라서 의료 분야의 머신러닝 모델에서는 단순 Accuracy만으로 성능을 평가하는 데 한계가 존재한다. 실제 의료 환경에서는 정상인과 환자의 비율을 인위적으로 맞추는 것보다, 수많은 데이터 속에서 실제 질환자를 정확하게 탐지하는 것이 중요하기 때문이다. 이러한 이유로 본 연구에서는 Recall 성능을 중요하게 분석하였다. 또한 머신 러닝 모델 학습 이후 Feature Importance 분석을 수행하였다. Logistic Regression과 Random Forest 모델 모두 BMI가 가장 높은 중요도를 나타냈으며, Logistic Regression에서는 0.408314, Random Forest에서

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

는 0.184123의 값을 보였다. 이를 통해 비만도가 당뇨병 예측에 중요한 영향을 미치는 요인임을 확인할 수 있었다.

표 3. 모델별 주요 Feature 비교

	1순위	2순위	3순위
LR	GenHlth:0.57	BMI:0.40	Age:0.38
RF	BMI:0.18	Age:0.12	Income:0.09
DT	BMI:0.14	Income:0.10	Age:0.09

<표 3>은 각 모델에서 중요하게 판단한 상위 Feature를 나타낸다. 세 모델 모두 BMI와 Age를 주요 변수로 판단하였다. 특히 BMI는 Random Forest와 Decision Tree에서 가장 높은 중요도를 나타냈으며, Logistic Regression에서도 상위 변수로 확인되었다. 또한 Age와 Income, GenHlth 역시 높은 중요도를 보여 생활 습관 및 건강 상태와 관련된 다양한 요소들이 당뇨병 예측에 영향을 미치는 것을 확인할 수 있었다.

IV. 결론

본 연구에서는 생활 습관 및 건강 지표 데이터를 활용하여 당뇨병을 예측하고, 머신러닝 모델별 성능 차이를 비교 분석하고자 하였다. 이를 위해 Logistic Regression, Decision Tree, Random Forest 모델을 활용하여 당뇨병 예측 성능을 비교하였다. 실험 결과 Decision Tree는 가장 높은 Recall 성능을 보였으나 Train Accuracy와 Test Accuracy의 차이가 크게 나타나 과적합 문제가 발생하였다. Random Forest는 보다 안정적인 성능을 보였지만 Logistic Regression과 마찬가지로 Recall 성능이 낮게 나타나 실제 당뇨 환자를 탐지하는 데 한계가 존재하였다. 이러한 결과는 실제 의료 데이터에서 발생하는 클래스 불균형 문제의 영향을 받은 것으로 해석할 수 있다. 또한 Feature Importance 분석 결과 모든 모델에서 BMI가 높은 중요도를 나타냈으며, 이를 통해 비만도가 당뇨병 예측에 중요한 영향을 미치는 요소임을 확인하였다. BMI 외에도 Age와 Income 등의 변수 역시 비교적 높은 중요도를 보였으며, 다양한 생활 습관 및 건강 지표가 당뇨병 예측에 영향을 미치는 것을 확인할 수 있었다. 본 연구를 통해 의료 데이터 환경에서는 Accuracy뿐만 아니라 실제 환자 탐지 능력을 나타내는 Recall을 함께 고려한 성능 평가가 중요함을 확인하였다. 향후 연구에서는 클래스 불균형 문제를 보다 효과적으로 해결할 수 있는 추가적인 머신러닝 기법과 데이터 처리 방법에 대한 연구가 필요할 것으로 생각된다.

참고문헌

[1] World Health Organization, Diabetes, WHO Fact Sheet (2024). https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] M.M. Chowdhury, R.S. Ayon, M.S. Hossain, An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset, Healthcare Analytics 5 (2024) 100297. https://doi.org/10.1016/j.health.2023.100297

[3] S. Gündoğdu, Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique, Multimedia Tools and Applications 82 (2023) 24225–24243. https://doi.org/10.1007/s11042-023-15165-8

[4] A. Teboul, Diabetes Health Indicators Dataset, Kaggle (2021). https://www.kaggle.com/datasets/alexteboul/diabetes-

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

health-indicators-dataset

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

AI Agent를 활용한 대학생 맞춤 정보 시스템 설계 및 구현에 대한 연구

최부영*, 배현욱*, 남재욱*, 전예은*, 홍승필*
* 한신대학교 AI·SW대학
e-mail: {ra020509, smfwn531, bh396582, jjh1221jjh, sphong}@hs.ac.kr

A Study on the Design and Implementation of an AI Agent-Based Personalized Information System for University Students

Bu-Young Choi*, Hyun-Wook Bae*, Jae-Wook Nam*, Yeeun Jun*, Seung-Pil Hong*
*School of Computing and Artificial Intelligence, HanShin University

요 약 대학 공지·학사·장학 정보는 여러 채널에 분산되어 학생이 필요한 정보를 적시에 찾기 어렵다. 본 연구는 한신대학교 재학생을 위한 AI 학교생활 도우미 UniMate를 제안한다. UniMate는 정보를 제공하는 동시에 사용자의 상호작용 신호(채팅 질의·북마크·일정/과제·관심 태그)를 수집하여 사용자 관심 프로파일을 갱신하고, 이를 다시 추천에 반영하는 폐루프(closed-loop) 구조를 채택한다. 시스템은 OpenAI gpt-4o 기반 LangChain Tool-Calling 에이전트와 pgvector 벡터 검색(RAG)으로 구성된다. 600명 규모의 합성 사용자 시뮬레이션에서 누적 상호작용이 증가할수록 Precision@5가 콜드스타트 0.19에서 0.62로(nDCG@10 0.75) 향상되어, 비개인화 베이스라인 대비 약 3.7배의 추천 정확도를 달성하였다.

키워드: AI 에이전트, 개인화 추천, 검색 증강 생성(RAG), 벡터 임베딩, 폐루프 개인화

I. 서론

대학생은 학기 중 장학·학사·취업·교내 행사·국제교류 등 성격이 상이한 정보를 끊임없이 확인해야 한다. 그러나 이들 정보는 학교 포털, 학과 게시판, 공지 메일 등 여러 채널에 분산되어 있어, 학생이 자신과 관련된 정보를 놓치지 않고 적시에 확인하기란 쉽지 않다. 정보의 양이 많아질수록 필요한 정보를 찾는 비용이 오히려 증가하는 정보 과부하 문제가 발생한다.

본 연구는 이 문제를 해결하기 위해 한신대학교 재학생을 위한 AI 기반 학교생활 도우미 서비스 UniMate를 설계·구현하였다. UniMate는 흩어진 정보를 자동으로 수집해 한곳에 모으고, 자연어 질의에 대해 대규모 언어모델(LLM) 에이전트가 적절한 도구(tool)를 스스로 선택하여 답변을 생성한다. 그러나 단순한 정보 검색·요약 서비스는 모든 사용자에게 동일한 결과를 제공한다는 한계가 있다. 1학년 신입생과 졸업을 앞둔 4학년, 공학 계열과 인문 계열 학생이 필요로 하는 정보는 근본적으로 다르다.

이에 본 연구는 데이터사이언스 관점에서 UniMate를 **정보 수집-제공 순환 구조(closed-loop collect-and-serve)**로 재정의한다. 핵심 아이디어는 다음과 같다. (1) 에이전트가 사용자에게 정보를 제공하는 모든 상호작용은 동시에 사용자의 관심을 드러내는 데이터 수집 행위이다. (2) 채팅 질의·북마크·일정/과제 등록·관심 태그 선택 등의 신호를 누적해 사용자 관심 프로파일을 임베딩 공간상의 벡터로 추정한다. (3) 추정된 프로파일을 이후 정보 제공·추천 단계에 다시 주입하여, 사용할수록 개인에게 더 정밀하게 맞춰지는 선순환을 형성한다.

본 논문의 기여는 다음과 같다. 첫째, LLM Tool-Calling 에이전트와 pgvector 기반 검색 증강 생성(RAG)을 결합한 대학 정보 도우미의 시스템 아키텍처를 제시한다. 둘째, 정보 제공과 데이터 수집을 하나의 순환으로 묶는 폐루프 개인화

구조를 정의하고, 어떤 상호작용 신호가 어떻게 사용자 프로파일로 변환되는지를 구체화한다. 셋째, 합성 사용자 시뮬레이션으로 개인화 추천 정확도가 누적 데이터에 따라 향상됨을 정량 검증하고 비개인화 베이스라인과 비교한다.

II. 관련 연구

2.1 LLM 에이전트와 도구 사용

최근 LLM은 외부 함수(도구)를 호출하여 검색·계산·DB 조회 등을 수행하는 도구 사용(tool use) 능력을 갖추게 되었다[4]. 에이전트는 사용자 질의를 분석해 어떤 도구를 어떤 인자로 호출할지 스스로 결정하고, 도구 실행 결과를 종합해 최종 답변을 생성한다. 이 방식은 모델 내부 파라미터에만 의존하는 생성보다 사실성이 높고 최신 데이터를 반영할 수 있다. UniMate는 LangChain의 도구 바인딩으로 세 개의 검색 도구를 에이전트에 연결한다.

2.2 검색 증강 생성(RAG)과 벡터 검색

검색 증강 생성(RAG)은 외부 문서를 임베딩하여 벡터 데이터베이스에 색인하고, 질의와 의미적으로 유사한 문서를 검색해 LLM 입력 컨텍스트로 제공하는 기법이다[1,2]. 이는 환각(hallucination)과 지식의 노후화 문제를 완화한다. 본 시스템은 PostgreSQL의 pgvector 확장을 사용해 공지와 학칙 문서를 1536차원 벡터로 색인하고, 코사인 유사도 기반 근사 최근접 탐색(IVFFlat)을 수행한다[3].

2.3 개인화 추천

추천 시스템은 사용자의 과거 행동으로부터 선호를 추정하여 관련 항목을 선별한다[5]. 협업 필터링은 사용자-항목 상호작용 행렬에 의존해 신규 항목·신규 사용자에 대한 콜드스타트 문제가 크다. 반면 콘텐츠 기반 접근은 항목 임베딩과 사용자 프로파일의 유사도로 추천하므로, 매일 새로운 공지가 유입되는 본 도메인에 적합하다. UniMate는 공지 임베딩과 사용자 관심 벡터의 코사인 유사도로 콘텐츠 기반 개인화를 수행하되, 사용자의 명시적·암묵적 상호작용을 모두 신호로 활용한다는 점에서 차별성을 가진다.

III. 제안방법

UniMate는 모바일 클라이언트, 백엔드 API, AI 에이전트, 데이터 수집 파이프라인의 네 부분으로 구성된다. 클라이언트는 React Native(Expo) 기반이며, 백엔드는 비동기 FastAPI로 구현되어 모든 데이터베이스 접근을 async/await로 처리한다. 데이터 저장소는 pgvector 확장이 활성화된 PostgreSQL과 캐시·작업 큐 역할의 Redis로 이원화된다.

AI 계층의 중심은 ScholarshipAgent이다. 이 에이전트는 OpenAI gpt-4o[6]를 LangChain 도구 바인딩으로 감싼 형태이며, 세 가지 도구를 보유한다. *fetch_notices*는 공지 제목·본문 키워드(ILIKE) 검색을, *search_by_deadline*은 최근 N일 내 마감 임박 공지 조회를, *answer_faq*는 notices와 qa_documents 임베딩에 대한 벡터 유사도 검색(코사인 0.75 이상)을 수행한다. 에이전트는 질의 성격에 따라 도구를 선택·조합하고 결과를 종합해 답변을 생성한 뒤 SSE로 토큰을 스트리밍한다.

응답 지연과 비용을 줄이기 위해 동일 질의 답변은 Redis 키 chat:cache:{md5(query)}로 1시간 캐싱하며, Redis 장애 시 인메모리 폴백한다. 임베딩 모델은 text-embedding-3-small(1536차원)[7]로 통일하여 공지와 질의가 동일 벡터 공간을 공유하도록 한다.

IV. 정보 수집-제공 순환 구조

본 절은 UniMate의 핵심 설계인 폐루프 개인화를 설명한다. 그림 1은 데이터가 어떻게 수집되고, 사용자 프로파일로 변환되며, 다시 맞춤형 정보 제공에 반영되는지를 보여준다. 제공(provision) 경로와 수집(collection) 경로가 하나의 순환을 이룬다는 점이 핵심이다.

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원 사업(경기RISE사업)의 연구결과 입니다.

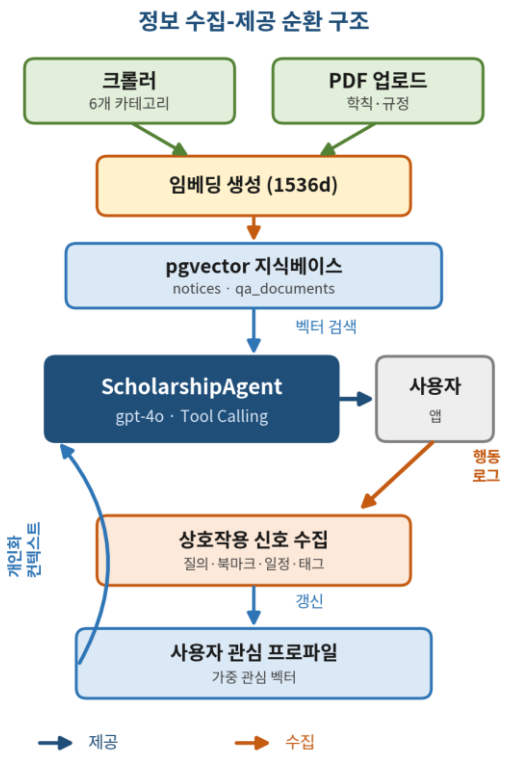


그림 1. UniMate의 정보 수집-제공 순환 구조.

파란 경로는 정보 제공, 주황 경로는 데이터 수집을 나타낸다.

4.1 데이터 수집 파이프라인 (콘텐츠 측)

콘텐츠 데이터는 두 경로로 유입된다. 첫째, Celery Beat 스케줄러가 운영 시간대(07–18시)에 하루 6회 한신대 포털의 6개 카테고리 공지를 크롤링한다. 수집된 공지는 source_url 기준으로 notices 테이블에 UPSERT되고, 저장 직후 본문을 임베딩하여 notices.embedding 컬럼을 갱신한다. 임베딩 실패 시에도 공지 본문 저장은 유지되도록 graceful하게 처리한다. 둘째, 관리자가 업로드한 학칙·규정 PDF는 텍스트 추출 후 1000자 단위(200자 중첩)로 분할되고, 각 청크가 임베딩되어 qa_documents 테이블에 적재된다. 두 테이블 모두 IVFFlat(lists=100) 코사인 인덱스를 가져 대규모 벡터 검색을 지원한다.

4.2 상호작용 신호 수집 (사용자 측)

정보 제공 과정에서 발생하는 모든 사용자 행동은 관심을 드러내는 신호로 기록된다. UniMate가 수집하는 신호와 그 의미는 표 1과 같다. 신호는 사용자가 의도를 분명히 표현한 명시적(explicit) 신호와, 행동에서 간접적으로 추론되는 암묵적(implicit) 신호로 구분된다. 명시적 신호일수록 사용자의 진짜 관심을 더 신뢰성 있게 반영한다.

표 1. UniMate가 수집하는 상호작용 신호와 개인화 해석

상호작용 신호	저장 위치	개인화 해석
관심 태그	user_interest_tags	가입 시 초기 관심(콜드스타트 완화)
공지 북마크	user_notice_bookmarks	명시적 관심 — 신뢰도 최상
채팅 질의	chat_messages	준명시적 — 현재 관심사 반영
일정/과제	user_schedules 등	학사 맥락(시험·마감) 시의성
공지 열람	로그/notifications	암묵적 — 약하나 풍부한 신호

4.3 사용자 관심 프로파일 구축

수집된 신호는 사용자 관심을 임베딩 공간상의 단일 벡터로 요약하는 데 사용된다. 사용자 u가 상호작용한 항목 집합을 I_u라 할 때, 관심 프로파일 p_u는 각 항목 임베딩 e_i를 신호 유형별 가중치 w_t로 가중 평균하여 추정한다: $p_u = normalize(\Sigma\ w_t \cdot e_i)$. 명시적 신호(북마크)에는 큰 가중치를, 암묵적 신호(열람)에는 작은 가중치를 부여한다. 이 추정치는 상호작용이 누적될수록 사용자의 잠재 관심에 점근적으로 수렴하며, 신규 사용자는 가입 시 선택한 관심 태그로 초기화되어 콜드스타트를 완화한다.

4.4 맞춤형 정보 제공

추정된 프로파일 p_u는 두 가지 방식으로 정보 제공에 반영된다. 첫째, 개인화 랭킹: 신규 공지 각각의 임베딩 e_i와 p_u의 코사인 유사도 cos(p_u, e_i)를 계산하여 상위 항목을 우선 노출한다. 둘째, 개인화 컨텍스트 주입: 에이전트가 답변을 생성할 때 사용자의 학과·학년·관심 분야를 컨텍스트로 함께 제공하여, 동일 질의라도 사용자에 맞는 정보를 우선 강조한다. 제공된 결과에 사용자가 다시 반응하면 그 행동이 또 다른 신호로 수집되어

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원 사업(경기RISE사업)의 연구결과 입니다.

4.2–4.4가 하나의 순환으로 닫힌다.

V. 개인화 추천 방법론

개인화 추천을 콘텐츠 기반 매칭 문제로 정식화한다. 모든 공지와 사용자 프로파일은 동일한 d=1536 차원 단위 벡터로 표현된다. 신규 공지 후보 집합 C에 대해 사용자 u에게 제시할 순위는 점수 함수 $s(u,i)=\cos(p_u, e_i)$의 내림차순으로 결정한다. p_u는 4.3절의 가중 평균으로 갱신되며, 온라인 환경에서는 새 신호가 들어올 때마다 누적합과 누적 가중치를 갱신하는 증분 방식으로 O(d) 비용에 업데이트된다.

이 설계는 두 가지 데이터사이언스적 함의를 가진다. 첫째, 추천 품질이 수집된 상호작용 데이터의 양·질에 직접 의존하므로 정보 제공 자체가 데이터 자산을 축적하는 행위가 된다. 둘째, 공지가 매일 새로 유입되어도 사용자-항목 행렬을 재학습할 필요 없이 임베딩 유사도만으로 즉시 추천 가능하므로 신규 항목 콜드스타트에 강건하다.

VI. 실험: 개인화 추천 정확도 시뮬레이션

6.1 시뮬레이션 설계

실서비스의 장기 상호작용 로그를 확보하기 전 단계에서 페루프 개인화의 효과를 검증하기 위해, 본 도메인 구조를 반영한 합성 사용자 시뮬레이션을 수행하였다. 한신대 크롤러의 6개 카테고리에 대응하는 6개 토픽 프로토타입 벡터를 정의하고, 600명의 합성 사용자 각각에게 소수 토픽에 집중된 잠재 관심 벡터(Dirichlet 분포)를 부여하였다. 매 시점 새로운 공지 후보 40건이 유입되며, 각 공지는 한 카테고리에 속하되 잡음을 갖는다.

사용자는 자신의 잠재 관심과 유사한 공지에 더 높은 확률로 반응(로지스틱 반응 모델)하고, 그 반응이 신호로 수집된다. 수집된 신호는 4.3절의 가중 평균으로 프로파일을 갱신한다. 신호 유형은 표 1을 따라 명시적·암묵적으로 구분되며, 암묵적 신호일수록 관심을 잘못 반영할 확률(노이즈)이 높게 모델링되었다. 각 평가 시점마다 별도의 신규 공지로 추천을 수행하여 학습-평가 누수를 방지하였다.

6.2 평가 지표와 베이스라인

지표로는 추천 상위 5건 중 실제 관심 공지의 비율인 Precision@5와, 순위 품질을 반영하는 nDCG@10[8]을 사용하였다. 비교 대상은 (i) 본 연구의 개인화 에이전트, (ii) 전역 인기도만 사용하는 인기순 베이스라인, (iii) 게시 순서를 따르는 최신순(사실상 무작위) 베이스라인이다. 무작위 추천의 기대 Precision@5는 8/40 = 0.20이다.

6.3 결과: 누적 데이터에 따른 정확도 향상

그림 2는 누적 상호작용 수가 증가함에 따른 추천 정확도 변화를 보여준다(음영은 95% 신뢰구간). 개인화 에이전트의 Precision@5는 콜드스타트 시점 0.19(무작위 수준)에서 시작하여 약 10회 상호작용 후 0.55, 20회 후 0.60에 도달하고 0.62 부근으로 수렴하였다. nDCG@10 역시 0.75까지 향상되었다. 반면 비개인화 인기순·최신순 베이스라인은 데이터가 쌓여도 각각 0.17, 0.20 수준에 정체하였다.

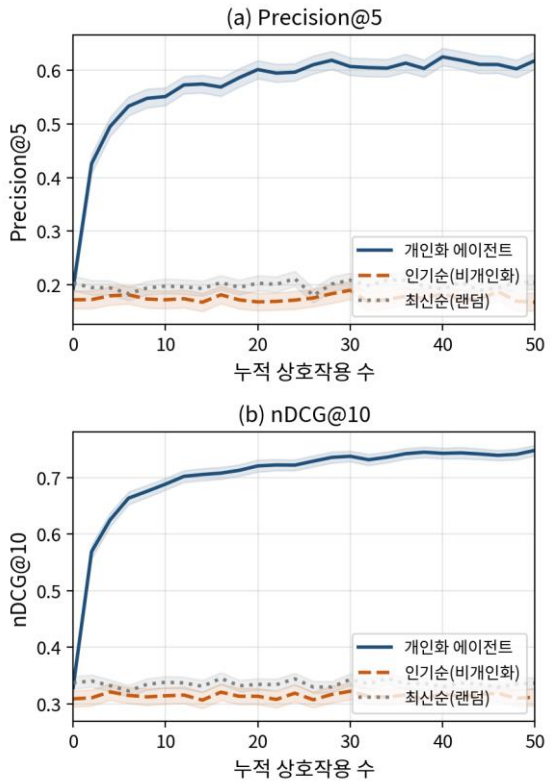


그림 2. 누적 상호작용 대비 (a) Precision@5, (b) nDCG@10. 개인화 에이전트는 데이터 누적에 따라 향상되나 비개인화 베이스라인은 정체한다.

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원 사업(경기RISE사업)의 연구결과 입니다.

수렴 시점에서 개인화 에이전트는 인기순 베이스라인 대비 약 3.7배, 무작위 대비 약 3.1배 높은 Precision@5를 달성하였다. 이는 정보 제공 과정에서 수집된 상호작용 신호가 실제로 개인화 품질을 끌어올리며 폐루프 구조가 의도대로 작동함을 시사한다. 표 2는 주요 시점의 수치를 요약한다.

표 2. 주요 시점별 추천 정확도 요약

Precision@5	개인화	인기순	최신순
콜드스타트 (t=0)	0.19	0.17	0.20
상호작용 10회	0.55	0.17	0.20
상호작용 20회	0.60	0.17	0.20
수렴 (t=50)	0.62	0.17	0.20
nDCG@10 (t=50)	0.75	0.31	0.34

6.4 사용자 활동량과 개인화 수렴 속도

폐루프 구조의 핵심 가설은 수집되는 데이터가 많을수록 개인화가 좋아진다는 것이다. 이를 직접 검증하기 위해 사용자를 활동량(시점당 상호작용 빈도)에 따라 高·中·低 세 집단으로 나누어 수렴 속도를 비교하였다. 그림 3에서 보듯 활동량이 많은 탐색적 사용자는 더 적은 경과 시간에 높은 정확도에 도달한 반면, 소극적 사용자는 같은 수준에 이르는 데 더 오랜 시간이 걸렸다. 다만 모든 집단이 결국 무작위 기대선(0.20)을 크게 상회하였다.

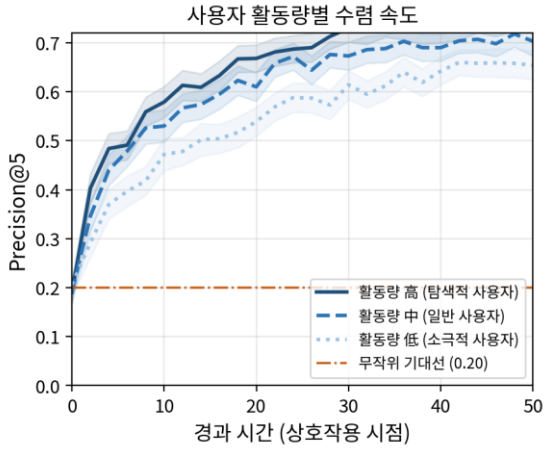


그림 3. 사용자 활동량별 개인화 수렴 속도. 활동량이 많을수록(=수집 데이터가 많을수록) 빠르게 수렴한다.

이 결과는 두 가지 운영상의 시사점을 준다. 첫째, 가입 시 관심 태그 입력 같은 명시적 초기 신호는 콜드스타트 구간의 정확도를 끌어올리는 데 효과적이다. 둘째, 소극적 사용자에게는 적절한 추천·알림으로 상호작용을 유도하는 것이 개인화 품질 개선의 지렛대가 된다.

VII. 논의 및 한계

7.1 실제 운영 환경에서의 검증

본 연구의 실험은 합성 사용자 시뮬레이션에 기반한다. 따라서 보고된 정확도 향상(콜드스타트 0.19 → 0.62, 비개인화 대비 약 3.7배)은 폐루프 개인화 구조가 통제된 조건에서 의도대로 작동함을 보이는 타당성 검증으로 해석해야 하며, 실제 학생의 관심 변화·계절성(수강신청·시험 기간 등)·반응 편향을 단순화한 결과다. 또한 단일 관심 벡터는 사용자의 다면적 관심(예: 장학과 취업을 동시에 중시)을 부분적으로만 표현하므로, 다중 프로파일 또는 시간 가중(time-decay) 프로파일로의 확장이 필요하다. 실제 운영 성능을 검증하기 위해 향후 연구에서는 (1) 한신대 재학생을 대상으로 한 파일럿 배포로 실제 상호작용 로그를 수집하고, (2) 개인화군과 비개인화군을 비교하는 온라인 A/B 테스트로 실데이터 기반 Precision@5·nDCG·클릭률(CTR)과 사용자 만족도를 측정하며, (3) 학기 주기에 따른 관심 변화와 신규 사용자 유입을 반영한 장기 효과를 분석한다.

7.2 개인정보 보호

상호작용 로그는 학번·연락처·열람 이력 등 민감한 학사 정보를 포함할 수 있으므로, 본 시스템은 개인정보보호법에 따라 동의 기반 수집과 목적 제한, 데이터 최소화를 전제로 설계한다. 구체적으로 학번·연락처 등 식별 정보는 가명화·암호화하여 저장하고, LLM API에는 개인 식별 원문을 전달하지 않도록 질의에서 개인정보를 마스킹한다. 또한 로그에 대한 접근 통제와 보관 기간 제한을 적용하고, 학생 본인이 자신의 수집 데이터를 열람·삭제할 수 있는 권리를 보장한다. 이를 통해 개인화에 필요한 신호를 활용하면서도 프라이버시 위험을 최소화한다.

7.3 추천 다양성과 필터 버블

개인화는 사용자의 기존 관심에 부합하는 정보만 반복 노출하여 필터 버블을 유발할 수 있다. 이를 완화하기 위해 추천 단계에 탐색(exploration)을 일정 비율 섞어(ε-탐색) 관심 밖 카테고리의 공지도

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원 사업(경기RISE사업)의 연구결과 입니다.

노출하고, 다양성 재순위화로 상위 결과의 주제 편중을 억제한다. 또한 장학·안전·학사 일정 등 전체 학생에게 중요한 공지는 개인화 점수와 무관하게 항상 노출되도록 보장한다. 운영 시에는 카테고리 커버리지·결과 내 다양성(intra-list diversity) 등 다양성 지표를 함께 모니터링하여, 추천 정확도와 정보 다양성 간의 균형(탐색-활용 균형)을 지속적으로 조정한다.

VIII. 결론

본 연구는 대학생을 위한 AI 정보 도우미 UniMate를 정보 수집-제공 순환 구조로 설계하고 그 개인화 효과를 시뮬레이션으로 검증하였다. 에이전트가 정보를 제공하는 모든 상호작용을 데이터 수집 기회로 활용하여 사용자 관심 프로파일을 갱신하고 이를 다시 추천에 반영하는 폐루프 설계는 누적 데이터에 따라 추천 정확도를 콜드스타트 0.19에서 0.62로(nDCG@10 0.75) 향상시켰으며, 비개인화 베이스라인 대비 약 3.7배의 정확도를 달성하였다. 또한 활동량이 많을수록 개인화가 빠르게 수렴함을 확인하였다. 향후 연구로는 한신대 파일럿 배포와 온라인 A/B 테스트를 통한 실사용자 로그 기반 검증, 동의 기반 수집·가명화 등 개인정보 보호 체계의 적용, 시간 가중·다중 관심 프로파일, 그리고 탐색-활용 균형을 통한 추천 다양성 확보를 계획한다.

참고문헌

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[2] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv:2312.10997, 2024.

[3] J. J. Pan, J. Wang, and G. Li, "Survey of Vector Database Management Systems," The VLDB Journal, vol. 33, no. 5, pp. 1591–1615, 2024.

[4] L. Wang et al., "A Survey on Large Language Model based Autonomous Agents," Frontiers of Computer Science, vol. 18, no. 6, 186345, 2024.

[5] Z. Zhao, W. Fan et al., "Recommender Systems in the Era of Large Language Models (LLMs)," IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 11, pp. 6889–6907, 2024.

[6] OpenAI, "GPT-4o System Card," OpenAI Technical Report, 2024.

[7] OpenAI, "New Embedding Models and API Updates (text-embedding-3)," OpenAI, 2024.

[8] K. Järvelin and J. Kekäläinen, "Cumulated Gain-based Evaluation of IR Techniques," ACM Transactions on Information Systems, vol. 20, no. 4, pp. 422–446, 2002.

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원 사업(경기RISE사업)의 연구결과 입니다.

모바일 온디바이스 환경에서 소형 언어 모델(SLM)의 [redacted] 체급 및 양자화 스케일이 인간 인지 임계 속도에 미치는 영향

손혁진(한신대학교), 판반 뚱람(한신대학교), 강민구(한신대학교)
한신대학교(AI.SW대학)
e-mail: sonhukjin78@hs.ac.kr , phamvantunglam@hs.ac.kr
kangmg@hs.ac.kr

The Impact of Small Language Model (SLM) Size and Quantization Scale on Human Cognitive Critical Speed in Mobile On-Device Environments

Son Hyuk-Jin, Phan Van Ttung Lam, Kang Min-gu
College of AI.SW, Hanshin University

요 약

본 연구에서는 모바일 기기에서 소형 언어 모델(SLM)을 직접 구동할 때 양자화 수준(Q4~Q8)이 속도와 메모리 사용량에 미치는 영향을 실측하였다. 동일 모델이라도 기기 스펙에 따라 성능이 달라지므로, 생성 파라미터(Context Window, Temperature, Top-K/P)는 고정하고 양자화 비트만 변경하여 초당 토큰 생성 수와 최대 메모리 점유량을 측정하였다. 또한 사람의 읽기 속도와 반응 지연에 관한 HCI 이론을 결합해, 기기 성능에 맞는 최적 양자화 모델 선택 기준을 제안한다. Primativo 등은 RSVP 실험에서 안구운동이 약 300 wpm 수준의 상한을 부과한다고 보고했으며, 이는 인간 인지적 한계와 모델 성능을 연결하는 중요한 근거가 된다

키워드: LLM , 양자화 , 사용자 만족도 , 초당 토큰 생성수

I. 서론

1. 서론 (Introduction)

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심대학지원사업(경기 RISE사업)의 연구결과입니다.

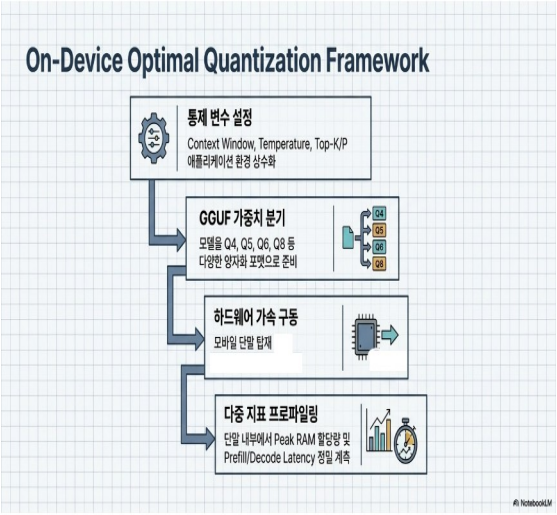
LLM이 빠르게 발전하면서 텍스트는 물론 이미지, 음성까지 처리하는 모델이 등장했지만, 수천억 개 파라미터를 대형 모델은 스마트폰 같은 저사양 자원 제약적인 디바이스는 독립형 구동자체가 불가능합니다. 더욱이 군사, 의료, 금융처럼 기밀성이 우선시 되는 도메인에서는 기존 외부 클라우드를 통한 송신하는 방식은 보안리스크를 유발합니다.

그래서 나온 것이 기기에서 직접 추론하는 온디바이스 LLM입니다. 삼성 갤럭시에 구글 Gemini Nano 가네이티브 탑재된 사례 역시 서버 운영비를 절감하고 개인정보를 디바이스 내에서 폐쇄하기위한 시도입니다.

그런데 모바일 환경에는 고유한 문제가 있습니다. 메모리사용량이 제한되어 있어서, LLM추론 연산시ART의 임계치를 초과할경우 안드로이드 커널 레벨에서 프로세스가 강제종료되는 결함을 가지고 있습니다. (OOM Killer)[1]. 이를 피하려고 모델 크기와 연산량을 줄이는 GGUF 양자화[2] 기술을 쓰는데, 양자화 비트를 낮출수록 메모리는 줄지만 속도나 품질에 트레이드오프가 생깁니다.

더욱이 이러한 성능저하는 단순 하드웨어 벤치마크 지표의 변동에 국한되지 않습니다. 이게 단순한 기술 지표 문제가 아니라는 점입니다. [3] jakob Nielsen 의 Response Time Limits 연구에 따르면, 사용자가 대화창에 집중할 수 있는 최대 시간은 약 10초입니다. 그보다 길어지면 사용자는 컴퓨터가 작업을 완료할 때까지 다른 작업을 하려고 할 것이므로, 컴퓨터가 언제 완료될 것으로 예상하는지 알려주는 피드백을 제공해야 합니다. 결국 온디바이스 LLM 최적화는 하드웨어 성능, 운영체제 자원 관리, 그리고 사람의 인지 특성을 동시에 고려해야 하는 문제입니다.

본 연구는 이러한 시스템 균형점 정량적으로 규명하는것을 목표로 합니다. 생성 파라미터를 고정하고 양자화 비트만 바꿔가며 성능을 측정한 뒤, 사람의 독서 속도 기준과 교차 분석해서 기기별로 최적 양자화 모델을 고르는 프레임워크를 제안합니다.



<그림 1. 논문에서 제안하는 온디바이스 LLM최적화 프레임워크 구조도>

II. 논문에서 제안하는 연구방법론

2.1 온디바이스 최적 양자화 프레임워크

이본 연구에서 제안하는 프레임워크는 모바일 단말기에서 소형 언어 모델(SLM)이 독립형 애플리케이션(Standalone Application) 형태로 서빙된다는 전제에 기반한다. 이에 따라 프레임워크는 그림 1과 같이 통제변수 설정, GGUF 양자화 분기 준비, 하드웨어 구동, 다중 지표 프로파일링의 4단계 파이프라인으로 구성된다.

첫째, 통제변수 설정 단계에서는 양자화 비트 수 변화의 순수한 영향만을 측정하기 위해 텍스트 생성에 관여하는 모든 하이퍼파라미터를 다음 과 같이 고정 상수로 설정한다. Context는 4,096 token, 온도(Temperature)는 0.7, Top-K는 40, Top-P는 0.9로 고정하였다. [4][5]의 논문을 참고하여 파라미터를 고정하였다. 온도는 모델이 다음 토큰을 선택할 때 확률 분포를 얼마나 날카롭게 또는 평탄하게 만들지를 조절하는 값으로, 낮을수록 결정적인 출력을, 높을수록 다양한 출력을 생성한다

Top-K와 Top-P는 각각 후보 토큰의 개수와 누적 확률을 기준으로 샘플링 범위를 제한하여 출력의 일관성을 확보한다. 이 값들을 상수로 고정함으로써

독립변수인 양자화 비트 수와 종속변수인 성능 지표 사이의 인과관계를 명확히 분리할 수 있다.

표 1. 온디바이스 LLM 실험 통제변수 고정값

Context size	4,096
Temperature	0.7
Top_K	40
Top_P	0.9

2.2 GGUF 양자화 모델별 분기 준비

두 번째 단계는 GGUF 양자화 분기 준비단계로 , 실험 대상 모델의 양자화 프로파일을 별도로 구축한다. (Gemma-3-1B, Gemma-3-4B, Qwen-0.8B, Qwen-2B, Qwen-4B)마다 Q3_K_M, Q4_K_M, Q5_K_M 세 가지 양자화를 별도로 변환해 준비합니다.[6][7][8][9][10] 즉, 하나의 베이스 모델에서 비트 수만 다른 세 개의 파일을 준비합니다. 각 분기는 독립적으로 로드되고 실험됩니다. 동동일 모델의 양자화 변형체를 개별 준비하는 목적은, 런타임 상의 동적 변환 오버헤드를 배제하고 각 양자화 수준이 실제 물리 파일 크기 및 순수 메모리 점유량에 미치는 영향을 정량적으로 계측하기 위함이다.

세 번째 단계는 실제 하드웨어에서의 구동입니다. 모바일 환경 고유의 발열 제약, 메모리 압박, 그리고 OS 스케줄러의 간섭 메커니즘을 온전히 재현하기 위해 에뮬레이터 환경을 배제하고 실기기에서 직접 추론을 수행한다. 실험 전 백그라운드 프로세스를 안정화한 유휴 상태를 베이스라인으로 잡고, 모델 로드 시점부터 추론 완료까지 전 과정을 기록합니다.

표 2. 실험 단말기기의 실험전 기초 상태

Total memory	11,903,212KIB	11.35GiB
Free Ram	8,099,185KIB	7.7 2GiB
USED	4,894,414KIB	4.67GiB

넷째, 다중 지표 프로파일링 단계에서는 단일 지표만으로는 온디바이스 LLM의 실사용 가능성을 판단할 수 없으므로, 아래 세 가지 종속변수를 동시에 측정한다.

TTFT(Time to First Token, ms)는 모델이 유휴 상태에서 프롬프트를 수신한 후 첫 번째 출력 토큰을 UI 스트림으로 반환할 때까지 걸리는 초기 지연 시간이다. 이는 사용자가 체감하는 반응 속도에 직접 대응한다.
Throughput(tokens/sec)은 첫 토큰 출력 시점부터 생성 완료 신호까지 초당 평균 생성 토큰 수를 나타낸다.

본 연구에서는 측정 일관성을 위해 영어 단어 단위를 기준 토큰으로 삼아 속도를 산출한다. Primativo 등은 RSVP 실험에서 마스킹이 읽기 속도를 크게 저하시하고, 단기기억 용량 초과와 의미적 맥락의 영향이 각각 속도를 제한하며, 특히 안구운동이 약 300 wpm 수준의 실질적 상한을 부과한다고 결론지었다.[10]
Memory Footprint(MB)는 모델이 메모리에 로드되는 순간의 순수 가중치 점유량(Model MB), 추론 중 발생하는 최대 커널 레벨 점유량(Peak MB), 그리고 가용 RAM 대비 실제 점유율(Usage %)의 세 항목으로 구성된다.

2.3 기능적 요구사항: 시스템 가용성

모바일 애플리케이션의 가장 기본적인 기능적 요구사항은 하드웨어 자원의 제약 속에서도 앱이 강제 종료 없이 안정적으로 실행 상태를 유지하는 것이다.
안드로이드 환경에서 온디바이스 LLM 추론이 시작되면 힙(Heap) 메모리 점유율이 급격히 상승하며, 이는 가비지 컬렉션(GC) 빈도를 증가시켜 시스템 오버헤드를 유발한다. 메모리 고갈이 커널 레벨에서 감지될 경우 LMK(Low Memory Killer)가 앱 프로세스를 강제 종료하여 시스템 가용성이 완전히 파괴된다. 따라서 GGUF 기반 가중치 양자화를 통한 메모리 풋프린트 통제는 단순한 최적화 수단이 아니라, 앱이 정상 작동하기 위한 필수 조건이다.

2-3 비기능적 요구사항: 인간 독서 인지 속도 기반 UX 성능

앱이 안정적으로 구동되더라도, 출력 속도가 사용자의 인지 흐름을 따라가지 못하면 서비스로서의 실용성을 잃는다.
Perceptual and cognitive factors imposing “speed limits” on reading rate: A study with the rapid serial visual

presentation.[11] 에 따르면, 인간이 텍스트를 읽고 이해할 수 있는 평균 독서 속도는 분당 최대 260 (WPM) 수준이다. 이를 온디바이스 LLM의 스트리밍 출력 환경에 대입하면, 최소 100 WPM 이상의 생성 속도가 유지되어야 대화의 인지적 흐름이 단절되지 않는다

즉, 양자화 비트를 낮추면 메모리 제약은 완화되지만 생성 속도가 임계선 아래로 떨어질 수 있고, 반대로 비트를 높이면 속도는 확보되지만 메모리 압박으로 앱이 종료될 위험이 생긴다. 본 프레임워크는 이 두 요구사항을 동시에 만족하는 균형점을 기기별로 정량적으로 규명하는 것을 목적으로 한다.

III. 실험

프롬프트 설계 측면에서는 모델 간 출력 조건의 균일성을 확보하기 위해 구조화 프롬프트(Structured Prompt)를 적용하였다. 출력 길이를 300자 이상 500자 이하로 제한하고, 번호·기호·소제목 사용을 금지하여 단일 단락 형태의 응답만을 생성하도록 명시하였다. 이를 통해 모델별 출력 토큰 수와 생성 패턴의 가변성을 통제하고, Throughput 및 TTFT 측정값의 비교 신뢰도를 확보하였다.

모델 체급 (Model Scale): 구조적 파라미터 크기가 모바일 CPU 연산 능력에 미치는 영향을 규명하고자 제품군의 유효 파라미터 분기인 1B(1 Billion) 및 2B(2Billion) 4B(4 Billion) 체급을 교차 실험한다.

종속 변수 (Dependent Variables)

TTFT (Time to First Token, ms): 모델이 대기 유휴(Idle) 상태에서 프롬프트 입력을 완전히 수용한 후, 첫 번째 출력 토큰을 외부 UI 스트림으로 반환할 때까지 걸린 초기 연산 지연 시간을 밀리초(ms) 단위로 정량 계측한다.

Throughput (Tokens/sec): 첫 토큰 출력 시점부터 생성 완료(EOS) 신호 도달 시점까지 초당 생성되는 평균 토큰 수를 정량적으로 산출한다. 본 논문에서는 통제 프로토콜의 일관성을 위해 영어 단어(Word) 세그먼트 단위를 벤치마크 표준 토큰 기준으로 책정하여 속도를 도출한다.

Memory Footprint (MB): 각 양자화 모델이 메모리에 온로드(On-load)되는 순간의 순수 모델 가중치 점유량(Model MB)과, 대화 추론 연산 중 활성화 오버헤드로 인해 발생하는 최대 커널 레벨 메모리 점유량(Peak MB) 및 가용 RAM 대비 실제 하드웨어 자원 점유율(Usage %)을 분석한다.

3, 실험 결과

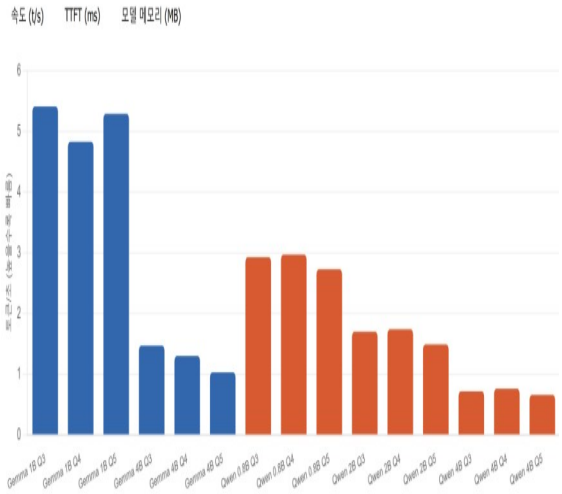


그림 2 [token /s]

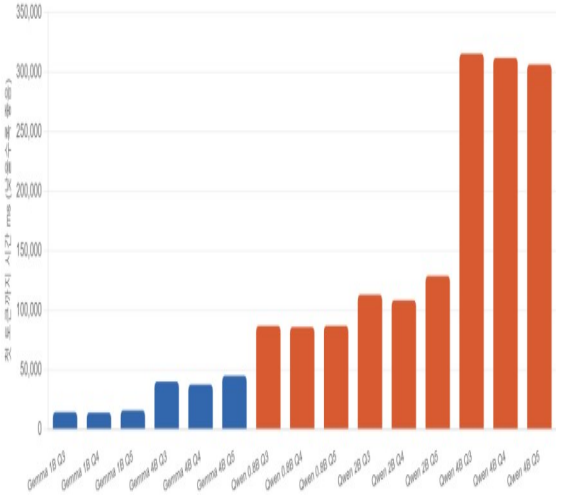


그림 3 [TTFT]

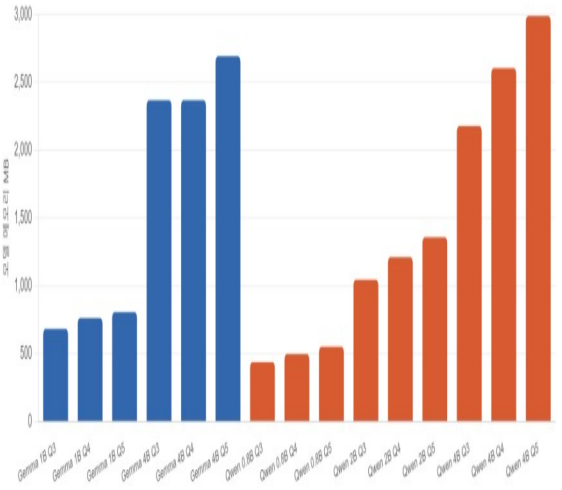


그림 4. 파라미터별 모델의 메모리 사용량

<표 3[전체 실험 통계]>

파일명	모델	파라미터	양자화	Speed (t/s)	TTFT (ms)	Total (s)	Model MB	Peak MB	Usage %
qwen3_5-0_8b_q3	Qwen3.5	0.8B	Q3_K_M	2.93	86941	103.00	438	706	6.1%
qwen3_5-0_8b_q4	Qwen3.5	0.8B	Q4_K_M	2.97	85915	109.50	497	763	6.6%
qwen3_5-0_8b_q5	Qwen3.5	0.8B	Q5_K_M	2.73	87043	179.74	552	819	7.0%
qwen3_5-2b_q3	Qwen3.5	2B	Q3_K_M	1.70	113127	217.57	1045	1284	11.0%
qwen3_5-2b_q4	Qwen3.5	2B	Q4_K_M	1.74	108410	179.28	1211	1446	12.4%
qwen3_5-2b_q5	Qwen3.5	2B	Q5_K_M	1.49	128943	208.75	1358	1593	13.7%
qwen3_5-4b_q3	Qwen3.5	4B	Q3_K_M	0.72	315493	530.67	2177	2477	21.3%
qwen3_5-4b_q4	Qwen3.5	4B	Q4_K_M	0.76	312033	568.41	2604	2902	25.0%
qwen3_5-4b_q5	Qwen3.5	4B	Q5_K_M	0.66	306499	549.71	2988	3285	28.3%
gemma3_1b_q3	Gemma3	1B	Q3_K_M	5.41	14415	41.03	683	1191	10.2%
gemma3_1b_q4	Gemma3	1B	Q4_K_M	4.83	14016	41.94	762	1191	10.2%
gemma3_1b_q5	Gemma3	1B	Q5_K_M	5.29	15886	33.48	806	1191	10.2%
gemma3_4b_q3	Gemma3	4B	Q3_K_M	1.47	40127	129.01	2368	3153	27.1%
gemma3_4b_q4	Gemma3	4B	Q4_K_M	1.30	37603	137.86	2368	3153	27.1%
gemma3_4b_q5	Gemma3	4B	Q5_K_M	1.03	44990	185.86	2692	3152	27.1%

IV. 실험 결과

본 실험에서는 보급형 모바일 단말기(Motorola G54) 환경에서 독서 인지 임계선인 100 WPM(Words Per Minute, 약 2.22 tokens/s) 충족 여부를 기준으로 실험 대상인 15개 모델 및 양자화 조합의 성능을 평가하였다.

4-1. 임계선 상회 모델 군 (Gemma-3-1B 및 Qwen3.5-0.8B) Gemma-3-1B는 전체 실험 대상 중 가장 높은 Throughput을 기록하며 압도적인 성능을 보였다. 양자화 레벨별로 Q3_K_M 사양에서 5.41 t/s, Q4_K_M에서 4.83 t/s, Q5_K_M에서 $5.29₩text{ t/s}$를 달성하였다. 이는 1 토큰을 0.75 단어로 환산할 경우 약 $217 ₩sim 243₩text{ WPM}$에 해당하여 실시간 대화 환경을 완벽히 충족하는 수치이다. 초기 지연 시간인 TTFT(Time-to-First-Token)는 140～158 ms 수준으로 매우 신속한 반응성을 보였으며, 피크 메모리(Peak Memory)는 전 양자화 구간에서 1,191MB로 동일하게 계측되었다.

Qwen3.5-0.8B는 Q3_K_M 설정에서 2.93 t/s, Q4_K_M에서 2.97 t/s, Q5_K_M에서 2.73token/s 를 기록하며 독서 인지 임계선(2.22 t/s)을 소폭 상회하였다. 다만, Gemma-3-1B 대조군과 비교했을 때 Throughput 효율은 약 절반 수준에 머물렀다. 대신 피크 메모리는 양자화 수준에 따라 706～819 MB 분포를 보여 Gemma-3-1B 대비 자원 소모량 측면에서 이점을 나타냈다.

4-2. 임계선 하회 모델 군 (2B 및 4B 체급 계열)

2B 이상의 고체급 모델들은 모바일 단말의 하드웨어 제약으로 인해 실시간 대화형 서비스에 부적합한 수준의 처리량을 보였다. Qwen3.5-2B는 Q4_K_M 기준 1.74 token/s를 기록하여 임계선 달성에 실패하였다. 4B 체급의 경우 성능 저하가 더욱 심화되었는데, Qwen3.5-4B는 전 양자화 구간에서 0.66 ~0.76token /s}의 극히 저조한 처리량을 나타냈으며, Gemma-3-4B 역시 Q3_K_M 1.47 token/s, Q5_K_M 1.03token /s 로 실효적인 대화 인터페이스를 구축하기 어려운 것으로 분석되었다.

초기 지연 시간(TTFT) 관점에서는 동급 4B 체급 내에서 확연한 격차가 관찰되었다. Gemma-3-4B의 TTFT는 약 37～44초(40,127～44,990 ms) 범주로 측정되어, 동급 Qwen3.5-4B의 TTFT인 약 306～315초 (306,499～315,493 ms) 대비 현저히 짧은 초기 반응 지연을 기록하였다. 이러한 프리필 단계의 지연 격차는 누적 연산 시간으로 이어져, 최종 작업 완료 시간(Total Time) 측면에서도 Gemma-3-4B(129～185초)가 Qwen3.5-4B(530～569초)보다 최소 3배 이상 빠른 연산 종결 능력을 입증하였다.

V. 고찰

5-1. 모델 아키텍처 특성이 온디바이스 추론 성능에 미치는 영향 본 실험에서 가장 주목할 만한 결과는 Gemma-3-1B가 Qwen3.5-0.8B보다 더 큰 파라미터 규모를 가짐에도 불구하고, 모든 양자화 구간에서 더 높은 토큰 생성 속도(Speed)를 기록했다는 점이다. Gemma-3-1B는 4.83~5.41 tokens/s의 처리량을 나타낸 반면, Qwen3.5-0.8B는 2.73~2.97 tokens/s 수준에 머물렀다. 일반적으로 파라미터 수가 증가할수록 연산량 또한 증가하므로 추론 속도가 감소하는 경향이 있으나, 본 실험 결과는 모델 규모만으로 온디바이스 추론 성능을 설명하기 어렵다는 사실을 보여준다. 동일한 모바일 환경에서는 파라미터 수뿐 아니라 모델 구조, 레이어 구성, 연산 그래프 최적화 수준, 런타임 구현 방식 등이 실제 처리량에 중요한 영향을 미칠 수 있다. 특히 Gemma-3-1B는 모든 양자화 수준에서 낮은 TTFT(Time To First Token)와 높은 토큰 생성 속도를 동시에 유지하였다. 이는 해당 모델이 모바일 CPU 환경에서 효율적으로 동작하도록 최적화되어 있거나, 추론 엔진이 모델 구조를 효과적으로 활용하고 있음을 시사한다.반면

Qwen3.5-0.8B는 상대적으로 작은 규모에도 불구하고 처리량이 낮게 측정되었다. 이는 모델 규모 외의 구조적 특성이 실제 추론 성능에 영향을 미칠 수 있음을 보여주며, 온디바이스 환경에서는 단순한 파라미터 수보다 모델 아키텍처와 런타임 최적화 수준이 더욱 중요한 요소가 될 수 있음을 시사한다.

5-2. 양자화 수준과 추론 성능 간의 비선형적 관계실험 결과, 양자화 비트 수가 감소할수록 추론 속도가 일관되게 향상되는 경향은 관찰되지 않았다. 오히려 일부 모델에서는 Q4_K_M 설정이 가장 높은 처리량을 기록하였다.Qwen3.5 계열의 경우 0.8B, 2B, 4B 모델 모두 Q4_K_M 환경에서 각각 2.97 tokens/s, 1.74 tokens/s, 0.76 tokens/s로 최고 성능을 보였다. 이는 모델 크기와 무관하게 동일한 경향이 반복되었다는 점에서 의미가 있다.일반적으로 양자화 비트 수가 낮아지면 모델 크기가 감소하여 메모리 대역폭 사용량이 줄어들지만, 동시에 가중치 복원 과정에서 추가적인 계산 비용이 발생한다. 따라서 실제 추론 성능은 메모리 접근 비용 감소 효과와 역양자화 연산 비용 증가 효과 사이의 균형에 의해 결정된다.본 실험 결과는 Q3_K_M과 같은 더 강한 압축 방식이 항상 높은 처리량을 보장하지 않음을 보여준다. 오히려 특정 하드웨어 및 추론 엔진 조합에서는 Q4_K_M이 메모리 사용량과 연산 효율 사이에서 가장 균형 잡힌 성능 지점을 형성할 가능성이 있다.이는 온디바이스 LLM 환경에서 양자화 수준을 결정할 때 단순히 모델 크기 감소만을 고려하기보다 실제 추론 처리량을 함께 평가해야 함을 시사한다.

5-3. 모델 규모 증가에 따른 지연시간 증가와 메모리 요구량 변화

본 실험에서는 모델 규모가 증가할수록 토큰 생성 속도는 감소하고 TTFT는 증가하는 전형적인 경향이 관찰되었다.

Qwen3.5 계열을 기준으로 살펴보면 0.8B 모델은 약 2.9 tokens/s 수준의 처리량을 보인 반면, 4B 모델은 약 0.7 tokens/s 수준으로 감소하였다. Gemma 계열 역시 1B 모델이 약 5 tokens/s 수준을 유지한 반면, 4B 모델은 약 1 tokens/s 수준으로 감소하였다.이러한 결과는 파라미터 수 증가에 따라 연산량과 메모리 접근량이 함께 증가하기 때문으로 해석할 수 있다. 특히 모바일 기기와 같이 연산 자원 및 메모리 대역폭이 제한된 환경에서는 대규모 모델의 성능 저하가 더욱 두드러지게 나타난다.메모리 사용량 측면에서도 모델 규모 증가에 따른 상승이 확인되었다. Qwen3.5 계열은 0.8B 모델에서 약 700~800MB 수준의 피크 메모리를 사용한 반면, 4B 모델에서는 약 3GB 수준까지 증가하였다. Gemma 계열 역시 1B 모델 대비 4B 모델에서 약 2.5배 이상의 메모리 사용량 증가가 관찰되었다.특히 Gemma 계열에서는 동일 모델 내에서 양자화 수준이 달라져도 피크 메모리 사용량이 거의 일정하게 유지되는 특성이 확인되었다. 이는 모델 가중치 외에도 추론 과정에서 사용되는 버퍼, 캐시 및 런타임 메모리 구조가 전체 메모리 사용량의 상당 부분을 차지하고 있음을 시사한다.결과적으로 온디바이스 환경에서는 모델 규모 증가가 추론 품질 향상에 기여할 수 있으나, 동시에 응답 지연과 메모리 사용량 증가를 초래한다. 따라서 실제 서비스 환경에서는 모델 성능뿐 아니라 응답 속도와 자원 사용량을 함께 고려한 균형 있는 모델 선택이 필요할 것으로 판단된다.

VI. 결론

본 연구는 보급형 모바일 단말기(Motorola G54) 환경에서 소형 언어 모델(SLM)의 파라미터 규모 및 양자화 수준 변화가 추론 속도, 초기 응답 지연, 메모리 사용량에 미치는 영향을 실증적으로 분석하고, 인간의 독서 인지 임계선(100 WPM, 약 2.22 tokens/s)을 기준으로 실시간 대화형 서비스에 적합한 모델 선정 기준을 제시하였다.

첫째, 모바일 온디바이스 환경에서 사용자의 인지적 흐름을 단절시키지 않는 실시간 대화형 서비스를 구현하기 위한 실효적 파라미터 상한선은 1.5B 이하로 나타났다. 실험에 포함된 2B 이상 체급의 모델은 양자화 수준과 무관하게 100 WPM 임계선을 충족하지 못하였으며, 이는 현재 보급형 모바일 하드웨어의 연산 및 메모리 대역폭 제약 하에서 고체급 모델을 그대로 엣지에 배치하는 접근이 실용적이지 않음을 시사한다.

둘째, 동일한 모바일 환경에서 추론 성능을 결정하는 요인은 파라미터 수만으로 설명되지 않았다. Gemma-3-1B는 Qwen3.5-0.8B보다 더 큰 파라미터 규모를 가짐에도 전 양자화 구간에서 약 2배 높은 처리량(4.83~5.41 t/s vs 2.73~2.97 t/s)을 기록하였으며, TTFT 또한 140~158ms 수준으로 가장 낮게 측정되었다. 이러한 결과는 온디바이스 환경에서의 추론 성능이 모델 구조, 레이어 구성, 연산

그래프 최적화, 런타임 구현 수준 등 파라미터 수 외의 요인에 의해 상당 부분 좌우될 수 있음을 보여준다. 이상의 실험 결과를 종합할 때, 본 연구가 제시하는 1.5B 이하 체급 내 최적 모델은 Gemma-3-1B이며, 해당 모델은 200WPM 이상의 안정적인 스트리밍 출력을 통해 실시간 대화에 요구되는 비기능적 요구사항을 충족하는 것으로 판단된다.

셋째, 양자화 비트 수와 추론 속도 간에는 일관된 단조적 관계가 관찰되지 않았다. Qwen3.5 계열에서는 0.8B, 2B, 4B 전 체급에서 Q4_K_M이 가장 높은 처리량을 기록하여, 양자화 비트를 낮추는 것이 항상 속도 향상으로 이어지지는 않음을 보여주었다. 이는 메모리 접근량 감소 효과와 역양자화 연산 비용 증가 효과 사이의 균형 지점이 양자화 수준별로 달라질 수 있음을 시사하며, 온디바이스 모델 배포 시에는 압축률뿐 아니라 실제 처리량을 함께 측정하여 양자화 수준을 결정해야 함을 보여준다. 다만 본 실험에서 Gemma-3-1B의 경우 Q3_K_M과 Q5_K_M이 Q4_K_M보다 다소 높은 처리량을 보였는데, 이러한 모델별 차이의 원인은 본 연구의 범위에서 명확히 규명되지 않았으며 추후 추론 엔진의 커널 구현 수준에서의 추가 검증이 필요하다.

넷째, 모델 규모가 증가할수록 처리량 감소, TTFT 증가, 메모리 사용량 증가라는 전형적인 경향이 확인되었다. Qwen3.5 계열은 0.8B에서 4B로 갈수록 처리량이 약 2.9 t/s에서 0.7 t/s 수준으로, 피크 메모리는 약 700~800MB에서 약 3GB 수준으로 증가하였으며, Gemma 계열 역시 1B 대비 4B에서 처리량 감소와 약 2.5배 이상의 메모리 증가가 관찰되었다. 한편 Gemma 계열은 동일 모델 내에서 양자화 수준이 달라져도 피크 메모리가 거의 일정하게 유지되는 특성을 보였는데(예: Gemma-3-1B 전 양자화 구간 1,191MB), 이는 모델 가중치 외에 런타임 버퍼, 캐시 등 고정 오버헤드가 전체 메모리 사용량의 상당 부분을 차지할 가능성을 시사한다. 아울러 4B 체급 내에서도 Gemma-3-4B의 TTFT(37~44초)가 Qwen3.5-4B(306~315초)보다 현저히 짧게 측정되었는데, 이는 둘째 항목에서 논의한 아키텍처 및 런타임 최적화 수준의 차이가 모델 규모와 무관하게 일관되게 작용할 수 있음을 보여주는 추가적인 사례로 해석할 수 있다.

다만 본 연구는 단일 보급형 단말기(Motorola G54)를 대상으로 수행되었으며, 각 모델·양자화 조합에 대한 측정이 단발성으로 이루어졌다는 한계를 가진다. 따라서 본 결과가 다른 디바이스 스펙이나 반복 측정 조건에서도 동일하게 재현되는지에 대해서는 추가적인 검증이 필요하다.

결론적으로, 본 연구의 실증 결과는 온디바이스 LLM 배포 시 파라미터 규모의 단순한 확대나 압축률 위주의 양자화 선택보다, 실제 하드웨어 환경에서의 처리량·지연·메모리를 종합적으로 고려한 아키텍처 인지형(architecture-aware) 모델 선정이 필요함을 보여준다. 특히 1.5B 이하 체급에서는 파라미터 수가 아닌 모델 구조와 런타임 최적화 수준이 사용자 체감 성능을 좌우하는 핵심 요인으로 작용할 수 있으며, 이러한 관점은 향후 모바일 환경에 특화된 초경량 모델 설계 및 추론 엔진 최적화 연구에 있어 중요한 시사점을 제공한다.

REFERENCE

[1]Dhaouadi M, Oakes BJ, Famelis M. Rationale dataset and analysis for the commit messages of the Linux Kernel Out-of-Memory Killer. Proceedings of the International Conference on Mining Software Repositories (MSR). Université de Montréal & Polytechnique Montréal. 2022.

[2]Sparrenberg L, Deußer T, Berger A, Sifa R. Small and fast LLMs on commodity hardware: Post-training quantization in llama.cpp. Proceedings of the International Conference on Machine Learning (ICML) Workshop on Efficient Natural Language Processing. 2023.

[3] Nielsen J. Response times: The 3 important limits. Usability Engineering. 1993 Jan 1; Chapter 5. Summary available at Nielsen Norman Group.

[4] Holtzman A, Buys J, Du L, Forbes M, Choi Y. The curious case of neural text degeneration. *Proceedings of the International Conference on Learning Representations (ICLR)*. 2020.

[5]Noarov G, Mallick S, Wang T, Joshi S, Sun Y, Xie Y, Yu M, Dobriban E. Foundations of Top-k Decoding for Language Models. *Proceedings of the International Conference on Learning Representations (ICLR)*. 2024.

[6]Hugging Face. MaziyarPanahi/gemma-3-4b-it-GGUF. Available from: https://huggingface.co/MaziyarPanahi/gemma-3-4b-it-GGUF

[7]Hugging Face. unsloth/gemma-3-1b-it-GGUF. Available from: https://huggingface.co/unsloth/gemma-3-1b-it-GGUF

[8]Hugging Face. unsloth/Qwen3.5-2B-GGUF. Available from: https://huggingface.co/unsloth/Qwen3.5-2B-GGUF [Accessed 2026 Jun 7].

[9]Hugging Face. bartowski/Qwen_Qwen3.5-0.8B-GGUF. Available from: https://huggingface.co/bartowski/Qwen_Qwen3.5-

0.8B-GGUF [Accessed 2026 Jun 7].

[10]Hugging Face. bartowski/Qwen_Qwen3.5-0.8B-GGUF. Available from: https://huggingface.co/bartowski/Qwen_Qwen3.5-0.8B-GGUF [Accessed 2026 Jun 7].

[11]Primativo S, Spinelli D, Zoccolotti P, De Luca M, Martelli M. Perceptual and cognitive factors imposing “speed limits” on reading rate: A study with the rapid serial visual presentation. *PLOS ONE*. 2016 Apr 18;11(4):e0153786. Editor: Kevin Paterson.

본논문은 한신대학교 졸업논문에 투고된 논문을 기반으로 작성된논문입니다.

사회 취약계층 건강 및 복지 서비스 효과성 평가

배윤진, 김유신
한신대학교 빅데이터융합학과
e-mail: y00nb@hs.ac.kr , yoosin25@hs.ac.kr

Evaluation of the effectiveness of health and welfare services for socially vulnerable groups

Yoonjin Bae, Yoosin Kim
Dept. of Big Data Convergence, Hanshin University

요 약

본 연구는 사회 취약계층을 대상으로 제공되는 건강 및 복지 서비스의 효과성을 분석하고, 이를 바탕으로 정책적 개선 방향을 제시하는 데 목적이 있다. 사회 취약계층은 경제적, 사회적 제약으로 인해 건강관리와 복지서비스 이용에 어려움을 겪고 있으며, 이에 따라 정부와 지방자치단체는 다양한 건강 및 복지 지원 정책을 시행하고 있다. 본 연구에서는 관련 선행연구를 검토하여 건강 및 복지 서비스가 취약계층의 건강 수준 향상, 의료 접근성 확대, 사회적 고립 감소 및 삶의 질 향상에 미치는 영향을 분석했다. 분석결과, 건강관리 서비스는 만성질환 예방과 건강 증진에 긍정적인 효과를 나타냈으며, 복지서비스는 정서적 안정과 사회 참여 확대에 기여하는 것으로 확인되었다. 반면, 서비스 간 연계 부족, 지역 간 지원 격차, 전문 인력 부족 등의 한계도 존재하는 것으로 나타났다. 따라서 사회 취약계층의 건강권 보장과 복지 증진을 위해서는 보건, 복지 통합서비스 체계 구축, 지역 간 서비스 균형 강화, 지속적인 성과평가 체계 마련이 필요하다. 본 연구는 향후 취약계층 지원정책의 효과성을 높이기 위한 기초자료를 제공한다.

키워드: 사회 취약계층, 건강서비스, 복지서비스, 효과성 평가, 의료 접근성, 삶의 질

I. 서론

현대 사회에서 사회 취약계층은 노인, 장애인, 저소득층, 한부모 가정, 독거노인 등을 포함하며 경제적·사회적 제약으로 인해 건강관리와 복지서비스 이용에 어려움을 겪고 있다. 특히 건강 불평등은 사회적 불평등과 밀접하게 연관되어 있으며, 취약계층의 건강 수준은 일반

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

인구에 비해 낮은 경향을 보인다.

정부는 이러한 문제를 해결하기 위해 방문건강관리사업, 지역사회 통합돌봄, 기초생활보장제도, 노인맞춤돌봄서비스 등 다양한 정책을 시행학고 있다. 그러나 실제 서비스가 대상자의 건강 개선과 복지 향상에 어느 정도 효과를 보이는지에 대한 지속적인 평가가 필요하다. 따라서 본 연구는 사회 취약계층 건강 및 복지 서비스의 효과성을 분석하고 정책적 시사점을 도출하고자 한다.

II. 이론적 배경

2.1 사회 취약계층의 개념

사회 취약계층은 경제적 빈곤, 신체적 장애, 고령화, 사회적 배제 등의 요인으로 인해 사회적 지원이 필요한 집단을 의미한다. 이들은 의료서비스 이용과 건강관리 능력이 상대적으로 낮으며, 사회적 자원의 접근성 또한 제한적이다.

2.2 건강 및 복지 서비스의 역할

건강 및 복지 서비스는 취약계층의 건강 증진과 사회적 안정망 구축을 목표로 한다. 건강서비스는 질병 예방, 건강관리, 의료 접근성 향상을 제공하며, 복지서비스는 생활지원, 심리 상담, 사회 참여 기회 확대 등을 통해 삶의 질을 향상시키는 역할을 수행한다.

III. 건강 및 복지 서비스 효과성 평가

3.1 건강 측면의 효과

건강서비스는 만성질환 관리와 예방 중심의 건강관리 체계를 통해 취약계층의 건강 수준 향상에 기여하고 있다. 방문건강관리서비스와 스마트 헬스케어 프로그램은 혈압과 혈당 관리 개선과 건강행태 변화에 긍정적인 영향을 미치는 것으로 보고되었다. 또한 의료기관 접근성이 향상되면서 조기 진단과 지속적인 치료가 가능해졌다. 이러한 결과는 건강 불평등 완화에 기여하는 중요한 성과로 평가된다.

3.2 복지 측면의 효과

복지서비스는 경제적 지원뿐만 아니라 정서적 안정과 사회적 관계 형성에도 긍정적인 영향을 미친다. 특히 독거노인과 저소득층을 대상으로 한 돌봄서비스는 우울감 감소와 사회적 고립 완화 효과를 나타내고 있다. 또한 지역사회 중심의 복지 프로그램은 취약계층의 사회참여를 확대하고 자립 역량을 향상시키는 데 기여하고 있다.

3.3 한계점

그러나 건강 및 복지 서비스는 몇 가지 한계를 가지고 있다. 첫째, 보건과 복지 서비스가 분절적으로 운영되어 통합적인 지원이 어려운 경우가 많다. 둘째, 지역 간 재정 및 인프라 차이로 인한 서비스 수준의 격차가 발생한다. 셋째, 전문 인력 부족과 서비스 전달체계의 비효율성으로 인해 지원 효과가 제한될 수 있다. 이러한 문제는 서비스 만족도와 지속 가능성에 부정적인 영향을 미친다.

IV. 결론

본 연구는 사회 취약계층을 대상으로 한 건강 및 복지 서비스의 효과성을 검토했다. 분석 결과, 건강관리 서비스는 의료 접근성 향상과 건강 상태 개선에 기여하였으며, 복지 서비스는 사회적 고립 감소와 삶의 질 향상에 긍정적인 효과를 보였다. 그러나 서비스 간 연계 부족과 지역별 격차 등의 한계도 존재했다.

따라서 향후 정책 방향으로는 첫때, 보건과 복지를 통합한 서비스 전달체계를 구축해야 한다. 둘째, 지역 간 서비스 격차를 줄이기 위한 재정 지원 확대가 필요하다. 셋째, 데이터 기반의 성과평가 체계를 마련하여 서비스 효과를 지속적으로 점검해야 한다. 이를 통해 사회 취약계층의 건강권 보장과 복지 향상을 더욱 효과적으로 실현할 수 있을 것이다.

참고문헌

[1] 김원진. (2012). 보건 및 복지 연계서비스 모형개발을 통한 지역사회의 방문보건사업의 활성화 방안. 한

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

국행정학보, 26(2), 245-260.
[2] 강은진, 이윤진. (2019). 영유아 양육 및 보건의료 서비스 통합을 위한 탐색적 연구- 국내 사례 비교 분석을 중심으로. 사회복지정책, 46(3), 5-32.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

히트곡 멜로디의 전개 및 화성 진행 패턴 데이터베이스 구축과 AI 작곡 보조 기법

허진명, 김유신
한신대학교 첨단융합계열
e-mail: molykorea3214@gmail.com, yoosin@hs.ac.kr

Development of hit song melody and harmony progress pattern database and AI composition assistance technique

Author*, Jinmyeong Her, Yoosin Kim
Division of Advanced Convergence, Hanshin University

요 약

본 논문은 대중에게 친숙함과 심리적 안정감을 주는 히트곡 고유의 화성 진행 및 멜로디 패턴을 시각적 데이터로 모델링하고, 이를 기반으로 작곡가를 위한 인터렉티브 추천 시스템을 구상하는 것을 목표로 한다. 이를 위해 국내외 주요 음원차트(Melon, Billboard 등)에서 상위권을 기록한 곡들의 MIDI 파일 및 화성 진행 데이터와 더불어, 해당 곡들의 장르, 발매 연도, 최고 순위 등의 메타데이터를 수집하여 분석에 활용한다.분석의 핵심 방법으로는 곡의 조(Key)와 관계없이 보편적인 비교가 가능하도록 화성을 로마자나 숫자로 변환하는 '화성 진행 인코딩'을 먼저 수행합니다. 이후 마르코프 연쇄(Markov Chain) 분석을 통해 대중음악에서 특정 코드 진행이 일어날 확률적 빈도를 파악하고, RNN 및 LSTM 모델을 활용하여 멜로디의 선율적 음정 변화 패턴을 정밀하게 학습시킨다.

키워드: RNN(Recurrent Neural Network), LSTM(Long Short - Term Memory), Markov Chain, MIDI, 작곡보조AI

1.서론

현대 대중음악 산업은 스트리밍 플랫폼의 활성화와 글로벌화로 인해 그 어느때보다 새로운 곡들의 소비 주기가 빨라졌으며, 이에 따라 창작자들에게 요구되는 곡 작업의 속도와 효율성 또한 극대화되고 있다. 대중음악, 특히 이른바 '히트곡'이라고 불리는

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

상위 차트 진입 곡들은 대중에게 심리적 안정감과 귀에 익은 친숙함을 제공하는 특정 화성적 규칙과 멜로디 패턴을 공유하는 경향이 있다. 최근 생성형 인공지능 기술이 급격히 발전함에 따라 데이터 기반의 자동 작곡 시스템 연구가 활발히 진행되고 있으나, 현재의 완전 자동 작곡 모델은 인간 작곡가의 고유한 예술적 의도나 미세한 감성을 온전히 반영하기 어렵다는 한계를 지닌다. 따라서 AI에게 창작의 모든 부분을 맡기는 것이 아닌, 창작의 초기단계에서 인간 작곡가의 창의성을 보조하고 작업 시간을 단축할 수 있는 효율적인 기술적 도구의 개발이 요구된다고 볼 수 있다. 본 연구의 목적은 대중에게 귀에 익고 안정감을 주는 국내외 히트곡의 화성 방법 및 멜로디 진행 패턴을 시각적 데이터로 모델링하고, 이를 바탕으로 인간 작곡가와 협업할 수 있는 음악 진행 추천 시스템을 설계하는 것이다. 이를 위해 국내외 음원 차트(Melon, Billboard 등)의 상위권 곡들을 대상으로 음악 구조 데이터 및 메타 데이터를 수집한다. 수집된 데이터는 조성(Key)에 구애받지 않도록 인코딩 과정을 거친 뒤, 마르코프 연쇄 및 RNN/LSTM 신경망 모델을 통해 분석된다. 최종적으로 이 분석 결과를 시각화 하고, 작곡가가 입력한 초기 음악적 모티프에 대응하여 최적의 다음 진행을 제안하는 프로토타입 시스템의 프레임워크를 정립하고자 한다.

2. 관련 연구

빌보드 차트 수백곡의 화성 진행을 디지털 데이터셋으로 정량화한 맥길 대학교의 '빌보드 프로젝트(McGill Billboard Project)'와 대중음악의 코드 전이 확률을 시각화 하여 제공하는 '훅이오리(Hooktheory)' 데이터 분석이 방법론적 기초가 된다. 또한, 순차적 데이터 학습을 통해 자연스러운 선율 패턴을 생성하는 구글 마젠타의 '멜로디 RNN'과 사용자의 멜로디 수정에 맞춰 실시간으로 화성을 채워주는 '딥바흐'연구는 멜로디 및 화성 학습 엔진의 기술적 토대를 제공한다. 마지막으로 사용자가 설정한 스타일을 바탕으로 AI가 화성과 멜로디 아이디어를 제안하고 작곡가가 이를 실시간으로 수정, 조합할 수 있도록 돕는 소니 CSL의 '플로우컴포저'는 본 연구가 지향하는 '작곡가를 위한 대화형 추천 시스템'의 인간-컴퓨터 상호작용(HCI) 관점을 가장 잘 뒷받침하는 선행사례이다.

3. 제안방법

본 연구에서 제안하는 히트곡 패턴 분석기반의 작곡가용 인터랙티브 추천 시스템은 대중음악 고유의 음악적 특징을 정량화하고 이를 시각적 데이터로 모델링하여 창작자와 상호작용하는 것을 목표로 한다. 본 시스템의 전체적인 연구 프로세스는 데이터 수집 및 정규화, 화성 및 선율적 패턴의 복합 모델링, 그리고 이를 활용한 사용자 인터랙티브 추천 알고리즘 설계의 3단계로 구성된다.

3.1.

본 연구의 신뢰성을 확보하기 위해 국내외 주요 음원차트에서 상위권을 기록한 히트곡의 메타 데이터와 MIDI 및 화성 진행 데이터를 수집한다. 수집된 이종 데이터의 보편적 비교 분석을 위해 화성진행 인코딩과 시퀀스 추출이라는 전처리 과정을 수행한다.

3.2

대중음악이 주는 친숙함과 심리적 안정감의 정량적 근거를 구축하기 위해, 화성의 통계적 흐름과 멜로디의 시간적 맥락을 동시에 학습하는 하이브리드 모델링인 마르코프 연쇄 분석과, RNN 및 LSTM 기반 선율 학습을

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

제안한다.

3.3

학습된 데이터 모델을 기반으로 작곡가에게 시각적 영감을 제공하고 대화형으로 소통할 수 있는 추천 시스템을 구축한다. 시각적 데이터 모델링과 인터랙티브 추천 알고리즘이 그 종류이다.

4. 결론

본 연구에서는 대중에게 친숙함과 심리적 안정감을 주는 히트곡 고유의 음악적 특징을 정량화하고, 이를 기반으로 창작자를 보조하는 AI 기반 인터랙티브 및 MIDI, 화성 데이터를 활용하여 대중음악의 흥행 공식을 데이터 모델로 성공적으로 구축하였다. 분석 과정에서 곡의 조성에 영향을 받지 않는 보편적 비교를 통해 히트곡 시퀀스 내 특정 코드 전이 확률의 통계적 빈도를 명확히 파악하였다. 이에 더해 RNN 및 LSTM 모델을 적용하여 선율의 시계열적 문맥과 장기 의존성을 정밀하게 학습시킴으로써, 대중적 친숙도를 반영한 자연스러운 멜로디 흐름을 예측 및 생성할 수 있음을 확인하였다. 본 연구의 핵심적 성과는 이와 같은 복합 신경망 학습 모델을 단순한 자동 작곡 엔진에 머무르게 하지 않고, '시각적 데이터 모델링'과 '인터랙티브 추천 인터페이스'로 확장한다는 점이 있다. 작곡가가 시스템과 실시간으로 상호작용하며 화성과 멜로디의 선율적 변화를 시각적으로 인지하고 선택할 수 있도록 설계함으로써, 인공지능이 인간의 창의성을 대체하는 것이 아닌 상호 보완적인 파트너로서 창작 프로세스의 효율성을 극대화할 수 있는 가능성을 제시하였다. 하지만 본 연구는 주로 화성과 멜로디의 음정 패턴에 집중되어 있으나, 실제 대중음악의 흥행과 심리적 안정감에는 편곡, 음색, 가사의 감성적 키워드 등 다양한 요소가 복합적으로 작용한다. 향후 연구에서는 오디오 신호 처리를 통해 음색 데이터를 추출하고 NLP 기반의 가사 분석을 통합하는 다중 모델 학습 모델로의 고도화가 필요하다는 한계점과, 제안된 인터랙티브 시스템이 실제 작곡가의 창작 프로세스에 미치는 긍정적 영향과 사용성을 검증하기 위해, 프로 작곡가 및 아마추어 창작자 집단을 대상으로 한 FGI(포커스 그룹 인터뷰) 및 통계적 UX 평가를 후속 연구로 진행하여 시스템의 실효성을 보완해야 한다는 점을 지니고 있다.

참고문헌

[1] 임세혁, (2023) K-POP 표현 기법의 분석 연구 : 빌보드 200차트를 중심으로
[2] 이상엽, (2024) 음향적, 가사적 특성을 이용한 음악 흥행 예측 연구 : 장르와 시대별 차이를 중심으로
[3] 하재용, (2025) 인공지능 활용 음악 창작자의 역량 모형 개발 연구

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기 RISE사업)의 연구결과입니다.

시계열 인과발견 결과의 실무적 해석을 위한 LPCMCI 기반 동적 요인 기여 프레임워크

김현우[O], 최서헌[*], 김성은[**]
[O, *]한동대학교 AI융합학부
[**]한동대학교 AI컴퓨터전자공학부
e-mail: {rejames01[O], csheon43[*], 2200093[**]}@handong.ac.kr

A Dynamic Driver Contribution Framework Based on LPCMCI for Practical Interpretation of Time-Series Causal Discovery Results

Hyeonwoo Kim[O], Seoheon Choi[*], Seoungeun Kim[**]
[O, *]School of Applied AI and Entrepreneurship, Handong Global University
[**]School of Artificial Intelligence, Computer and Electrical Engineering, Handong Global University

요 약

가격 변동 요인의 식별은 단순한 가격 예측을 넘어 실무적 의사결정의 근거를 제공한다. 이러한 원인 해석을 위해 시계열 자료에서 변수 간 지연 관계를 탐색하는 인과발견 방법이 활용되고 있지만, 기존 인과발견 알고리즘에서 산출되는 조건부 독립성 검정 통계값(val)은 조건부 집합 내에서 변수 간 관계가 얼마나 뚜렷하게 나타나는지를 보여주는 지표에 가깝다. 따라서 특정 시점에 실제로 발생한 원인 변수의 변화 방향과 크기를 함께 반영하지 못하며, 이를 가격 상승·하락에 대한 요인별 기여로 직접 해석하기에는 한계가 있다. 본 연구는 이러한 해석 간극을 줄이기 위해 PCMCI 계열의 확장 방법인 LPCMCI 와 슬라이딩 윈도우 분석을 결합한 동적 요인 기여 프레임워크를 제안한다. 구체적으로 val에 해당 시차의 원인 변수 표준화 값을 결합하여 동적 요인 기여 점수를 산출하고, 이를 통해 가격 변동 국면별 주요 기여 요인을 해석한다. 제안 프레임워크를 LME 구리 현물가격의 월별 변동률에 적용한 결과, 동적 요인 기여 점수는 기존 val 기반 점수보다 가격 변동 방향과의 정합성이 높게 나타났다. 또한 수요, 공급, 재고, 거시 요인의 월별 기여를 분해함으로써 가격 변동 국면별 주요 기여 요인을 식별할 수 있음을 확인하였다.

키워드: 시계열 인과발견, LPCMCI, 구리 가격, 요인 해석, 동적 요인 기여 점수

I. 서론

원자재 가격은 수요, 공급, 재고, 거시 및 지정학적 요인의 영향을 복합적으로 받으며, 이러한 영향은 즉각적으로 반영되기도 하지만 일정한 시차를 두고 나타난다[1][2]. 이처럼 가격 변동이 여러 요인의 시차적 작용을 통해 형성되기 때문에, 원자재 시장의 실무적 의사결정에서는 가격 변동을 예측하는 것뿐 아니라 해당 변동이 어떤 시장 요인의 변화와 연결되어 나타났는지를 식별하고 해석하는 과정이 필요하다[2]. 본 연구는 이러한 가격 변동 해석 문제의 적용 사례로 LME 구리 현물가격을 분석한다. 구리는 제조업, 건설, 전력망, 배선 등 다양한 산업과 연결된 대표적인 산업 금속이다[3]. 최근에는 AI 확산에 따른 데이터센터 증설과 전력망·배선 인프라 확대 속에서 중요성이 커지고 있으며, 지정학적 갈등, 공급망 불안 등이 겹치면서 가격 변동성도 확대되고 있다[4].

구리를 포함한 원자재 가격 변동을 해석하기 위해서는 변수 간 동시적 관계뿐 아니라 시차적 관계를 함께 고려할 필요가 있다[1]. 이를 위해 본 연구는 시계열 자료에서 변수 간 지연 관계를 탐색하는 인과발견 방법을 활용한다. 다만 기존 인과발견 알고리즘이 산출하는 조건부 독립성 검정 통계값(이하 val)은 특정 시점에서 원인 변수의 실제 변화 방향과 크기를 직접 반영하지 못하므로, 가격 상승 또는 하락에 대한 요인별 기여로 해석하는 데 한계가 있다[5]. 본 연구는 이러한 한계를 보완하기 위해 LPCMCI의 지연 간선 통계값과 원인 변수의 표준화 값을 결합한 동적 요인 기여 프레임워크를 제안한다.

본 연구의 주요 기여는 다음과 같다. 첫째, 기존 인과발견 결과 해석 및 활용에서의 한계를 제시하고, 이를 보완하기 위한 동적 요인 기여 점수를 제안한다. 둘째, 기존 해석 방식과의 비교를 통해 제안 점수의 유효성을 검증한다. 셋째, LME 구리 현물가격에 대한 요인군별 월별 기여를 분해하여 가격 변동 국면별 주요 요인을 해석할 수 있는 방법론적 기반을 마련한다.

II. 관련 연구 및 이론적 배경

2.1 시계열 인과발견의 정의와 발전 과정

인과발견은 관측 자료 또는 개입 자료로부터 변수들 사이의 가능한 인과 구조를 추정하는 방법이다[5]. 시계열 인과발견은 인과발견 문제를 시간 순서가 존재하는 자료로 확장한 것으로, 특정 변수의 변화가 즉시 또는 일정한 시차를 두고 다른 변수에 반영될 수 있음을 고려하여 변수 간 방향성뿐 아니라 시간 지연을 포함한 인과 그래프를 추정한다[1]. 예를 들어 $X_{(t-3)} \rightarrow Y_t$는 원인 변수 X의 3시점 전 값이 현재 시점의 대상 변수 Y와 관련된 지연 관계를 의미한다.

시계열 인과발견의 출발점인 그레인저 인과는 한 변수의 과거 정보가 다른 변수의 미래 예측에 도움이 되는지를 기준으로 변수 간 관계를 판단하는 방법이다[7]. 그러나 그레인저 인과는 주로 예측력 개선 여부에 초점을 두기 때문에, 여러 변수가 동시에 작용하는 고차원 시계열에서 직접적 연결 구조를 구분하는 데에는 한계가 있다[8].

이러한 한계를 보완하는 접근으로, PCMCI 계열 알고리즘은 조건부 독립성 검정과 그래프 탐색 절차를 결합하여 고차원 시계열에서 자기상관과 시차 효과를 고려한 인과관계 후보를 탐색할 수 있도록 제안되었다[9]. PCMCI 계열 알고리즘의 결과는 변수 간 지연 간선을 포함하는 인과 그래프와 각 간선에 대한 val로 나타난다. 여기서 지연 간선은 특정 원인 변수가 일정한 시차를 두고 대상 변수와 연결되는 관계를 의미하며, val은 해당 지연 관계가 다른 변수들의 영향을 고려한 이후에도 해당 관계

가 데이터상에서 얼마나 뚜렷하게 나타나는지를 보여준다[10].

2.2 인과발견에서 지켜져야 하는 가정

시계열 인과발견 결과를 인과적 관계로 해석하기 위해서는 몇 가지 주요 가정이 전제된다[5].

첫째, 인과 마르코프 가정(Causal Markov Condition)은 어떤 변수의 직접 원인에 해당하는 부모 변수들을 조건으로 두었을 때, 그래프상에서 분리된 다른 변수들과 조건부 독립이 성립한다고 보는 가정이다[6].

둘째, 인과 충실성 가정(Causal Faithfulness Assumption)은 데이터에서 관찰되는 조건부 독립성이 우연한 효과 상쇄의 결과가 아니라, 인과 그래프의 분리 구조에 의해 나타난다고 보는 가정이다[6].

셋째, 인과 충분성 가정(Causal Sufficiency Assumption)은 분석 대상 변수들 사이에 존재하는 모든 공통 원인이 관측 변수 집합에 포함되어 있다고 보는 가정이다[6].

넷째, 인과 정상성 가정(Causal Stationarity Assumption)은 변수 간 인과관계와 작동 메커니즘이 일정 기간 동안 시간에 따라 변하지 않는다고 보는 가정이다[5].

그러나 현실 자료에서는 지정학적 사건, 공급망 충격, 투자 심리 등 관측되지 않은 공통 원인이나 외생 요인이 존재할 수 있어 인과 충분성 가정을 완전히 만족하기 어렵다[6]. 이에 따라 인과발견 방법은 기존 가정이 엄격히 성립하지 않는 상황을 보완하는 방향으로 확장되어 왔으며, LPCMCI는 잠재 교란이 존재할 수 있는 관측 시계열에서도 시차 관계를 함께 추정할 수 있도록 설계된 PCMCI 계열의 확장 방법이다[11]. 본 연구는 구리 시장 자료에 관측되지 않은 외생 요인이 존재할 가능성을 고려하여 LPCMCI를 사용한다.

또한 구리 가격 변동은 전체 분석 기간 동안 하나의 고정된 인과구조로 설명되기 어렵다. 시장 국면에 따라 변수 간 관계가 달라질 수 있으므로, 인과 정상성 가정 역시 현실 자료에서 엄격히 성립한다고 보기는 제한적이다[5]. 이러한 한계를 보완하기 위해 본 연구는 전체 기간에 대해 하나의 인과구조를 추정하는 대신, 36개월 길이의 분석 윈도우를 1개월 단위로 이동시키며 각 구간마다 LPCMCI를 반복 적용하는 슬라이딩 윈도우 방식을 사용한다.

III. 동적 요인 기여 프레임워크

3.1 기존 인과발견 결과 해석에서의 한계

시계열 인과발견 결과로 산출되는 인과 그래프는 변수 간 연결 구조와 시차적 관계를 파악하는 데 유용하다[1]. 특히 PCMCI 계열 알고리즘은 특정 변수가 몇 개월의 시차를 두고 대상 변수와 연결되는지를 추정함으로써, 구리 가격 변동과 관련된 후보 요인을 구조적으로 탐색할 수 있게 한다[9].

그러나 이러한 결과를 실제 시장 해석이나 의사결정에 바로 활용하기에는 한계가 있다. 해당 알고리즘에서 제공하는 *val*은 특정 지연 관계가 데이터에서 얼마나 뚜렷하게 관측되는지를 나타낸다[10]. 즉, *val*은 변수 간 연결의 상대적 강도를 비교하는 데에는 유용하지만, 특정 시점에서 원인 변수가 실제로 어느 정도 변화했는지, 그리고 그 변화가 가격 상승 또는 하락 방향과 어떻게 연결되는지는 직접적으로 보여주지 않는다[5].

따라서 기존 *val* 기반 산출값만으로는 시점별 가격 변동 요인을 실무 의사결정에 활용 가능한 형태로 해석하는 데 한계가 있다. 이에 본 연구는 인과발견 결과를 원인 변수의 시점별 변화와 결합하여, 가격 변동 방향과 상대적 기여를 해석할 수 있는 동적 요인 기여 점수로 나타내고자 한다.

3.2 동적 요인 기여 점수

본 연구는 LPCMCI 결과를 가격 변동 해석에 활용하기 위해 동적 요인 기여 점수를 정의한다. 이 점수는 각 슬라이딩 윈도우에서 선택된 지연 간선의 부호가 있는 val과 해당 시차의 원인 변수 표준화 값을 결합하여 산출된다. 여기서 조건부 독립성 검정 통계값인 val은 해당 지연 관계가 조건부 독립성 검정에서 양 또는 음의 방향으로 얼마나 뚜렷하게 나타나는지를 의미하며, 원인 변수의 표준화 값은 해당 시점에서 원인 변수가 평소 수준 대비 얼마나 상승 또는 하락했는지를 나타낸다.

구체적으로, 윈도우 w의 종료 시점을 T_w, 원인 변수를 i, 종속 변수인 LME 구리 현물가격 전월 대비 변동률을 y, 시차를 τ라고 할 때, 개별 지연 간선의 동적 요인 기여 점수는 (1)과 같이 정의한다.

$$D_{(i,y,\tau)}^{(w)} = val_{(i,y,\tau)}^{(w)} \cdot Z^{(w)}\left(X_{(i,T_w-\tau)}\right) \qquad (1)$$

여기서 $\mathrm{val}_{(i,y,\tau)}^{(w)}$는 윈도우 w에서 LPCMCI가 산출한 조건부 독립성 검정 통계값이며 표준화 값 $Z^{(w)}X_{(i,T_w-\tau)}$는 해당 윈도우 내부의 평균과 표준편차를 기준으로 표준화한 지연 시점의 원인 변수값이다. 표준화 값은 (2)와 같이 계산한다.

$$Z^{(w)}\left(X_{(i,T_w-\tau)}\right) = \frac{X_{(i,T_w-\tau)} - \mu_i^{(w)}}{\sigma_i^{(w)}} \qquad (2)$$

$\mu_i^{(w)}$와 $\sigma_i^{(w)}$는 각각 윈도우 w내부에서 계산한 원인 변수 i의 평균과 표준편차이다. 따라서 동적 요인 기여 점수는 해당 윈도우 내에서 원인 변수가 평소 수준 대비 어느 방향으로 얼마나 벗어났는지를 함께 반영한다.

동적 요인 기여 점수의 부호는 가격 변동 방향에 대한 해석에 사용한다. 동적 요인 기여 점수가 양수이면 해당 요인이 구리 가격 상승 방향과 연결된 신호로, 음수이면 구리 가격 하락 방향과 연결된 신호로 해석한다. 또한 점수의 절댓값이 클수록 해당 시점에서 해당 변수 또는 요인군이 가격 변동 해석에 상대적으로 크게 기여한 것으로 본다.

3.3 전체 프레임워크

본 연구의 전체 파이프라인은 ① 입력 데이터 정의, ② 슬라이딩 윈도우 기반 동적 인과구조 추정, ③ LPCMCI 그래프 추정 및 대상 변수 유입 간선 추출, ④ 동적 요인 기여 점수 산출의 절차로 구성된다.

① 먼저 종속 변수와 원인 변수를 구성한다. 이때 LPCMCI는 예측모형처럼 사전에 종속 변수와 독립 변수의 관계를 고정하여 학습하는 방식이 아니라, 전체 변수 집합을 노드로 두고 변수 간 가능한 방향 및 시차 관계를 탐색하는 구조학습 기반 방법이다[11]. 따라서 본 연구에서 종속 변수는 추정된 전체 인과 그래프 중 해석의 기준이 되는 대상 노드를 의미한다. 또한 각 변수는 월별 변동률 또는 전년 대비 변동률 형태로 변환하여 비정상성을 완화하였다[5].

② 다음으로 슬라이딩 윈도우를 이동시키며 각 구간에 LPCMCI를 반복 적용한다. 이는 각 윈도우 내부에서 변수 간 관계가 비교적 안정적이라는 전제하에, 시기별 인과구조 변화를 추정하기 위한 절차이다.

③ 각 구간에 LPCMCI를 적용하여 변수 간 시차적 인과구조를 추정한 뒤, 최종 그래프에서

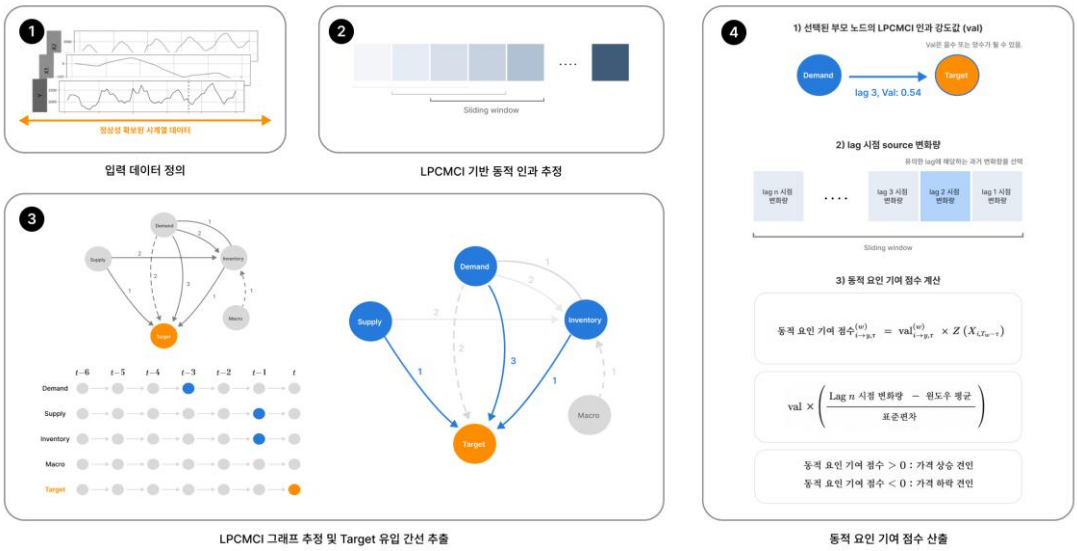


그림 1. 동적 요인 기여 프레임워크

종속변수로 유입되는 유의한 지연 간선만을 추출한다.

④ 마지막으로, 추출된 지연 간선에 대해 LPCMCI가 산출한 *val*과 해당 시차의 원인 변수 표준화 값을 결합하여 식 (1)과 식 (2)에 따라 동적 요인 기여 점수를 계산한다. 이후 수요, 공급, 재고, 거시의 요인군 단위로 집계하여, 각 시점의 가격 상승 또는 하락이 어떤 시장 요인과 정합적으로 연결되어 있었는지를 해석하는 데 사용한다.

IV. 실험

4.1 사용 데이터 및 파라미터 설정

표 1. 분석 변수 구성

요인군	변수	데이터 출처
종속변수	LME 구리 현물 가격	*London Metal Exchange (LME)*
수요	중국 고정자산투자	*National Bureau of Statistics of*
	중국 산업생산	*China*
	미국 건축허가 건수	*U.S. Census Bureau*
공급	글로벌 구리 생산량	*International Copper Study Group (ICSG)*
	ICSG 구리 광산 생산능력	
	ICSG 정련구리 총생산량	
재고	CME 구리 재고	*CME Group*
	LME 구리 재고	*London Metal Exchange (LME)*
	SHFE 구리 재고	*Shanghai Futures Exchange (SHFE)*
거시	미국 달러지수	*Barchart*
	상하이종합지수	
	S&P 500 지수	

본 연구는 2016년 1월부터 2025년 8월까지의

월별 구리 시장 데이터를 사용하였다. 전체 데이터는 116개월의 관측치로 구성되며, 날짜 변수를 제외한 총 13개의 시계열 변수를 분석 대상으로 하였다. 이 중 LME 구리 현물가격의 전월 대비 변동률을 종속 변수로 설정하고, 나머지 12개 변수는 구리 가격 변동과 관련된 후보 원인 변수로 구성하였다. 후보 원인 변수는 경제적 의미에 따라 수요, 공급, 재고, 거시의 네 가지 요인군으로 구분하였다[12]. 이를 통해 구리 가격 변동을 개별 변수 수준뿐 아니라 시장 요인군 단위에서도 해석할 수 있도록 하였다.

실험에 사용한 분석 파라미터는 아래와 같이 설정되었다.

표 2. 파라미터 설정

항목	값
분석 빈도	월별
데이터 기간	2016.01-2025.08 (116개월)
윈도우 크기	36개월
윈도우 스텝	1개월
tau_min, tau_max	1개월 ~ 6개월
pc_alpha	AIC 기반 자동 선택
조건부 독립성 검정	ParCorr

윈도우 크기는 추정 안정성과 국면 변화 포착 가능성을 함께 고려하여 설정하였다. 윈도우 길이가 짧을 경우 조건부 독립성 검정과 간선 추정의 불안정성을 높일 수 있고, 반대로 너무 긴 윈도우는 시기별 인과구조 변화를 희석할 수 있다[5][9]. 이에 본 연구에서는 월별 원자재 시장에서 약 3년 수준의 시장 국면을 반영하면서도, 최대 시차 6개월 적용 후 약 30개월의 유효 관측치를 확보할 수 있도록 윈도우 길이를 36개월로 설정하였다.

최대 시차(tau_min, tau_max)는 수요, 공급, 재고 및 거시 요인의 영향이 가격에 분기 또는 반기 수준의 지연을 두고 나타날 수 있음을 고려하되, 후보 간선 증가와 추정 불안정성을 방지하기 위해 1개월에서 6개월 범위로 설정하였다.

조건부 독립성 검정에는 ParCorr를 사용하였으며[9], 이때 val은 조건화 변수의 선형 영향을 제거한 뒤 남은 두 잔차 간 피어슨 상관계수로 해석할 수 있다[10].

PC 단계의 부모 후보 선별에 사용되는 임계 유의수준인 pc_alpha는 첫 번째 윈도우에서 AIC 기준으로 자동 선택한 뒤 전체 윈도우에 동일하게 적용하였다[10]. 각 윈도우의 LPCMCI 결과 중 최종 그래프에 포함된 링크를 분석 대상 간선으로 정의하되, 자기회귀 간선은 제외하고 다른 시장 변수에서 LME 구리 현물가격 변동률로 유입되는 시차 간선만 동적 요인 기여 점수 계산에 활용하였다.

4.2 동적 요인 기여 점수의 가격 변동 정합성 검증

동적 요인 기여 프레임워크의 해석 가능성을 평가하기 위한 첫 번째 실험으로, LPCMCI 결과를 가격 변동 해석에 활용하는 세 가지 점수화 방식을 비교하였다. 본 실험의 목적은 LPCMCI가 산출한 간선 정보만으로 가격 변동을 설명할 수 있는지, 그리고 원인 변수의 시점별 변화량을 결합할 때 실제 구리 가격 변동률과의 정합성이 개선되는지를 확인하는 것이다.

비교 대상은 val 단독 점수, val과 원인 변수값의 결합 점수, 동적 요인 기여 점수의 세 가지이다. 첫 번째 방식은 각 윈도우에서 LME 구리 가격 변동률로 유입되는 유의한 지연 간선들의 val을 집계하여 해당 시점의 점수로 사용한다. 두 번째 방식은 val에 해당 시차의 원인 변수값을 결합하되 변수별 표준화는 수행하지 않는다. 세 번째 방식은 val에 해당 시차의 원인 변수 표준화 값을 결합한 점수로, 본 연구에서 제안하는 동적 요인 기여 점수이다.

각 방식의 정합성은 실제 구리 가격 전월 대비 변동률과의 피어슨 상관계수, 스피어만 상관계수, 방향 일치율로 평가하였다. 피어슨 상관계수는 점수의 크기가 실제 가격 변동률의 크기와 함께 움직이는 정도를, 스피어만 상관계수는 값의 순위를 기준으로 두 시계열의 변동 패턴이 유사한 정도를 나타낸다. 방향 일치율은 점수와 실제 가격 변동률의 부호가 일치한 비율을 의미한다.

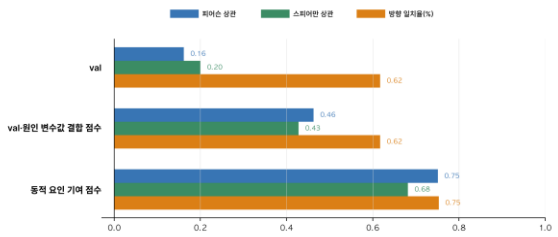


그림 2. 비교 대상별 가격 변동 정합성 비교

분석 결과, *val* 단독 점수는 피어슨 상관계수 0.16, 스피어만 상관계수 0.20, 방향 일치율 0.62로 세 방식 중 가장 낮은 정합성을 보였다. 이는 LPCMCI가 선택한 간선의 *val*만으로는 특정 시점의 가격 상승 또는 하락을 충분히 설명하기 어렵다는 것을 의미한다.

반면 *val*과 원인 변수값의 결합 점수는 피어슨 상관계수 0.46, 스피어만 상관계수 0.43으로 개선되어, 간선 정보에 원인 변수의 시점별 움직임을 결합하는 것이 가격 변동 해석에 중요함을 보여준다. 다만 방향 일치율은 0.62로 *val* 단독 점수와 동일하였는데, 이는 전체 일치 비율이 같다는 의미일 뿐 방향이 일치한 시점이 완전히 동일하다는 의미는 아니다.

제안한 동적 요인 기여 점수 기준은 피어슨 상관계수 0.75, 스피어만 상관계수 0.68, 방향 일치율 0.75로 세 지표 모두에서 가장 높은 정합성을 보였다. 이는 선택된 지연 간선의 통계적 관계와 해당 시점 원인 변수의 표준화된 변화를 함께 고려하는 방식이 실제 가격 변동률과 가장 잘 부합함을 의미한다. 또한 *val*과 원인 변수값의 결합 점수보다 높은 결과를 보였다는 점은, 단순한 원인 변수값의 증감보다 윈도우 내부 기준에서 현재 변화가 평균 대비 얼마나 강하거나 약한지를 반영하는 과정이 필요함을 보여준다. 즉, 원인 변수의 변화량이 양이더라도 해당 윈도우의 평균적 변화 수준에 비해 낮다면 상대적으로 약한 상승 신호로 해석될 수 있으며, 이러한 표준화 과정이 가격 변동 방향과의 정합성을 높이는 데 기여한 것으로 볼 수 있다. 이러한 결과는 전체 기간에 대한 시계열 비교에서도 확인된다.

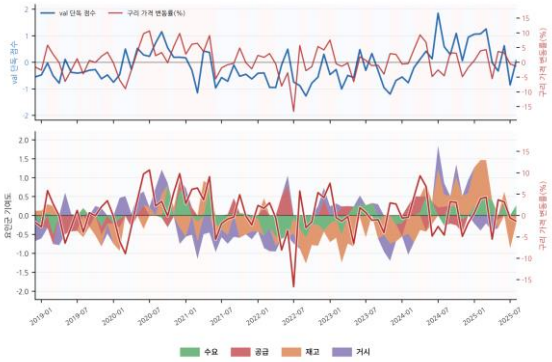


그림 3. *val* 단독 점수와 구리 가격 변동률(%)

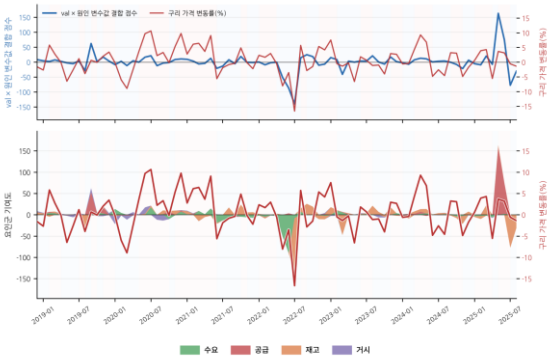


그림 4. *val*과 원인 변수값의 결합 점수와 구리 가격 변동률(%)

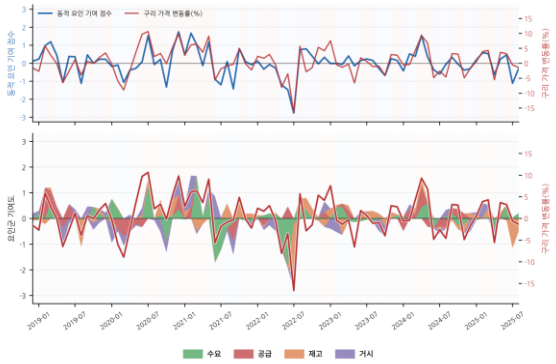


그림 5. 동적 요인 기여 점수와 구리 가격 변동률(%)

이러한 결과는 전체 기간에 대한 시계열 비교에서도 확인된다. *val* 단독 점수는 특정 시점에서 원인 변수의 실제 변화 방향과 크기를 충분히 반영하지 못해, 구리 가격 변동률의 흐름과 불일치하는 구간이 나타났다. 반면 동적 요인 기여 점수 기준은 실제 구리 가격 변동률과 더 유사한 흐름을 보였으며, 요인군별 기여를 함께 제시함으로써 각 시점의 가격 상승 또는 하락이 어떤 시장 요인과 연결되어 있었는지를 해석할 수 있다.

4.3 요인군별 등장 빈도 및 해석 안정성 검증

두 번째 실험에서는 앞서 정의한 수요, 공급, 재고, 거시의 네 개 요인군을 기준으로 동적 요인 기여 점수를 집계하고, 요인군별 등장 빈도와 방향 일치율을 분석하였다. 본 실험의 목적은 제안 프레임워크가 개별 변수 수준을 넘어 요인군 단위에서도 안정적인 해석을 제공하는지 확인하는 데 있다.

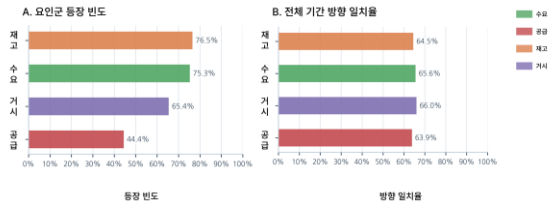


그림 6. 요인군별 등장 빈도와 방향 일치율 비교

여기서 등장 빈도는 평가 기간 동안 해당 요인군이 가격 변동 해석에 포함된 비율을 의미하며, 방향 일치율은 해당 요인군이 등장한 시점에서 요인군 기여 점수의 부호와 실제 구리 가격 변동률의 부호가 일치한 비율을 의미한다. 본 연구에서는 등장 빈도를 통해 요인군별 해석 활용도를 확인하고, 방향 일치율을 통해 해당 해석 신호가 실제 가격 상승·하락 방향과 정합적으로 작동하는지를 평가하였다.

전체 기간을 기준으로 요인군별 등장 빈도와 방향 일치율을 비교한 결과, 재고와 수요 요인군은 각각 76.5%, 75.3%로 상대적으로 높은 등장 빈도를 보였다. 거시 요인군은 65.4%로 중간 수준이었으며, 공급 요인군은 44.4%로 가장 낮게 나타났다.

방향 일치율은 재고, 수요, 거시, 공급 요인군에서 각각 64.5%, 65.6%, 66.0%, 63.9%로 나타났으며, 요인군 간 최대 차이는 2.1%p에 그쳤다. 이는 특정 요인군에서만 방향 정합성이 높게 나타난 것이 아니라, 동적 요인 기여 점수가 요인군 전반에서 비교적 유사한 방향 안정성을 보였음을 의미한다.

등장 빈도 측면에서는 재고와 수요 요인군이 상대적으로 자주 선택되었다. 이는 본 연구의 분석 기간과 변수 구성에서 재고 및 수요 관련 지표가 구리 가격 변동 해석에 반복적으로 포함되었음을 보여준다. 반면 공급 요인군은 상대적으로 낮은 등장 빈도를 보였는데, 이는 공급 변수가 글로벌 생산량 및 생산능력과 같은 집계 지표 중심으로 구성되어 단기적 공급 충격이나 개별 생산 차질을 충분히 반영하지 못했을 가능성과 관련될 수 있다.

종합하면, 실험 2를 통해 동적 요인 기여 점수가 요인군 단위에서도 유사한 방향 안정성을 보였으며, 본 연구의 데이터 구성에서는 재고와 수요 요인군이 가격 변동 해석에 상대적으로 높은 활용도를 보였음을 확인하였다.

V. 결론 및 한계

본 연구는 시계열 인과발견 결과를 실무 차원의 가격 변동 해석에 활용하기 위해, 동적 요인 기여 프레임워크를 제안하였다. 본 연구의 의의는 시계열 인과발견 결과를 단순한 구조 추정 결과에 머무르게 하지 않고, 가격 상승·하락에 대한 시점별 요인 기여 신호로 변환했다는 점에 있다. 또한 관측 가능한 시장 데이터가 존재하는 원자재 카테고리에서 개별 변수 수준의 해석뿐 아니라 요인군 단위의 해석을 가능하게 함으로써, 방법론적 분석 결과와 도메인 해석 사이의 간극을 줄였다는 점에서도 의미가 있다.

다만 본 연구는 다음과 같은 한계를 가진다. 첫째, 본 연구에서는 조건부 독립성 검정으로 선형 관계를 가정하는 ParCorr를 사용하였다. 따라서 변수 간 관계가 비선형적이거나 비정규적 의존성이 강한 경우에는 지연 관계를 충분히 포착하지 못할 수 있다[9][10]. 둘째, 인과발견은 지도학습과 달리 명확한 정답 구조나 손실함수가 주어지지 않는 경우가 많아, 하이퍼파라미터 선택과 결과 검증이 상대적으로 까다롭다[13]. 본 연구에서도 윈도우 크기, 최대 시차, pc_alpha와 같은 설정은 추정되는 인과구조와 동적 요인 기여 점수에 영향을 줄 수 있으므로, 이러한 설정에 따른 결과의 변동 가능성을 고려할 필요가 있다.

향후 연구에서는 다음과 같은 방향으로 확장할 수 있다. 첫째, GPDC, CMIknn 등 비선형 조건부 독립성 검정을 함께 적용하여 ParCorr 기반 결과와 비교함으로써, 변수 간 비선형 관계를 고려한 프레임워크의 적용 가능성을 검토할 필요가 있다. 둘째, 윈도우 크기, 최대 시차, pc_alpha 등 주요 하이퍼파라미터 설정에 따른 민감도 분석과 안정성 기반 평가를 수행하여 결과의 강건성을 검토할 필요가 있다. 셋째, 뉴스 데이터와 실제 시장 사건을 활용한 정성적 검증을 추가하고, 데이터 범위와 변수 품질을 확장함으로써 구리뿐 아니라 원자재 시장 전반에 적용 가능한 가격 변동 해석 프레임워크로 발전시킬 수 있을 것이다.

참고문헌

[1] C. K. Assaad, E. Devijver, and E. Gaussier, “Survey and evaluation of causal discovery methods for time series,” Journal of Artificial Intelligence Research, Vol. 73, pp. 767–819, Feb. 2022.

[2] J. Baffes and P. Nagle, Eds., “Commodity Markets: Evolution, Challenges, and Policies,” World Bank, Dec. 2022.

[3] U.S. Geological Survey, “Copper statistics and information,” National Minerals Information Center, Accessed Jun. 5, 2026.

[4] International Energy Agency, “Energy and AI,” IEA, Apr. 2025.

[5] J. Runge, A. Gerhardus, G. Varando, V. Eyring, and G. Camps-Valls, “Causal inference for time series,” Nature Reviews Earth & Environment, Vol. 4, No. 7, pp. 487–505, Jul. 2023.

[6] C. Glymour, K. Zhang, and P. Spirtes, “Review of causal discovery methods based on graphical models,” Frontiers in Genetics, Vol. 10, Article 524, Jun. 2019.

[7] C. W. J. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” Econometrica, Vol. 37, No. 3, pp. 424–438, Aug. 1969.

[8] M. Eichler, “Granger causality and path diagrams for multivariate time series,” Journal of Econometrics, Vol. 137, No. 2, pp. 334–353, Apr. 2007.

[9] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, “Detecting and quantifying causal as-

sociations in large nonlinear time series datasets," Science Advances, Vol. 5, No. 11, Article eaau4996, Nov. 2019.

[10] J. Runge, E. Gillies, E. V. Strobl, and S. Palachy-Affek, "jakobrunge/tigramite: Tigramite 5.2" [Computer software], Zenodo, Version 5.2, Mar. 2023.

[11] A. Gerhardus and J. Runge, "High-recall causal discovery for autocorrelated time series with latent confounders," Advances in Neural Information Processing Systems, Vol. 33, pp. 12615–12625, Dec. 2020.

[12] A. Kabundi, G. Vasishtha, and H. Zahid, "The nature and drivers of commodity price cycles," in Commodity Markets: Evolution, Challenges, and Policies, J. Baffes and P. Nagle, Eds., World Bank, pp. 185–219, Dec. 2022.

[13] K. Biza, I. Tsamardinos, and S. Triantafillou, "Tuning causal discovery algorithms," Proc. 10th International Conference on Probabilistic Graphical Models, Proceedings of Machine Learning Research, Vol. 138, pp. 17–28, Sep. 2020.

멀티모달 융합 매칭 기반 지능형 분실물 검색 시스템

김수연[O], 김동호*, 민병석*
[O]세종대학교 데이터사이언스학과, *(주) 마크애니
*세종대학교 인공지능 데이터사이언스학과
e-mail : rlatndusgu@sju.ac.kr[O], kimdh@markany.com*, bmin@sejong.ac.kr*

Multimodal Fusion Matching based Intelligent Lost and Found Search System

Sooyeon Kim[O], Dong ho Kim*, Byeongseok Min*
[O]Dept. of Data Science, SeJong University, *MarkAny Co. Ltd
*Dept. of AI & Data Science, SeJong University

요 약

본 논문에서는 기존 분실물 서비스의 구조적 한계를 극복하기 위해 멀티모달 융합 매칭 기반의 지능형 분실물 검색 시스템을 제안한다. 기존 시스템은 텍스트 키워드 검색과 단순 필터링에만 의존하여 자동 매칭 기능이 전무하고, VLM 기반 자동 매칭이 일부 도입되었으나 카테고리와 색상 수준의 제한적 분석에 그쳐 로고, 브랜드, 흠집 등 세부 외관 특징은 사용자가 직접 입력해야 하며, 이미지 간 직접 유사도 비교 및 매칭 근거 제공 기능 또한 부재하다. 본 연구는 이를 개선하기 위해 (1) 한국어 특화 VLM을 파인튜닝하여 이미지 속 객체 세부 외관 묘사를 자동 추출하는 이미지 캡셔닝 모듈, (2) 분실물 도메인 특화 Re-ID, (3) Re-ID와 bge-m3 하이브리드 임베딩을 결합한 멀티모달 유사도 계산의 세가지 핵심 모듈을 설계하고, RAG 구조를 통해 분실자와 습득물 정보를 함께 참조한 매칭 근거를 자동 생성하여 사용자에게 제공하는 통합 파이프라인을 구축한다.

키워드 : 멀티모달, Re-ID, RAG, 이미지 캡셔닝, 임베딩

I. 서론

분실물은 일상에서 빈번하게 발생하는 사회적 문제이다. 서울교통공사에 따르면 2024년 서울 지하철 분실물 신고 건수는 약 15만 건에 달하며, 이 중 56.8%만이 주인에게 반환되었다. 반환되지 못한 분실물은 습득 기관에 일정 기간 보관 후 폐기 처리되어 자원 낭비로 이어지며, 분실자 입장에서도 금전적·심리적 손실이 크다. 이러한 낮은 반환율의 근본적인 원인 중 하나는 현행 분실물 검색 시스템의 구조적 한계에 있다.

현재 국내 대표 공공 서비스인 경찰청 유실물 종합관리시스템은 물품명, 장소, 날짜, 색상 등을

활용한 텍스트 기반 검색만 제공하며 자동 매칭 기능이 없어 사용자가 직접 다수의 결과를 확인해야 한다. 텍스트 검색은 동일하거나 유사한 물품이 다수 등록된 경우 원하는 결과를 찾기까지 많은 시간이 소요된다. 민간 영역에서는 VLM을 활용한 분실물 매칭 서비스가 일부 제공되고 있으나, 카테고리와 색상 수준의 제한적인 분석에 머물러 로고, 브랜드, 흠집 등의 세부 외관 특징은 사용자가 직접 입력해야 한다. 또한 외부 상용 API 기반 구조는 민감정보 유출 위험이 있으며, 매칭 결과에 대한 근거를 제공하지 않아 신뢰성이 낮다는 한계를 가진다.

본 논문에서는 이러한 한계를 극복하기 위해 멀티모달 융합 매칭 기반의 지능형 분실물 검색 시스템을 제안한다. 제안 시스템은 분실자가 한국어 특화 VLM을 파인튜닝하여 이미지로부터 색상, 형태, 로고, 브랜드, 흠집 등 세부 외관 특징 전체를 자동 추출하고, Re-ID 기반 image-to-image 유사도와 bge-m3 하이브리드 텍스트 임베딩을 통합한 멀티모달 매칭을 수행한다. 최종적으로 RAG 구조를 통해 분실자 정보와 습득물 정보를 함께 참조한 매칭 근거를 자연어로 자동 생성하여 멀티모달 검색 기반 매칭 결과의 설명 가능성과 신뢰도를 높인다.

II. 관련 연구

2.1 관련 기술

객체 탐지는 이미지 내 관심 객체의 위치와 영역을 자동으로 검출하는 기술이다. YOLO 계열 모델은 높은 처리 속도와 정확도를 바탕으로 다양한 객체 탐지 분야에서 널리 활용되고 있다. Re-identification(Re-ID)은 서로 다른 환경에서 촬영된 동일 객체를 식별하는 기술로, 객체의 외관 특징을 기반으로 이미지 간 유사도를 계산하는 데 활용된다. TransReID는 트랜스포머 기반 구조를 활용하여 전역 및 지역 특징을 함께 학습함으로써 외관 변화와 도메인 차이에 강건한 특징 표현을 추출할 수 있다.

LLM은 자연어 이해 및 생성 능력을 바탕으로 비정형 텍스트를 구조화하거나 의미 정보를 추출하는 데 활용된다. 또한 LoRA(Low-Rank Adaptation)는 전체 모델 파라미터를 갱신하지 않고도 특정 도메인에 대한 성능을 효율적으로 향상시킬 수 있는 대표적인 경량 파인튜닝 기법으로 널리 사용되고 있다.

VLM은 이미지와 텍스트를 함께 처리하는 멀티모달 모델로, 이미지 캡셔닝, 시각 질의응답(VQA) 등 다양한 태스크에 활용된다. 최근 HyperCLOVA X, Kanana 등 한국어 특화 VLM이 등장하면서 국내 환경에서도 높은 활용 가능성을 보이고 있다.

텍스트 임베딩 기반 검색은 쿼리와 문서를 벡터 공간에 매핑한 후 유사도를 계산하여 관련 문서를 검색하는 방법이다. Dense retrieval은 의미적 유사도 파악에 강점을 가지며, Sparse retrieval은 정확한 키워드 매칭에 강점을 가진다. BGE-M3는 Dense와 Sparse 임베딩을 모두 지원하는 하이브리드 검색 모델로 다국어 환경에서도 우수한 성능을 제공한다.

RAG(Retrieval-Augmented Generation)는 검색된 정보를 컨텍스트로 활용하여 언어 모델이 사실 기반 응답을 생성하도록 하는 방법론이다. 검색 결과에 대한 설명 가능성을 제공할 수 있어 질의응답, 추천 시스템 등 다양한 분야에서 활용되고 있다.

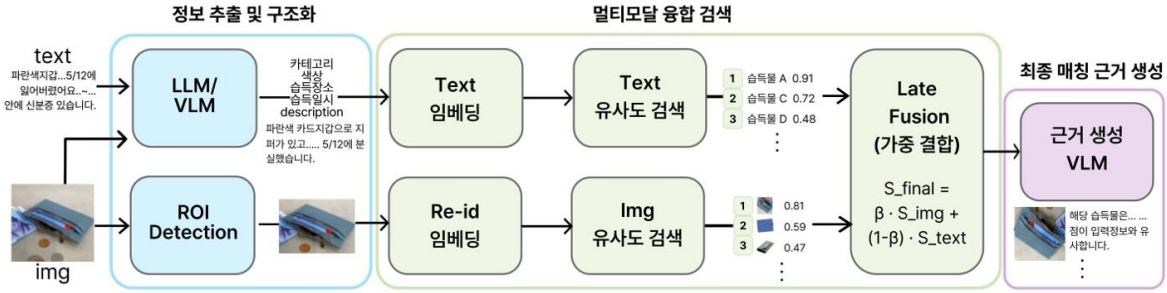


그림 1. 시스템 파이프라인

III. 제안방법

3.1 시스템 아키텍처

제안 시스템은 텍스트 및 이미지 기반 멀티모달 retrieval 구조로 설계되며, 입력 정보 구조화, 후보 검색, 매칭 근거 생성의 세 단계로 구성된다. 분실자의 텍스트 입력은 LLM을 통해 카테고리, 색상, 형태 등의 구조화된 정보로 변환되며, 이미지가 함께 입력된 경우에는 추가적으로 VLM 기반 캡셔닝 및 Re-ID 기반 이미지 특징 추출을 수행한다. 이후 텍스트 및 이미지 임베딩을 기반으로 벡터 DB에서 각각 retrieval을 수행하여 상위 후보를 검색하고, retrieval 결과를 late fusion 방식으로 통합하여 최종 ranking을 산출한다. 검색된 후보 정보와 유사도 결과는 컨텍스트로 활용되어 VLM이 매칭 근거를 생성한다.

3.2 객체 탐지 및 Re-ID 기반 유사도 계산

습득물 이미지에는 배경 잡음이 포함되어 있어 관심 객체 영역만을 정확히 검출하는 전처리가 필수적이다. 본 연구에서는 YOLO 모델을 공공 데이터포털에서 수집한 한국형 분실물 특화 데이터셋으로 파인튜닝하여 분실물 객체의 ROI를 정밀하게 추출한다.

ROI가 추출된 이미지에 대해 TransRe-ID 모델을 적용하여 분실물 이미지와 습득물 이미지 간의 직접적인 image-to-image 유사도를 계산한다. 기존 이미지 분류 방식과 달리 Re-ID 기반 접근은 동일 카테고리 내 서로 다른 두 이미지를 직접 비교하여 세밀한 외관 유사도를 산출할 수 있다. TransRe-ID은 Products-10K 데이터셋으로 파인튜닝하여 다양한 상품군의 외관 특징을 학습하고 분실물 도메인 적응 성능을 향상시킨다. 이를 통해 동일 카테고리 내에서도 차이를 반영한 특징 표현을 추출하며, 객체 간 유사도는 특징 벡터 간의 코사인 유사도를 기반으로 계산된다.

3.3 LLM 기반 텍스트 정제 및 카테고리 추출

분실자의 자유 형식 입력 텍스트는 맞춤법 오류, 비표준 표현, 모호한 서술을 포함할 수 있어 직접 매칭에 활용하기 어렵다. 본 연구에서는 한국어 특화 LLM을 LoRA로 파인튜닝하여 입력 텍스트를 카테고리, 색상, 형태, 브랜드, 특이사항 등의 구조화된 필드로 변환하는 텍스트 정제 모듈을 구성한다. 카테고리 추출에는 지갑, 가방, 전자기기, 의류, 열쇠 등으로 구성된 분실물 특화 데이터를 구축하여 활용한다. 정제된 텍스트는 bge-m3 모델로 임베딩하여 텍스트 유사도 계산에 활용되며, 카테고리 정보는 Retrieval 단계의 필터링에 활용된다. GT 데이터는 생성형 모델을 활용하여 구축하며, 직접 작성한 캡션과의 임베딩 유사도 비교를 통해 품질을 검증하였다.

3.4 VLM 기반 객체 묘사 및 카테고리 추출

ROI가 추출된 습득물 이미지로부터 VLM 캡셔닝 모듈이 색상, 형태, 카테고리, 흠집 유무, 로고, 브랜드, 키링 등 세부 외관 특징 전체를 자동으로 추출한다. 기존 이미지 분류 방식은 단일 카테고리

레이블만을 출력하는 데 그치나, VLM 캡셔닝은 다양한 세부 속성을 자연어 형태로 동시에 추출할 수 있어 매칭 정확도 향상에 유리하다. 한국어 특화 VLM을 로컬 환경에서 LoRA로 파인튜닝하여 민감정보 유출 위험을 최소화한다. GT 캡션 데이터는 생성형 모델을 활용하여 구축하고 품질을 검증하였으며, 생성된 캡션은 bge-m3로 임베딩하여 텍스트 유사도 계산에 활용된다. 캡셔닝 결과는 분실자의 입력 텍스트와 함께 RAG 컨텍스트를 구성하는 핵심 정보로 사용된다.

3.5 멀티모달 유사도 계산 및 매칭

텍스트 유사도는 dense 임베딩과 sparse 임베딩을 결합한 hybrid 방식으로 산출한다. 이미지 유사도는 TransRe-ID 모델로 계산하며, 최종 매칭 점수는 텍스트 유사도와 이미지 유사도를 가중 합산하여 산출한다. 텍스트 유사도는 다음과 같이 계산되고, α는 각 유사도의 가중치를 나타낸다.

$$S_{text} = \alpha \cdot S_{dense} + (1-\alpha) \cdot S_{sparse}$$

여기서 각 query에 대해 텍스트 retrieval과 이미지 retrieval을 독립적으로 수행한다. 텍스트 retrieval에서는 상위 100개 후보를 추출하고, 이미지 Re-ID retrieval에서도 상위 100개 후보를 추출한다. 이후 두 retrieval 결과의 union을 구성하여 최종 fusion 대상 후보 집합을 생성한다.

$$S_{fusion} = \beta \cdot S_{img} + (1-\beta) \cdot S_{text}$$

여기서 β는 이미지 유사도 가중치이며, $(1-\beta)$는 텍스트 유사도 가중치이다. Fusion score 계산 이후 후보들을 재정렬하고, 최종 상위 10개 후보를 대상으로 Recall@1, Recall@5, Recall@10 성능을 평가한다.

3.6 RAG 기반 검색 및 매칭 근거 생성

본 연구에서는 멀티모달 검색 결과를 활용하여 매칭 근거를 생성하는 RAG 기반 구조를 적용한다. 검색 단계에서 선정된 후보의 VLM 캡션, 카테고리, 유사도 점수와 분실자의 입력 정보를 함께 VLM에 제공하여 VLM은 이를 바탕으로 단순한 검색 결과를 넘어 두 물건이 매칭된 이유를 자연어 형태의 근거로 생성한다. 이를 통해 검색 결과의 설명 가능성과 사용자 신뢰도를 향상시킨다.

IV. 실험 및 성능 평가

공공 데이터포털에서 수집한 분실물 이미지와 생성형 모델 기반 이미지 내 객체 묘사 텍스트 GT 데이터를 활용하였다. 최종 매칭 평가를 위해 가방, 모자, 지갑, 휴대폰, 우산 등으로 구성된 query-found pair를 구축하였다. 학습·검증·테스트 셋은 8:1:1 비율로 분할하며 PyTorch 및 HuggingFace Transformers 기반으로 구현하였다.

4.1 객체 탐지 및 Re-ID 성능 평가

ROI 추출 성능은 객체 탐지 분야 표준 지표인 mAP로 평가하고, 여기서 N은 전체 클래스 수이며, AP_i는 클래스의 Average Precision를 나타낸다.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i$$

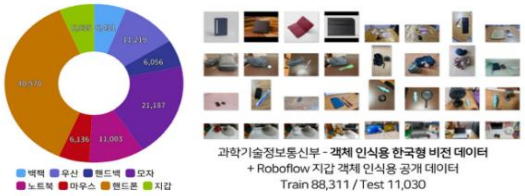


그림 2 분실물 특화 ROI 추출 학습 데이터셋

평가 지표	mAP@50(%)	mAP@50:95(%)
Yolo11x (Pre-Trained)	0.815	0.780
Yolo11x (Fine-Tuned)	0.957	0.928
Yolo26x (Pre-Trained)	0.850	0.819
Yolo26x (Fine-Tuned)	0.959	0.927

표 1. 모델별 ROI 추출 성능 비교

분실물 특화 데이터셋으로 파인튜닝한 모델은 사전학습 모델 대비 전반적으로 높은 탐지 성능을 보였으며, 특히 Yolo26x 모델에서 가장 우수한 결과가 확인되었다. 세부 성능 비교 결과는 표 1 과

같이 나타난다. Re-ID 성능은 Rank-K Accuracy 로 평가한다. Rank-K 는 전체 쿼리 Q 에 대해 상위 K 개 검색 결과 내 정답이 포함된 비율을 나타낸다.

$$Rank-K = \frac{1}{N}\sum_{i=1}^{N} 1(\mathrm{Gi} \cap \mathrm{TopKi} \neq \emptyset)$$

평가 지표	Rank-1	Rank-5	Rank-10
TransReID (Pre-Trained)	0.430	0.712	0.795
TransReID (Fine-Tuned)	0.755	0.946	0.976

표 2. Re-ID 기반 이미지 매칭 성능 비교

도메인 적응 학습을 적용한 TransReID 모델은 사전학습 모델 대비 모든 Rank-K 지표에서 향상된 성능을 보였다. 이는 분실물 도메인에 대한 파인튜닝이 객체 간 세밀한 외관 특징 학습에 효과적으로 기여할 수 있음을 시사하며, 구체적인 결과는 표 2 와 같이 확인된다.

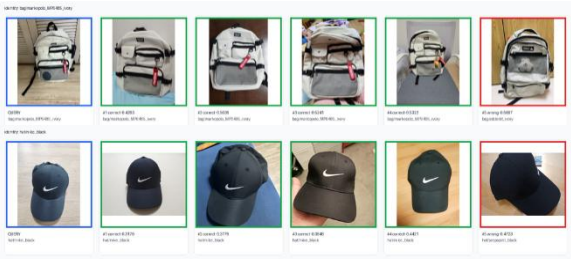


그림 3. TransReID 기반 쿼리-검색 결과 예시

4.2 LLM 및 VLM 성능 평가

$$Micro\ F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Micro F1 은 분실물 카테고리 데이터에서 정밀도와 재현율을 통한 분류 성능을 측정한다.

$$BERTScore\ F1 = \frac{2P_{\mathrm{bert}}R_{\mathrm{bert}}}{P_{\mathrm{bert}} + R_{\mathrm{bert}}}$$

BERTScore F1 은 예측 텍스트와 정답 간의 토큰 임베딩 유사도를 기반으로 의미적 일치도를 평가하며, 표현이 완전히 일치하지 않더라도 의미적으로 유사한 예측을 반영할 수 있어 자유형식 텍스트 출력 평가에 사용하였다. LoRA 기반 파인튜닝을 적용한 결과 기본 모델 대비 카테고리 분류 및 텍스트 정제 성능이 향상되는 경향을 보였다.

평가 지표	카테고리 Micro F1		BERTScore F1	
	Base	LoRA	Base	LoRA
Exaone_3B	0.94	0.99	0.86	0.95
Qwen2.5_7B	0.98	0.99	0.86	0.95

표 3. LLM 파인튜닝 전 후 성능 비교

VLM 카테고리 분류 성능은 Micro F1 으로 이미지 속 객체 묘사는 BERTScore F1 으로 평가하였다.

평가 지표	카테고리 Micro F1		BERTScore F1	
	Base	LoRA	Base	LoRA
HyperCLOVAX_3B	0.94	0.99	0.80	0.89
kanana1.5_3B	0.95	0.98	0.71	0.91

표 4. VLM 파인튜닝 전 후 성능 비교

VLM 파인튜닝 결과 이미지 속 객체 카테고리 분류와 세부 묘사 성능이 사전학습 모델 대비 전반적으로 향상된 것을 확인하였다.

4.3 최종 멀티모달 매칭 성능 평가

텍스트 유사도 기반 retrieval 성능과 최종 멀티모달 융합 retrieval 성능은 Recall@K 로 평가한다. Recall@K 는 전체 query 에 대해 상위 개의 검색 결과 내에 정답이 포함된 비율을 의미한다.

$$Recall@K = \frac{1}{N}\sum_{i=1}^{N} 1(y_i \in \mathrm{Top}K_i)$$

여기에서 N 은 전체 qeury 수, y_i 는 query 의 정답, i는 상위 개의 retrieval 결과를 의미한다.

텍스트 retrieval 성능은 BGE-M3 기반 dense, sparse, hybrid 코사인 유사도 기반으로 평가한다.

평가 지표	R@1(%)	R@5(%)	R@10(%)
dense	0.702	0.925	0.957
sparse	0.595	0.853	0.929
hybrid(0.9:0.1)	0.712	0.925	0.960

표 5. 텍스트 유사도 검색 성능 비교

Dense 와 sparse 를 0.9:0.1 비율로 결합한 hybrid 방식이 전 지표에서 우수한 검색 성능을 보였다. 최종 retrieval 성능은 text 와 image 결과를 late fusion 으로 통합한 최종 ranking 결과를 기준으로 평가한다. 본 연구에서는 Recall@1, Recall@5, Recall@10 지표를 사용한다. 평가는 두 가지 gallery 환경에서 수행하였다. 첫 번째는 정답 found 이미지 376 개만 포함하는 Positive-only 환경이며, 이를 통해 모델의 기본 매칭 성능을 평가하였다. 두 번째는 Positive gallery 에 추가 distractor 이미지 1,330 개를 합한 총 1,706 개 규모의 gallery 환경으로, 실제 서비스 환경과 유사한 검색 조건에서의 성능을 평가하였다.

평가 지표	R@1(%)	R@5(%)	R@10(%)
텍스트 매칭	0.712	0.925	0.960
이미지 매칭	0.755	0.946	0.976
img/text 5:5	0.821	0.973	0.994
img/text 6:4	0.819	0.973	0.992
img/text 7:3	0.835	0.976	0.986

표 6. 최종 매칭 성능 비교 (positive only)

단일 모달 방식과 멀티모달 융합 방식의 최종 매칭 성능을 비교한 결과, 멀티모달 융합 방식은 모든 평가 지표에서 단일 모달 방식보다 우수한 성능을 보였으며, 특히 이미지와 텍스트 유사도를 7:3 비율로 결합하였을 때 R@1 기준 최고 성능 0.835 를 달성하였다. 이를 통해 두 모달의 상호 보완적인 정보가 검색 정확도 향상에 기여할 수 있음을 확인하였다.

평가 지표	R@1(%)	R@5(%)	R@10(%)
텍스트 매칭	0.638	0.848	0.912
이미지 매칭	0.726	0.925	0.965
img/text 7:3	0.819	0.973	0.986

표 7. 최종 매칭 성능 비교 (distract 포함)

distract data 를 추가하여 동일한 매칭 실험을 수행 결과 멀티모달 융합 방식은 R@1 0.816 을 기록하였다. 이는 검색 공간이 약 4.5 배 증가한 환경에서도 성능 저하가 크지않음을 보여주며 제안한 멀티모달 융합 방식이 실제 검색 환경에서도 강건성을 유지함을 확인하였다.

V. 결론

본 논문에서는 기존 분실물 서비스의 한계를 극복하기 위해 멀티모달 융합 매칭 기반의 지능형 분실물 검색 시스템을 제안하였다. 제안 시스템은 YOLO 기반 ROI 추출, 한국어 특화 VLM 로컬 파인튜닝을 통한 세부 외관 특징 자동 추출, TransRe-ID와 bge-m3 하이브리드 임베딩을 결합한 멀티모달 매칭을 수행한다. 또한 RAG 구조를 통해 매칭 근거를 생성함으로써 결과의 설명 가능성과 사용자 신뢰도를 높인다. 실험 결과, 제안 시스템은 단일 모달 방식 대비 Recall@10 기준 유의미한 성능 향상을 달성하였다. 향후 연구에서는 더 많은 분실물 카테고리로의 확장 및 실 서비스 환경에서의 검증을 목표로 한다.

참고문헌

[1] He, S. et al., "TransReID: Transformer-Based Object Re-Identification", ICCV, 2021.

[2] Wang, C. et al., "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information", ECCV, 2024.

[3] Yoo, K. M. et al., "HyperCLOVA X Technical Report", arXiv, 2024.

[4] Hu, E. et al., "LoRA: Low-Rank Adaptation of Large Language Models", ICLR, 2022.

[5] Chen, J. et al., "BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation", arXiv, 2024.

[6] Lewis, P. et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", NeurIPS, 2020.

119 신고 발화자의 마음위기 분류(C-CASA) 프레임워크 제안 연구

강미주 *, 정유진, 이지영O, 홍승필
*가톨릭대학교 의과학과, 한신대학교(IT영상콘텐츠학과, O IT영상콘텐츠학과, AI.SW대학)
e-mail: ak06080111@gmail.com*, 5342dbwls@hs.ac.kr, leejy0723@hs.ac.krO,sphong@hs.ac.kr

A Study of Proposed Framework for Emotional Crisis Classification(C-CASA) of 119 Callers

Kang Miju* , Jung Yujin, Lee JiyoungO, Hong Seng-Phil
*Dept. of Medical Science, The Catholic University of Korea
Dept. of IT Transmedia Contents, Hanshin University
O Dept. of IT Transmedia Contents, Hanshin University
College of AI.SW, Hanshin University

요 약

본 연구는 119 응급 신고 발화 데이터에서 정신과적 응급상황, 특히 자살 및 자해 관련 마음위기 사례를 자동으로 선별·분류하기 위한 C-CASA 기반 위험 분류 프레임워크를 제안한다. AI-Hub 119 응급 음성 데이터셋 중 자살 및 약물중독 관련 신고 기록을 대상으로, KoNLPy Okt 기반 형태소 정규화와 어간 추출을 수행하고, 정규표현식 기반 규칙을 활용하여 Accident, Behavior, Fatality, Interrupted, Intent의 5대 임상 변수 선정 및 추출하였다. 이후 정신의학 분야의 표준화된 자살 사건 분류 체계인 C-CASA를 119 병원 전 응급 환경에 맞게 재구성하여 Completed Suicide, Suicide Attempt, Preparatory Behavior, Suicide Ideation 등 8단계 위험 분류 체계를 설계하였다. 또한 물리적 위기 증거를 우선 평가하는 Risk-First 결정 트리와 Anchor 기반 다중문장 병합 규칙을 적용하여, 단일 문장 단위의 과잉 분류를 방지하고 신고 발화 전체의 맥락을 보존하도록 하였다. 제안한 프레임워크는 비정형 119 신고 발화를 임상적으로 해석 가능한 구조화된 중증도 정보로 변환함으로써, 정신응급 환자의 초기 선별, 고위험군 식별, 구급대원 의사결정 지원 및 병원 전-병원 단계 간 정보 연계의 효율성을 향상시키는 데 기여할 수 있다.

키워드: 119 응급신고, 마음위기, 정신응급, C-CASA, 자살위험 분류, 위험 기반 결정 트리, AI-Hub 데이터셋

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 한신대학교가 수행된 경기 지역혁신중심대학지원사업(경기 RISE사업)의 연구결과입니다.

I. 서론

최근 119 응급 신고 시스템은 긴급 상황에서 국민의 생명과 안전을 보호하는 병원 전(Pre-hospital) 단계의 핵심 공공 의료 서비스로 자리 잡고 있다. 특별히 신고자가 전달하는 발화 데이터는 응급 상황 판단과 구조 대응을 위한 1차적이고 가장 중요한 임상 정보로 활용된다. 자연어 처리(NLP) 등 인공지능 기술의 발전과 더불어 실제 신고 음성을 기반으로 구축된 AI-Hub 119 응급 음성 데이터셋 등은 데이터 기반 응급 상황 분석 연구에 전기를 마련하고 있다.

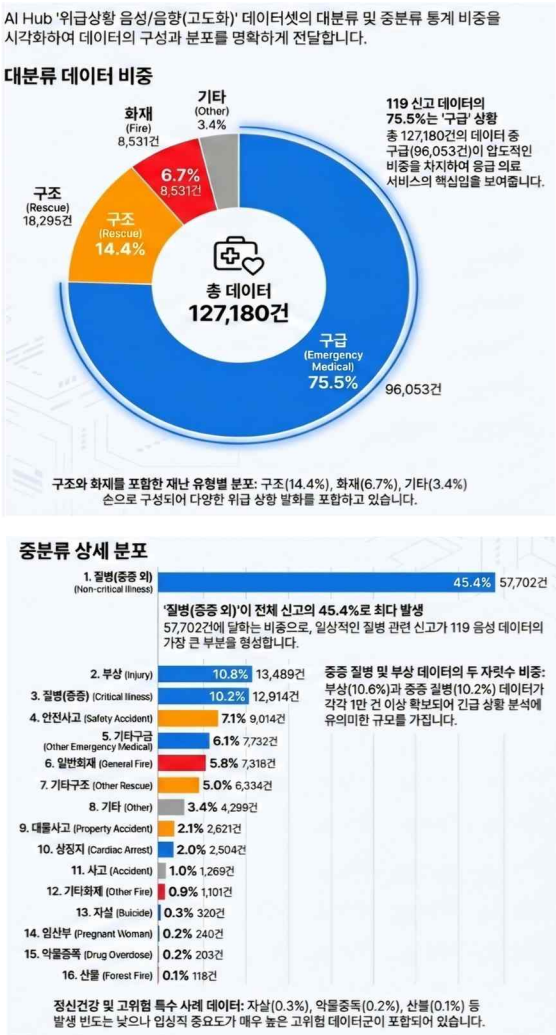
그러나 기존의 119 신고 데이터 분석 및 환자 분류 체계는 발화에서 표면적으로 드러나는 증상(symptom)이나 단순 결과 중심의 응급도(urgencyLevel) 라벨링에 의존하는 경향이 짙다. 기존 응급 환자 분류 체계(KTAS) 역시 호흡, 맥박 등 신체적 생체 징후 중심으로 중증도를 평가하여 정신과적 응급 위기에 대한 맥락과 위험성을 온전히 담아내지 못한다.

이때, 30초에서 180초 사이의 극심한 시간 압박 속에서 수집되는 현재의 비정형 119 신고 데이터는 환자의 명시적인 '자살 의도(Intent)'나 구체적인 '준비 행동(Behavior)'을 임상적으로 구분하지 못한다. 비자살적 자해, 단순 약물 과다 복용, 명백한 자살 시도 등이 환자의 실제 의도와 무관하게 '자살' 또는 '약물중독'이라는 단일 현상 범주로 묶여 처리되는 구조적 라벨 노이즈(Label Noise)를 유발하고 있다.

이러한 파편화된 데이터와 정보 소실은 주관적 직관 개입에 따른 분류 오류와 응급의료진 간 정보 일치도 저하라는 한계로 이어진다. [그림 1]는 AI Hub에서 '위급상황 음성/음향(고도화) 분야의 '119 지능형 신고접수 음성 인식 데이터셋'을 분석하고 구분한 내용이다.

[그림 1]은 Aihub의 음성수보 데이터 셋을 처리한 결과이다. 분석한 [그림 1]의 결과로 119 수보 음성의 응급 신고를 접수한 자살 관련 신고나 고위험 정신적 위기 상황에서는 단순 현상 위주의 정보만으로 환자의 임상적 위험도를 추정하고 골든 타임 내에 응급 자원을 효율적으로 배분하기 어렵다.

그림 1. AI Hub '위급상황 음성/음향 (고도화) 데이터 셋 분석결과 119 신고접수 음성 인식 데이터 셋' 대분류/중분류 결과분석



[그림 1]에서 여성과 청소년·청년층에서 증가폭이 크게 나타났다. 이는 응급 단계에서 정신위기(마음위기) 환자에 대

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 한신대학교가 수행된 경기 지역혁신중심대학지원사업(경기 RISE사업)의 연구결과입니다.

한 조기 선별 체계 구축의 필요성을 보여준다.

본 연구는 구두점 보존 및 형태소 어간 추출 등 특화된 전처리 모듈을 통해 비정형 구어체 노이즈를 통제하고, 신고의 임상적 상황 맥락을 선제적으로 반영할 수 있는 자동 분류 프레임워크를 제안한다. AI-Hub 119 응급 음성 데이터셋의 한계를 보완하기 위해 정신의학분야의 표준 평가 도구인 Columbia Classification Algorithm of Suicide Assessment (C-CASA) 분류 체계를 119 현장 특성에 맞게 8단계로 재구성하여 적용한다.

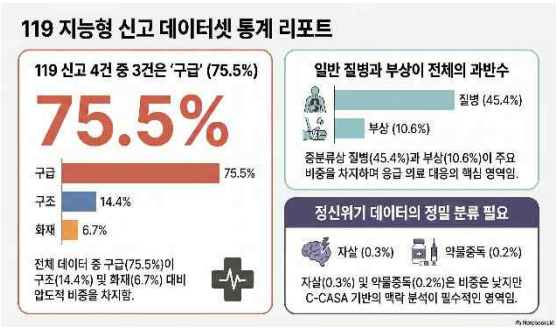


그림 2. AI-Hub 119 응급 음성 데이터셋 기반 재난 유형 분포와 정신응급 데이터의 정밀 분류 필요성

본 제안 모델은 비정형 텍스트에서 선정한 5대 임상 변수(Accident, Behavior, Fatality, Interrupted, Intent)를 다단계 정규식으로 추출하고, 사망 확인 및 구체적 자해 행동 등 물리적 위기 증거를 최우선으로 평가하는 위험 기반(Risk-First) 결정 트리 로직을 통해 신고자의 의도 단계와 다층적 위험 수준을 체계적으로 해석한다. 결과적으로 파편화된 119 비정형 기록을 임상적으로 신뢰할 수 있는 객관적 중증도 지표로 변환함으로써, 저위험군과 고위험군을 명확히 구분하고 구급대원의 의사결정 지원 및 정신응급 환자 이송의 효율성을 극대화하는 것을 목적으로 한다.

II. 응급 발화자 마음위기 선별/분류 연구

2.1 정신과 응급 위기개입과 초기선별 한계

최근 국내에서도 정신과적 응급상황에 대한 체계적인 대응과 데이터 기반 분석의 중요성이 대두되고 있다. 전북특별자치도정신건강복지센터의 3개년 정신위기개입 응급출동 데이터를 분석한 연구에 따르면, 정신위기개입 주요 대상 중 '자살상담'이 전체의 79.3%를 차지하여 '일반 정신응급'(20.7%)보다 압도적으로 높은 비중을 보였다. 특히 대상자의 67.2%가 구체적인 자살 계획이나 시도력을 보유한 고위험군이었으며, 응급개입의 긴급성이 높은 사례(위기 분류 A~C 유형)가 73.9%에 달했다. 여기에서 주목할 만한 점은 이러한 정신과적 응급출동의 의뢰 경로 중 소방 및 경찰 등 병원 전 단계의 핵심 행정기관이 전체의 90.4%를 차지한다는 것이다[1]. 또한 병원 전 단계 119 구급일지 기반 분석을 통해 기존 연구들이 병원 이송 환자 중심으로 수행되어 왔으며, 실제 병원 전 단계 정보의 중요성이 충분히 반영되지 못했다고 지적된 바 있다[4]. 이는 119 응급 신고 시스템이 정신위기 대상자를 최초로 선별하고 초기 대응을 결정하는 가장 중요한 관문임을 시사한다.

그러나 현재의 정신위기 평가는 주로 위기분류척도(CTRS)나 정신과적 위기분류 평정척도(CRI) 등을 활용하여, 현장에 출동한 정신건강전문요원의 수동적인 사후 면담에 의존하고 있다. 이러한 대면 중심의 평가 방식은 119 수보 단계에서의 극심한 시간 압박 속에서 환자의 중증도(Triage)를 선제적으로 분류하는 데에는 적용하기 어렵다는 구조적 한계를 지닌다. 즉, 접수 초기 단계인 '신고 발화' 시점에서 대상자의 자살 의도(Intent)와 물리적 행동(Behavior)을 신속하고 객관적으로 추출할 수 있는 정보학적 접근이 부재한 상황이다.

응급실 의료진이 수용 환자를 진료할 때, 이송 전 구급대원으로부터 제공받은 환자 정보와의 일치도가 낮게 나타나(3.9점) 병원 전 단계와 병원 단계 간 정보 연계에 심각한 단절이 존재함이 확인되었다.[2] 이는 기존 체계가 신체적 활력징후(Vital sign) 위주로 편중되어 있어, 정신과적 위기

상황의 복합적인 임상 맥락을 임상 맥락 기반 분류 로직이 필요함을 시사한다

2.2 Columbia Classification Algorithm of Suicide Assessment (C-CASA)

Posner et al. (2007) 은 미 식품의약국(FDA)의 항우울제 임상시험 데이터에서 자살 관련 부작용이 일관성 없이 기록되는 문제를 해결하기 위해 표준화된 자살 평가 시스템인 Columbia Classification Algorithm of Suicide Assessment (C-CASA)를 제안하였다[3]. 본 연구는 이러한 C-CASA의 임상적 신뢰성과 '의도 및 행동 기반의 구조화 추론 로직'이 119 비정형 발화 데이터 분석에 최적의 틀을 제공한다고 판단하였다.

기존 C-CASA가 주로 임상시험이라는 사후적이고 통제된 환경을 위해 개발된 반면, 본 연구는 이를 시간 압박과 정보 소실이 극심한 병원 전(Pre-hospital) 응급 신고 현장에 실시간 자연어 처리(NLP) 프레임워크로 최초 이식한다는 점에서 차별성을 갖는다. 119 텍스트에서 추출되는 5대 임상 변수를 C-CASA의 8단계 범주에 능동적으로 매핑함으로써, 현장 구급대원들이 단순 자해와 실제 자살 시도를 명확히 구분하고 환자의 중증도를 객관화할 수 있는 객관적 의사결정 지원 시스템으로 기능할 수 있다.

III. 119맞춤형 C-CASA 8단계 분류/적용

본 연구에서는 119 수보 기록 발화 텍스트의 임상적 맥락을 자동 분석하여 정신응급 환자의 중증도를 객관적으로 분류하는 C-CASA 8단계 자동 분류 파이프라인을 제안한다. 제안하는 시스템 아키텍처는 시각 자료에 제시된 바와 같이 데이터 입력, 자연어 전처리, 임상 변수 추출, 결정 트리 기반 분류, 그리고 맥락 보존을 위한 결과 병합의 5단계 모듈형 파이프라인으로 구성된다.

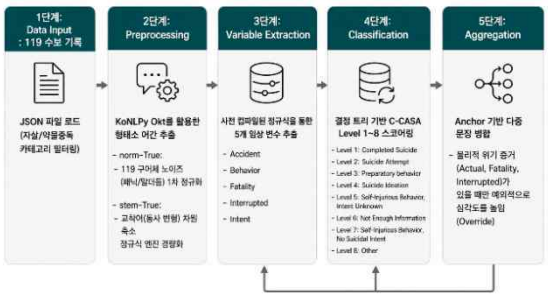


그림 3. 제안하는 C-CASA 기반 119 정신과적 응급신고 맥락분류 자동화 시스템 구조도 제시

3.1. 응급음성 수보데이터 입력과 필터링

본 파이프라인의 첫 단계는 119 수보 기록 원본 데이터를 시스템에 입력하여 분석 대상을 선별 과정이다. 원시 데이터 형식인 JSON 파일을 로드한 후, 재난 대분류 체계상 '자살' 또는 '약물중독' 카테고리로 운영 태깅된 레코드만을 필터링한다. 이를 통해 불필요한 연산 낭비를 방지하고 정신의학적 응급 상황에 특화된 데이터셋을 구축하여 후속 분석의 집중도를 높인다.

3.2. 노이즈 통제 위한 형태소 어간추출

119 수보 음성데이터는 긴박한 병원 전(Pre-hospital) 상황에서 수집된 발화 텍스트는 신고자의 패닉, 말더듬, 불규칙한 어순 등 심한 구어체 노이즈를 동반한다. 이를 제어하기 위해 KoNLPy 라이브러리의 Okt(Open Korean Text) 형태소 분석기를 도입하였다.

먼저 norm=True 파라미터를 적용하여 일그러진 비정형 구어체 텍스트를 1차적으로 교정 및 정규화한다. 이어 stem=True 파라미터를 통해 한국어 특유의 복잡한 동사 및 형용사 변형(교착어 특성)을 단일 기본형으로 복원하는 차원 축소를 수행한다. 이러한 어간 추출 전처리는 후속 정규

식 엔진의 복잡도를 획기적으로 낮춰 연산을 경량화하고 시스템의 키워드 탐지율을 극대화하는 핵심적인 역할을 수행한다.

선행 연구에서 응급실 인터뷰 언어 데이터를 기반으로 자연어처리(NLP)와 머신러닝을 활용한 자살 위험 예측모델의 가능성을 제시하였다[7]. 이는 신고 발화 기반 정신응급 위험도 분석의 실효성을 뒷받침한다.

3.3. 정규 표현식의 5대 임상 변수 추출

전처리 모듈을 통과한 텍스트는 도메인 지식을 바탕으로 사전 컴파일된(Pre-compiled) 정규표현식 계층을 거치며 5대 임상 변수로 치환된다. 추출되는 [그림 3]의 변수는 다음과 같다.

첫째, 사고 여부와 고의성을 판별하는 Accident,

둘째, 물리적 자해 및 약물 과다 복용 등 환자의 구체적 행동 양상을 평가하는 Behavior,

셋째, 심정지 및 생존 징후 소실 등 확정적 사망 상태를 파악하는 Fatality,

넷째, 제3자의 물리적 저지 등 자살 시도 직전의 상황을 포착하는 Interrupted,

다섯째, 자살 혹은 생존에 대한 신고자의 명시적 및 정황적 의사를 나타내는 Intent이다.

이는 실제 임상적 위기 수준을 계량화 근거가 된다

.

임상 변수 및 영문 명칭	임상적 정의 및 역할	정규식 탐지 및 발화 키워드 예시 (examples)
사고 여부 Accident	신고 상황의 고의성 여부와 발생 경위를 판별함.	"자전거 타고 가다가 갑자기", "차량 단독사고"
물리적 행동 Behavior	신체적 자해 및 약물 과다 복용 등 구체적인 가해 행동 양상을 평가함.	"수면제를 많이 먹어서", "손목을 그었는데요"
사망 징후 Fatality	심정지, 생존 징후 소실 등 현장 상태 확정적 사망에 준하는지 파악함.	"숨을 아예 못 쉬어요", "의식이 전혀 없으세요"
행동 저지 Interrupted	위험 장소 진입이나 제3자의 물리적 제지 등 자살 실행 직전 상황을 포착함.	"지금 붙잡고 계시거든요", "지구대에 보호조치"
명시적 의도 Intent	자살 혹은 생존에 대한 신고자의 주관적이고 명시적인 의사를 판단함.	"자살 충동 때문에", "살려 주세요", "죽고 싶어"

그림 4. 정규 표현식 기반 5대 임상 변수 정의 및 추출 기준 예시 분석

3.4. 결정트리 기반 C-CASA 8단계분류

추출된 5개의 임상 변수는 임상 지식 기반으로 설계된 결정 트리(Decision Tree) 알고리즘에 입력되어 최종적인 응급도 계층으로 스코어링된다.

평가는 정신의학 분야의 표준화된 도구인 C-CASA (Columbia Classification Algorithm of Suicide Assessment)를 119 현장에 맞게 최적화한 Level 1~8 체계로 이루어진다.

1. Completed Suicide (완수된 자살),
2. Suicide Attempt (자살 시도),
3. Preparatory behavior (준비 행동),
4. Suicide Ideation (자살 생각),
5. Self-Injurious Behavior, Intent Unknown (의도 불명의 자해 행동),
6. Not Enough Information (정보 불충분),
7. Self-Injurious Behavior, No Suicidal Intent (자살 의도 없는 자해 행동),
8. Other (기타)로 세분화된다.

이를 통해 단순 증상 위주의 기존 분류를 탈피하고 다층적 위험도를 체계적으로 평가한다.

C-CASA 계층	중증도 범주 명칭	119 현장 특성별 발화 맥락 및 판단 기준
Level 1	Completed Suicide (완수된 자살)	사망 징후 변수가 확정적으로 탐지되어 소생 가능성이 희박한 상태
Level 2	Suicide Attempt (자살 시도)	명확한 자살 의도를 가지고 신체적 손상을 가하는 물리적 행동을 실행함
Level 3	Preparatory behavior (준비 행동)	실제 신체 손상은 없으나 자해 도구를 확보하거나 위험 장소에 진입한 단계
Level 4	Suicide Ideation (자살 생각)	물리적인 자해 행동은 없으나, 주관적인 자살 충동이나 의도를 명시적으로 표출함
Level 5	Self-Injurious Behavior, Intent Unknown (의도 불명의 자해)	신체 손상이나 약물 복용 행동은 명확히 존재하지만, 자살하려는 고의성 여부가 모호한 상태
Level 6	Not Enough Information (정보 불충분)	발화 데이터의 정보량이 결여되어 5대 임상 변수를 통한 판단 및 분류가 불가능한 경우
Level 7	Self-Injurious Behavior, No Suicidal Intent (비자살적 자해)	환자가 평소 습관적으로 자해를 반복하는 양상을 보이며 명백한 자살 의도가 결여된 손상 상태
Level 8	Other (기타 정신응급)	자살, 자해 위험은 없으나 조현병 발작, 난동, 공황장애 등 급성 정신과적 구급 처치가 필요한 맥락

그림 5. 결정 트리 기반 119 맞춤형 C-CASA 8단계별 스코어링 체계의 분류와 판단기준 제시

3.5. Anchor 기반 수보음성의 다중문장 병합

119 응급음성의 수보 기록은 대개 다중 문장

으로 구성되므로, 시스템이 단일 문장의 심각도를 무조건적으로 취합할 경우 신고자의 맥락적 의도 철회나 부정(Denial)이 무시되는 과잉 분류(Over-triage)가 발생할 위험이 크다.

이를 방지하기 위해 [그림5]처럼 전체 발화 텍스트 기반의 Anchor 판정 레벨을 도출하여 기본값으로 보존하는 병합 전략을 도입하였다.

과잉 분류 방지를 위해, 개별 문장 평가 결과가 전체 맥락(Anchor)보다 더 심각하게 산출되더라도 무조건 상향하지 않으며, 오직 해당 문장에 명확한 '물리적 위기 증거(Actual, Fatality, Interrupted)'가 존재할 때만 예외적으로 심각도를 상향 조정(Override)하는 원칙을 엄격하게 적용한다.

그림 6. Anchor 기반 119 수보 응급음성의 다중문장 병합 알고리즘의 적용 예시 및 의사결정 흐름도

IV. 결론

본 연구는 병원 전(Pre-hospital) 응급 신고 단계에서 정신과적 응급 환자의 임상적 맥락을 파악하고 객관적인 중증도를 산출하기 위한 'C-CASA 8단계 자동 분류 프레임워크'를 제안하였다.

제시된 시스템은 119 수보 기록의 비정형 구어체 노이즈를 형태소 어간 추출(KoNLPy Okt)을 통해 통제하고, 사전 컴파일된 정규식 기반의 5대 임상 변수 추출 및 위험 기반(Risk-First) 결정 트리 로직을 적용하여 기존의 단순 현상 중심 라벨링이 지닌 구조적 한계를 극복하였다.

특별히, 텍스트 처리 과정에서 Anchor 기반의 다중 문장 병합 전략을 도입함으로써 신고자의 의도 철회나 명시적 부정(Explicit Denial) 맥락을 보존하여, 주관적 직관에 의해 발생할 수 있는 과잉 분류(Over-triage) 현상을 효과적으로 방어하고 실무적 의사결정 지원 도구로서의 가능성을 입증하였다.

그러나 본 연구의 파이프라인은 다음과 같은 한계점을 지닌다.

첫째, 전문가 지식에 기반한 규칙기반(Rule-based) 사전매칭에 의존하므로, 자연어를 온전히 포착하는 데 한계가 존재한다.

둘째, 실전 환경에서 오탈자가 발생할 경우, 형태소 분석 및 정규식 매칭이 연쇄적으로 실패하여 위음성을 유발할 취약성이 있다.

셋째, Anchor 기반 병합 로직이 문장 간 맥락은 방어하나, "어제 시도했지만 지금은 멀쩡하다"와 같이 위기 상황이 해소된 시계열적 서사를 완벽히 추론하여 반영하기에는 알고리즘적 제약이 따른다.

이러한 한계를 극복하기 위한 추후 과제는 다음과 같다. 단어 단위 매칭의 한계와 시계열적 맥락 추론의 제약을 해결하기 위해, 본 결정 트리 엔진에 거대언어모델(LLM)을 결합한 하이브리드 다중 에이전트(Multi-Agent) 아키텍처로 고도화할 계획이다. 선행 연구에서 텍스트 기반 위기 상담 데이터에서 NLP 기반 머신러닝 모델이 자살 위험도를 효과적으로 탐지할 수 있음을 보고하였다[8].

아울러 응급 상황 특유의 STT 노이즈에 강건한 도메인 특화 전처리 모델을 추가하고, 향후 실제 국가응급진료정보망(NEDIS) 데이터 등과 연계하여 산출된 C-CASA 중증도 지표의 임상적 타당성과 환자 예후를 후향적으로 검증하는 후속 연구를 수행이 필요하다.

참고문헌

[1] 한진우 외, “전북특별자치도 정신위기개입 응급출동 3개년 분석” 신경정신의학. Vol. 63, No. 4, pp. 260-270, Dec. 2024.

[2] 박준범 외, “응급이송체계의 문제에 대한 분석과 해결방안 연구” 의료정책연구원 연구보고서. Vol. 2024, No. 17,pp. 1-250, Mar.2025

[3] Posner K, Oquendo MA, Gould M, Stanley B, Davies M. “Columbia Classification Algorithm of Suicide Assessment (C-CASA): Classification of Suicidal Events in the FDA’s Pediatric Suicidal Risk Analysis of Antidepressants,” American Journal of Psychiatry, Vol. 164, No. 7, pp. 1035-1043, July. 2007.

[4]김기환 외, “일개 도 지역 자살 시도 환자에 대한 병원전 단계 분석,” 대한응급의학회지, Vol. 29, No. 6, pp. 687-698, Dec. 2018.

[5] 강창신 외, “한국형 병원전 응급환자 분류도구 시범사업 시행결과,” 대한응급의학회지, Vol. 35, No. 1, pp. 6-15, Feb. 2024.

[6] Kim S, Sung HK, Kim T, Ko SK, Kim SJ, Lee JH, “Trends in Emergency Department Visits for Suicide Attempts Before and During the COVID-19 Pandemic in Korea: A Nationwide Study, 2016-2021,” SSRN Electronic Journal, pp. 1-20, 2022.

[7] Cohen J, Wright-Berryman J, Rohlfs L, Trocinski D, Daniel L, Klatt TW, “Integration and Validation of a Natural Language Processing Machine Learning Suicide Risk Prediction Model Based on Open-Ended Interview Language in the Emergency Department,” Frontiers in Digital Health, Vol. 4, Article 818705, Feb. 2022.

[8] Broadbet M, Medina Grespan M, Axford K, Zhang X, Srikumar V, Kious B, Imel Z, “A Machine Learning Approach to Identifying Suicide Risk among Text-Based Crisis Counseling Encounters,” Frontiers in Psychiatry, Vol. 14, Article 1110527, Mar. 2023.

Sentinel AI: 규칙 기반 엔진과 멀티모달 Local LLM을 결합한 시니어 실신·배회 탐지 지능형 관제 시스템

류찬*(발표자), 성낙준*
*한신대학교 AI·SW대학
e-mail: {lyuchan, njsung}@hs.ac.kr

Sentinel AI: An Intelligent Monitoring System for Senior Syncope and Wandering Detection Combining a Rule Engine and Multimodal Local LLM

Chan Lyu*, Nak-Jun Sung*
*School of Computing and Artificial Intelligence, Hanshin University

요 약

본 연구는 요양시설·병동·복지시설 등에서 고령자와 환자의 실신 및 배회와 같은 이상 행동을 조기에 탐지하기 위한 온프레미스형 지능형 관제 시스템 Sentinel AI를 제안한다. 제안하는 Sentinel AI는 실시간 영상에서 YOLO11n과 ByteTrack으로 사람을 검출·추적하고, MediaPipe Pose Landmarker를 이용해 자세 특징점을 추출한 뒤, 공공 데이터를 활용해 학습한 규칙 기반 엔진을 통해 실신·배회 여부를 판단한다. 엔진을 통해 이상 상황이 감지되면 멀티모달 Local LLM 모델인 Gemma4를 이용해 대상자의 연령대, 성별, 행동, 위치 등을 포함한 구조화된 설명을 생성하여 관제 대시보드에 제공한다. 실험 결과 Sentinel AI는 이상 상황 감지와 관제자의 1차 판단 지원에 효과적임을 확인하였으며, 개인정보 보호와 현장 적용 가능성을 갖춘 시스템임을 확인하였다.

키워드: 지능형 관제 시스템, 실신 및 배회 감지, 멀티모달, Local LLM, Rule Engine, 온프레미스

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

I. 서론

고령자와 환자가 생활하는 요양시설, 병동, 복지시설 등에서는 실신, 배회와 같은 사람의 이상 행동을 조기에 탐지하고 신속히 대응하는 것이 중요하다. 그러나 기존 활용되는 CCTV 관제 방식은 사람에 의한 지속적인 모니터링에 의존하는 방식으로, 실시간 대응성과 상황 이해 측면의 한계가 존재한다. 따라서 이상행동을 자동으로 감지하고 관제자의 판단을 지원할 수 있는 지능형 관제 시스템이 필요하다. 동시에 요양시설과 병동의 영상 데이터는 민감한 개인정보를 포함하므로, 외부 전송 없이 시설 내부에서 처리할 수 있는 온프레미스 방식이 요구된다.
본 연구에서는 이를 해결하기 위해 멀티모달 인공지능 모델과 공공데이터를 활용한 온프레미스형 지능형 관제 시스템 Sentinel AI를 제안한다. 제안하는 Sentinel AI는 실신과 배회의 행동적 특성을 구분하여 이상 상황을 판정하고, 감지된 이벤트에 대한 구조화된 상황 설명을 제공함으로써 기존 인력 중심 CCTV 관제의 한계를 보완하고자 한다.

II. 관련 연구

영상 기반 이상행동 탐지는 CCTV 영상에서 사람의 위치, 자세, 이동 궤적을 분석하여 비정상 상황을 탐지하는 방향으로 연구되어 왔다. 실신 및 낙상 감지 연구에서 객체 검출과 자세 추정을 결합하여 분석하는 방법이 활용되었으며[1], 배회 감지 연구에서는 시간에 따른 이동 궤적 이력을 기반으로 움직임 패턴을 분석하는 접근이 제안되었다[2]. 최근에는 VLM(Vision Language Model)과 LLM(Large Language Model)을 활용하여 영상 장면을 텍스트로 변환하고, 이를 바탕으로 이상상황을 탐지하려는 연구도 활발히 진행되고 있다[3].

III. 제안방법

3.1 전체 시스템 구조

본 연구에서 제안하는 Sentinel AI는 고령자와 환자의 실신 및 배회 상황을 조기에 탐지하고, 관제자의 1차 판단을 지원하기 위한 온프레미스형 지능형 관제 시스템이다. 그림 1은 Sentinel AI의 전체 시스템 구조를 나타낸다. Sentinel AI는 영상 입력, 객체 검출 및 자세 추정, 규칙 기반 엔진을 활용한 이상행동 판정, Local LLM을 활용한 상황 설명 생성, 관제 대시보드 표시의 5단계로 구성된다.

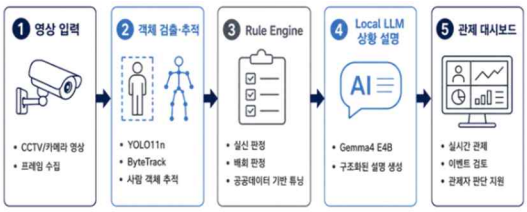


그림 1. Sentinel AI 시스템 구조

입력된 CCTV 영상은 프레임 단위로 처리되며, 이후 객체 검출·추적 모듈과 자세 추정 모듈을 통해 사람 단위의 데이터로 변환된다. 변환된 데이터는 규칙 기반 엔진으로 전달되어 실신 및 배회 여부를 판정하고, 이상 이벤트가 감지되면 Local LLM이 해당 이벤트에 대한 구조화된 상황 설명을 생성한다. 마지막으로 관제 대시보드는 이벤트 영상, 스냅샷, 감지 결과, 상황 설명을 함께 제공하여 관제자가 신속하게 상황을 검토할 수 있도록 한다. Local LLM은

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

3.2 객체 검출 및 자세 추정

객체 검출을 위해 널리 사용되는 YOLO 계열의 모델 중 경량화된 구조를 가지는 YOLO11n을 이용하여 사람 객체를 검출한다. 이후 ByteTrack을 이용하여 동일 인물의 이동 흐름을 추적한다. 검출된 사람 영역에 대해 MediaPipe Pose Landmarker를 적용하여 신체 특징점을 추출한다. 이를 통해 몸통 각도, 자세 변화량, landmark 평균 이동량 등을 계산한다.

배회 감지는 자세보다 이동 경로가 중요하므로, 사람 bounding box의 하단 중앙점을 이동 기준점으로 사용하고, 동일 track ID의 위치 이력을 누적하여 반복 이동 패턴을 분석한다.

3.3 규칙 기반 엔진의 실신·배회 판별

규칙 기반 엔진은 객체 검출 및 추적 단계에서 생성된 사람의 track ID, bounding box, 자세 landmark, 위치 이력을 입력으로 받아 실신 및 배회 여부를 판별한다. 본 연구에서는 실신과 배회 행동의 판별 기준을 정의하기 위해 AI-Hub의 이상행동 CCTV 영상 데이터셋을 활용하였다. 해당 데이터셋은 실신, 배회, 정상 행동 구간을 포함하고 있으며, 본 연구에서는 각 구간에서 관찰되는 자세 변화, 위치 이동, 체류 시간 등의 특징을 기반으로 구간별 판별 알고리즘을 구성하였다.

그림 2. 규칙 기반 엔진의 실신·배회 판별 과정

실신 판별은 짧은 시간 내 발생하는 자세의 붕괴와 이후 움직임의 지속 여부를 중심으로 동작한다. 중심점 하강, 몸통 각도 변화, 수평 자세 조건, 자세 landmark의 평균 이동량을 함께 확인하여 이벤트를 발생시킨다. 배회 판별은 사람 객체의 bounding box 하단 중앙점을 이동 기준점으로 사용하고, 동일 track ID의 위치 이력을 시간에 따라 누적하여 수행된다. 같은 영역에 머물면서 방향 전환과 반복 이동이 나타나고, 총 이동거리에 비해 시작점과 종료점 사이의 변위가 작은 경우를 배회 의심 상황으로 판단한다.

3.4 Local LLM 기반 상황 설명 생성

규칙 기반 엔진을 통해 실신과 배회 이벤트가 감지되면, 시스템은 해당 이벤트의 대표 이미지와 메타데이터를 추출하여 온프레미스 환경에서 동작하는 멀티모달 Local LLM인 Gemma4-E4B 모델에 전달하고, 구조화된 상황 설명을 생성한다. 해당 모델은 Apple M5, 10GB Unified Memory 환경에서 추론을 수행한다. 이를 통해 Sentinel AI는 감지된 이상 상황에 대해 대상자의 행동, 위치, 상태, 위험 가능성 등을 텍스트 형태로 요약하여 관제자가 상황을 빠르게 이해하고 1차 판단을 수행할 수 있도록 지원한다.

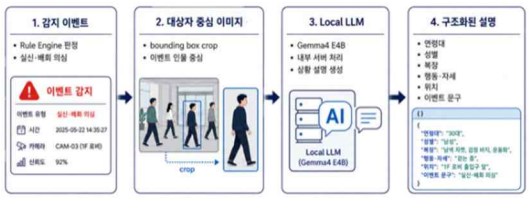


그림 3. Local LLM을 활용한 상황 설명 생성 과정

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

상황 설명 생성 결과는 대상자의 연령대, 성별, 복장, 행동 및 자세, 위치, 이벤트 문구 등을 포함한다. 인파가 많은 장면에서는 이벤트가 발생한 인물을 관제자가 즉시 식별하기 어려울 수 있으므로, 대상자 중심의 설명은 현장 확인과 신고 판단을 보조하는 요소로 활용될 수 있다. 또한 LLM 입력에는 이벤트 대상자의 bounding box 영역을 크롭한 이미지가 사용되어, 여러 사람이 함께 등장하는 상황에서 LLM이 다른 인물을 설명하는 문제를 방지했다.

3.5 관제 대시보드

대시보드는 실시간 관제 화면과 이벤트 검토 화면을 중심으로 구성된다. 실시간 관제 화면에서는 실시간 CCTV, 이벤트 발생 여부 같은 정보를 함께 표시한다. 이상 이벤트가 발생하면 이벤트 영상, LLM 상황 설명을 함께 제공하여 관제자가 즉시 상황을 파악할 수 있도록 한다. 이때 이벤트 발생 알림과 LLM 상황 설명은 비동기로 처리하여 상황 설명 생성이 늦어지더라도 관제자가 상황을 늦게 인지하게 되는 위험을 방지했다.

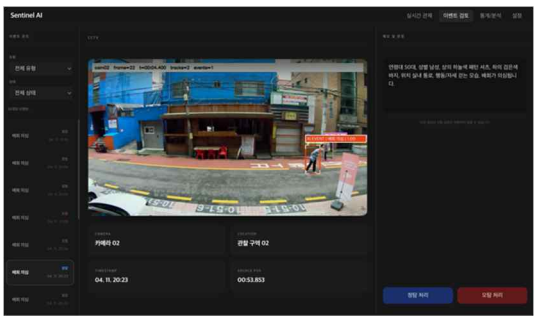


그림 4. 대시보드 이벤트 검토 화면

이벤트 검토 화면에서는 누적된 실신·배회 의심 이벤트를 관제자가 확인할 수 있다. 관제자는 이벤트 유형, 상태, 카메라 기준으로 목록을 확인하고, 각 이벤트의 영상 클립, 스냅샷, 구조화된 상황 설명을 검토한 뒤 정탐 또는 오탐 여부를 기록할 수 있다.

그림 5. 대시보드 통계/분석 화면

통계/분석 화면에서는 시스템 운영 현황과 감지 품질을 확인할 수 있다. 유형별·상태별·카메라별 분포, 시간대별 추세 등을 제공한다.

IV. 결론

본 연구에서는 실신 및 배회 의심 상황을 감지하고, 관제자의 1차 판단을 지원하기 위한 온프레미스형 지능형 관제 시스템 Sentinel AI를 제안하였다. Sentinel AI는 영상 기반 시니어 이상 행동 탐지 기술과 멀티모달 LLM을 활용한 구조화된 상황 설명 기능을 결합함으로써 단순한 감지를 넘어 관제자가 상황을 이해하고 판단하는 데 필요한 정보를 제공할 수 있음을 확인하였다.

참고문헌

[1] E. Tîrziu, A.-M. Vasilevschi, A. Alexandru, and E. Tudora, “Enhanced Fall Detection Using YOLOv7-W6-Pose for Real-Time Elderly Monitoring,” *Future Internet*, Vol. 16, No. 12, 472, 2024.

[2] Khaertdinov, Bulat, Yusuf Can Semerci, and Stylianos Asteriadis. "Dementia wandering recognition using classical machine learning and deep learning techniques with skeletal trajectories." *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference*. 2021.

[3] L. Zanella et al., “Harnessing Large Language Models for Training-free Video Anomaly Detection,” *Proceedings*

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

2026 한국데이터사이언스학회 하계 종합학술대회 논문집

공공데이터 융합 기반 고령층 의료 이동 취약지역 발굴을 위한 K-MMVI 프레임워크

신현진*, 김병재O
*한신대학교 소프트웨어융합학부 O한신대학교 응용통계학과
e-mail : rlaqudwo@hs.ac.kr, shjj8769@hs.ac.kr

An Optimization Framework for Korea Medical Mobility Vulnerability Index (K-MMVI) Using Public Data Convergence

Hyeonjin Shin*, Byeongjae KimO
*Dept. of Software Convergence, Hanshin University ODiv. of AI & Software, Hanshin University

요약

본 연구는 초고령화 사회 진입에 따른 지방 소멸 및 지역 간 의료 격차 문제를 해결하기 위해, 인구 통계와 보건의료 접근성, 대중교통 인프라 데이터를 복합적으로 융합한 '한국형 의료 이동 취약성 지수(K-MMVI)' 기반 프레임워크를 제안한다. 기존의 의료 취약지 분류 체계는 공급자 중심의 단순 물리적 직선거리나 행정구역별 의사 수에 의존하여, 실제 고령층이 체감하는 대중교통 배차 간격 및 노선 밀도 등 실질적 이동 제약 맥락을 충분히 반영하지 못하는 한계가 있다. 이를 해결하기 위해 주민등록인구 데이터, 국립중앙의료원 헬스맵 통계, 공공보건의료 통계 데이터를 수집하고 다차원 정규화 알고리즘을 적용하였다. 또한, 지리적 원격성이 극심한 공급 단절 지역을 우선 스크리닝하는 '물리적 차단 선제 플래그(Auxiliary Flagging Rule)' 알고리즘을 설계하여 지수의 왜곡을 제어하고 최종 복합 K-MMVI 스코어를 산출하여 현실적인 공급 사각지대를 보완하고자 한다. 본 연구는 파편화된 공공데이터를 정책 수립자가 즉각 해석 가능한 객관적 취약도 지표로 변환함으로써, 수요응답형 교통(DRT) 노선 최적화 및 찾아가는 의료 복지 자원의 효율적 배분 효율성 향상에 기여할 것으로 기대된다.

Ⅰ. 서론

우리나라는 세계에서 유례없이 빠른 속도로 고령화가 진행되고 있다. 행정안전부 주민등록 인구 기준 2024년 12월 65세 이상 인구가 전체의 20%를 돌파하며 한국은 초고령사회에 진입하였고, 이는 2017년 고령사회 진입 이후 약 7년 4개월 만의 변화로 일본 등 주요국과 비교해도 가장 빠른 속도이다. 이러한 경향이 지속될 시 향후 2036년 30%, 2050년에는 40%를 넘어설 것으로 전망된다. 이러한 인구구조의 변화는 의료 수요의 급격한 증가로 이어지며, 특히 만성질환과 응급상황 발생 빈도가 높은 고령층에게 의료 서비스에 대한 물리적 접근성은 생존과 직결되는 핵심 요소가 된다.

그러나 의료 자원은 지역적으로 균등하게 분포하지 않으며, 고령인구의 분포와도 불일치한다. 상급종합병원과 응급의료기관은 대도시에 집중되어 있는 반면, 고령인구 비율이 높은 농어촌 및 지방 중소도시는 의료기관까지의 거리가 멀고 이를 연결하는 대중교통 또한 취약한 경우가 많다. 이로 인해 고령층이 거주 지역에서 적시에 의료 서비스를 이용하지 못하는 "의료 이동 취약성"이라는 구조적 문제가 발생한다. 이러한 지역을 정확히 식별하는 것은 한정된 보건의료 자원을 효율적으로 배분하고 정책 우선순위를 결정하기 위한 필수 선행 과제이다.

기존의 의료 취약지 분석은 병원 수, 의사 수와 같은 단일 공급 지표나 행정구역 단위의 단순 통계에 의존하는 경향이 있었다. 이러한 접근은 의료 자원의 절대량은 파악할 수 있으나, 고령인구의 분포, 의료기관까지의 실제 이동 여건, 대중교통 연결성 등 복합적 요인이 결합되어 나타나는 '의료 이동 취약성'을 종합적으로 반영하지 못하는 한계를 지닌다. 더욱이 이러한 정보는 행정안전부, 건강보험심사평가원, 국가교통정보센터 등 다양한 출처에서 제각각의 형식과 단위로 산재되어 있어, 이를 의미 있는 정책 지표로 통합하기 어렵다. 즉, 의료 자원이 일정 수준 존재하더라도 고령층이 그 자원에 실제로 도달하기 어려운 지역의 구조적 취약성을 정량적으로 드러내고, 이를 객관적 우선순위로 변환하기 위한 체계적 방법론이 부재한다.

이에 본 연구는 이질적인 공공데이터의 융합을 통해 고령층의 의료 이동 취약성을 정량적으로 측정하는 한국형 의료 이동 취약성 지수(K-MMVI, Korea Medical Mobility Vulnerability Index)와 이를 산출하는 분석 프레임워크를 제안한다. 본 제안 모델은 인구 밀집도, 의료 접근성, 교통 공급량이라는 3대 핵심 차원 변수를 다단계 계층 구조로 추출하고, 정규화(normalization)와 엔트로피 기반 가중합산을 통해 지역별 취약 수준을 정량적으로 산출한다. 정량 산출에 그치지 않고 의료기관 전무 지역이나 대중교통망 단절과 같은 결정적 요인을 사전에 감지하여 최우선 사각지대로 매핑하는 물리적 차단 선제 플래그(Auxiliary Flagging Rule) 메커니즘을 결합함으로써 지역별 취약 수준을 체계적이고 실전적으로 해석한다. 결과적으

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심대학지원사업(경기RISE사업)의 연구결과입니다.

로 파편화된 공공데이터 기록을 통합하여 예산 투입의 최우선 순위 지역을 명확히 구분하고, 수요응답형 복

지 인프라 확충의 효율성을 극대화하는 것을 목적으로 한다. 또한 산출된 지수를 국립중앙의료원 헬스맵 등에서 제공하는 공식 의료취약지 지표와 비교하여 타당성을 검증하고, 가장 취약한 것으로 식별된 지역을 대상으로 읍면동 단위의 심층 분석을 수행하여 지역 내부의 미시적 격차를 규명한다.

Ⅱ. 관련 연구

2.1 국내 의료 취약지 현황과 기존 평가 방식의 한계

국내에서 의료 접근성의 지역 간 불균형이 심화되면서, 이를 데이터에 기반해 진단하려는 노력이 이어지고 있다. 보건복지부와 국립중앙의료원은 의료취약지 모니터링 연구를 통해 응급·분만·소아청소년과 등 분야별 의료취약지를 선정하고, 그 결과를 헬스맵(Health Map) 서비스로 공개하고 있다 [1]~[2]. 이러한 취약지 지정은 관내 의료기관까지 일정 시간 내에 도달 가능한 인구의 비율 등을 기준으로 하며, 응급의료를 비롯한 필수의료의 공백이 농어촌 및 지방 중소도시에 집중되어 있음을 보여준다.

한편 지역 간 의료 격차의 실태도 여러 통계를 통해 확인된다. 국가통계연구원의 분석에 따르면, 우리나라의 의료시설과 의료인력은 지역 간 및 의료기관 유형 간에 불균형한 분포를 보이며, 비수도권은 병원 수와 병상 수에 비해 의사 수가 상대적으로 부족한 것으로 나타났다 [3]. 이는 단순한 시설의 양적 지표만으로는 지역 주민이 실제로 체감하는 의료 접근성을 충분히 설명하기 어렵다는 점을 시사한다.

그러나 현재의 취약지 평가는 행정구역별 의사 수나 병상 수와 같은 공급자 중심의 단편적 인프라 통계에 의존하는 경향이 있으며, 정기 조사 역시 사후적으로 수행되는 한계가 있다. 이러한 방식은 고령인구의 미시적 분포, 의료기관까지의 실제 이동 거리와 소요 시간, 대중교통 연결성과 같은 수요·공급·이동성의 복합 요인을 유기적으로 통합하지 못한다. 즉, 정책 수립의 초기 단계에서 대상 지역의 실질적 수요(고령인구 위험성)와 물리적 제약(교통 공급량 및 이동 거리)을 객관적으로 결합하여 사각지대를 선제적으로 규명할 수 있는 데이터 융합형 접근이 부재한 상황이다.

2.2 공간 다변량 분석과 데이터 융합을 통한 취약 지역 평가

지역의 취약성을 정량적으로 식별하는 연구 영역에서는 다출처의 이질적 데이터를 통합하여 복합 위험도를 산출하는 공간 다변량 분석이 활발히 적용되어 왔다. 이명호 외는 GIS 공간분석과 취약성 지표를 결합하여 사회 취약지역을 정량적으로 도출하는 방법론을 제시한 바 있다 [4]. 이러한 연구들은 단일 지표가 아닌 다수의 지표를 결합함으로써 지역 취약성을 다차원적으로 평가할 수 있음을 보여준다.

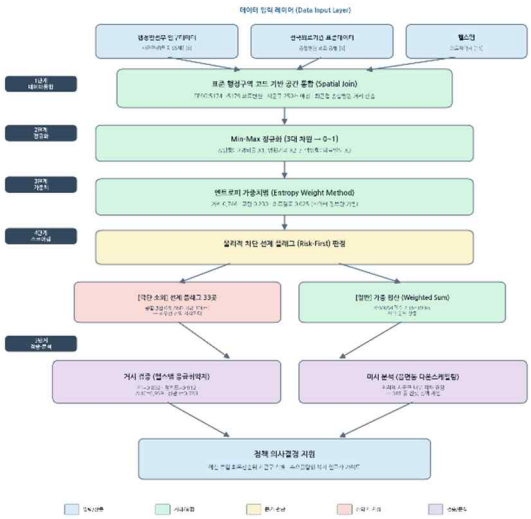
이때 다수의 지표를 하나의 종합 지수로 결합하는 과정에서 각 지표의 가중치를 어떻게 결정하는가가 핵심 쟁점이 된다. 분석가의 주관이 개입하는 가중치 부여 방식은 결과의 객관성을 저해할 수 있는데, 엔트로피 가중치법(Entropy Weight Method)은 각 지표가 지닌 데이터의 분산 정도, 즉 정보량을 바탕으로 가중치를 산출함으로써 이러한 주관적 편향을 최소화하는 방법으로 알려져 있다.

그러나 기존의 GIS 기반 공간분석 연구들은 행정안전부의 주민등록인구, 의료기관 위치 정보, 대중교통 데이터 등 출처와 수집 단위가 상이한 공공데이터를 표준화된 체계로 통합하지 못해 정보의 단절을 야기해 왔다. 본 연구는 이러한 한계를 극복하기 위해, (1) 이질적 공간 데이터를 표준화된 행정구역 코드를 기준으로 통합하는 파이프라인을 구축하고, (2) 인구 위험성, 의료 접근성, 교통 공급량을 K-MMVI의 세 가지 차원으로 매핑하며, (3) 엔트로피 가중치법을 통해 객관적으로 가중합산함으로써 단순 인프라 부족 지역과 교통망 단절형 소외 지역을 정량적으로 구분한다는 점에서 차별성을 가진다.

Ⅲ. 제안방법

본 연구에서는 다출처 공공데이터를 유기적으로 통합하고 고령층이 체감하는 복합적 의료 이동 취약지역을 정밀하게 식별하기 위해 'K-MMVI(한국형 의료 이동 취약성 지수) 5단계 분석 프레임워크'를 제안한다. 제안하는 시스템 아키텍처는 공간적 수요와 공급, 그리고 이동성의 제약 요인을 수학적으로 결합하여 정책적 우선순위를 도출할 수 있도록 모듈형 파이프라인으로 설계되었다. 본 방법론의 전체적인 데이터 흐름과 레이어 구조는 그림 1과 같으며, 구체적으로는 데이터 입력 및 공간 결합, 다차원 정규화, 엔트로피 기반 가중치 산출, 선제 플래그 기반 스코어링, 그리고 공식 지표 검증 및 미시 분석의 5단계 연속적인 모듈로 구성된다.

그림 1. K-MMVI 5단계 분석 파이프라인 아키텍처



3.1 데이터 입력 및 공간 데이터 통합 (1단계)

K-MMVI의 다차원 분석을 구성하기 위해 데이터 입력 레이어(Data Input Layer)에서 세 가지 이질적 공

공데이터셋을 수집한다. 첫째, 인구 위험성 도메인을 위해 행정안전부 주민등록 인구통계 [5]로부터 시군구 및 읍면동 단위별 연령별 인구 데이터를 추출하여 전체 인구 대비 65세 이상 고령인구 비율을 산출한다. 둘째, 의료 원격성 도메인을 위해 보건복지부 전국의료기관 표준데이터 [6]에서 종별코드를 활용해 종합병원급 의료기관의 위치 좌표 및 종별 정보를 필터링한다. 셋째, 검증 기준 및 의료 공급 밀도 보조 지표로 국립중앙의료원 헬스맵(Health Map) [8]의 시군구 단위 의료취약지 지표를 활용한다.

1단계 데이터 통합 모듈에서는 수집된 데이터의 지리정보학적 스케일을 통일하기 위해 표준 행정구역 코드를 공통 키(Common Key)로 설정하여 공간 결합(Spatial Join)을 수행한다. 특히 기관별로 상이한 데이터의 지리적 위치 정보를 단일 좌표체계 위에 투영하기 위해 의료기관의 기준 좌표계를 국내 표준 투영좌표계인 EPSG:5179(UTM-K)로 통일하는 좌표변환(Reprojection)을 수행한다 [7]. 변환이 완료되면 전국 시군구 행정구역 중심점 기반의 공간 매칭을 통해 각 지역별 '최근접 종합병원 거리'와 '관내 종합병원 수'를 도출함으로써 다차원 공간 분석을 위한 베이스라인 데이터셋을 구축한다.

또한 분석의 시점 정합성을 확보하기 위해 인구·행정구역경계·검증지표는 2023년 기준으로 통일하였으며, 의료기관 현황은 과거 시점 자료가 제공되지 않아 가용한 최신 표준데이터를 사용하였다. 분석 대상인 종합병원급 의료기관은 단기간 위치 변동이 미미하여 시점 차이가 결과에 미치는 영향은 제한적이다.

3.2 다차원 변수 정규화 및 엔트로피 가중치 산출 (2~3단계)

2단계 정규화 모듈에서는 측정 척도와 단위가 상이한 공간 지표들을 상호 비교할 수 있도록 식 (1) 및 식 (2)의 Min-Max 정규화(Normalization)를 적용하여 0과 1 사이의 무차원 값으로 변환한다. 이때 지표의 특성에 따라 값이 커질수록 지역의 취약성을 가중시키는 정방향 변수(고령인구 비율 X_1 , 최근접 종합병원 거리 X_2)와 값이 커질수록 취약성을 완화시키는 역방향 변수(면적당 종합병원 수 X_3)를 구분하여 처리한다.

$$X'_j = \frac{X_j - X_{min}}{X_{max} - X_{min}} \quad \text{(정방향 지표)} \qquad (1)$$

$$X'_j = \frac{X_{max} - X_j}{X_{max} - X_{min}} \quad \text{(역방향 지표)} \qquad (2)$$

K-MMVI의 3대 핵심 공간 변수와 그 정책적 정의는 표 1과 같다.

표 1. K-MMVI 3대 핵심 공간 변수

핵심 공간 변수	영문 명칭	정책적 정의 및 역할
고령인구 비율	Demography (X_1)	행정구역 내 잠재적 의료 이용 고위험 수요층의 집중도를 반영함
종합병원 거리	Medical Distance (X_2)	골든타임 내 응급의료기관 도달을 가로막는 물리적 지리 저항을 평가함
의료기관 밀도	Medical Density (X_3)	행정구역 면적 대비 의료 공급의 충분성을 보장하는 인프라 공급량을 파악함

3단계 가중치 산출 모듈에서는 연구자의 주관적 편향이나 가설에 의한 왜곡을 배제하고 데이터가 자체적으로 내포한 정보량에 기반하여 가중치를 최적화하는 엔트로피 가중치법(Entropy Weight Method)을 실행한다. 각 공간 지표의 속성 가치에 대한 통계적 불확정성(Entropy)을 계산하고, 데이터의 분산이 커서 정보 송신량이 많은 지표일수록 더 높은 결합 강도를 가지도록 차원별 최종 가중치(W□)를 통계적으로 결정한다.

3.3 물리적 차단 선제 플래그 및 K-MMVI 복합 스코어링 (4단계)

4단계 스코어링 모듈에서는 다기준 의사결정 모델의 고질적 한계인 '변수 간 상쇄 효과(Compensation Effect)'를 제어하기 위해, 극단적 위험 지역을 사전에 스크리닝하는 물리적 차단 선제 플래그(Auxiliary Flagging Rule) 판정 규칙을 도입한다.

[선제 플래그 판정 규칙] 관내 종합병원급 의료 인프라가 전무하고(X_3 관련 종합병원 수 = 0), 동시에 최근접 거점 병원까지의 지리적 이동 거리가 물리적 한계 임계치(D_thr)를 초과하는 경우.

위 임계 조건을 충족하는 행정구역은 복합 가중합산 점수의 높고 낮음에 관계없이 시스템에 의해 '최우선 취약 사각지대'로 선제 매핑(Flagging)된다. 이를 통해 극단적 의료 공백 지역이 타 지표의 양호함에 의해 은폐되는 현상을 방지한다.

반면, 선제 플래그 조건에 해당하지 않는 일반 지역들은 식 (3)과 같이 정규화된 3대 차원 변수와 엔트로피 분석으로 도출된 최적 가중치의 선형 가중합산(Weighted Sum) 알고리즘을 거쳐 최종 K-MMVI 스코어를 부여받는다. 점수가 1(또는 100)에 가까울수록 고령층의 의료 접근성과 이동성이 결핍된 중증 취약 지역으로 분류된다.

$$K\text{-}MMVI = \sum_{j=1}^{3} (W_j \times X'_j) \qquad (3)$$

3.4 공식 지표 비교 검증 및 미시적 다운스케일링 분석 (5단계)

마지막 5단계 검증 및 응용 모듈은 제안 지수의 신뢰성을 학술적으로 확보하고, 정책적 실천력을 극대화하기 위해 거시와 미시 체계로 이원화하여 전개된다.

첫째, 거시 검증(Macro-level Validation) 단계에서는 본 파이프라인을 통해 산출된 시군구 단위의 K-

MMVI 스코어와 국가 공인 지표인 국립중앙의료원 헬스맵의 응급의료 취약지 지정 현황을 상호 대조한다. 혼동 행렬(Confusion Matrix) 기반의 평가지표인 F1-Score, 분류 정확도(Accuracy), AUC(곡선하면적)를 산출하고 점이연 상관계수(Point-biserial Correlation)를 통해 제안 모델이 실제 정부의 공간 진단과 지니는 통계적 타당성을 검증한다.

둘째, 미시 분석(Micro-level Downscaling) 단계에서는 거시 검증을 통과한 프레임워크를 기반으로 고위험 시군구를 식별한 뒤, 분석의 공간적 단위를 읍면동 수준으로 하향 분할(Downscaling)하는 2차 심층 분석을 수행한다. 거시적 행정구역 통계에 가려져 있던 지자체 내부의 미시적 사각지대를 정밀 규명함으로써, 향후 수요응답형 교통(DRT) 노선 최적화 및 핀셋형 복지 인프라 배치 등 실효성 있는 정책 의사결정 지원(Policy Decision Support) 가이드라인을 도출한다.

Ⅳ. 분석 결과

4.1 K-MMVI 산출 및 엔트로피 가중치

제안한 프레임워크를 전국 250개 시군구에 적용하여 K-MMVI를 산출하였다. 엔트로피 가중치법을 통해 도출된 3대 차원의 가중치는 의료 원격성(병원까지의 거리)이 0.746으로 가장 높았고, 인구 위험성(고령인구 비율) 0.230, 의료 공급 밀도 0.025 순으로 나타났다. 이는 지역 간 의료 이동 취약성의 격차를 결정하는 가장 핵심적인 요인이 의료기관까지의 물리적 거리임을 의미한다. 이는 분석가의 주관이 아니라 데이터의 정보량에 근거하여 도출된 객관적 결과이다. 의료 공급 밀도의 가중치가 상대적으로 낮게 산출된 것은 면적당 종합병원 수가 다수의 농어촌 지역에서 0에 수렴하여 지역 간 변별력이 작았기 때문으로 해석된다.

산출된 K-MMVI는 최소 2.66점에서 최대 89.65점의 분포를 보였으며, 전국 평균은 18.81점이었다.

그림 2는 전국 시군구별 K-MMVI를 단계구분도로 시각화한 것이다. 수도권 및 광역시 도심부는 취약성이 낮게(연한 색) 나타난 반면, 경상북도 북동부 산간, 전라남도 도서 지역, 강원 산간 지역에서 취약성이 높게(진한 색) 분포하여 의료 이동 취약성이 뚜렷한 공간적 편중을 보임을 확인할 수 있다.

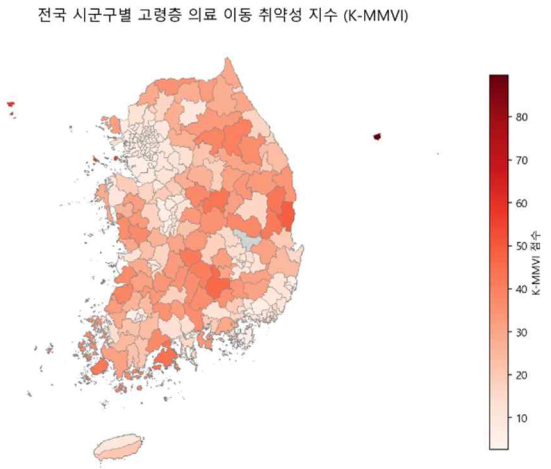


그림 2. 전국 시군구별 K-MMVI 단계구분도

4.2 의료 이동 취약 지역 식별

K-MMVI 점수를 기준으로 산출한 취약성 상위 지역은 그림 3과 같다. 가장 취약한 지역은 경상북도 울릉군(89.65점)으로, 도서 지역 특성상 관내 종합병원이 부재하고 육지의 의료기관까지의 거리가 150km를 초과하여 압도적으로 높은 취약성을 보였다. 이어 옹진군(58.68점), 영덕군(48.73점), 합천군(45.83점), 고흥군(43.54점) 순으로 나타났다. 상위 10개 지역은 모두 헬스맵의 응급의료 취약지 분류와 일치하였으며, 취약 상위 지역 대부분은 고령인구 비율이 높은 동시에 관내 종합병원이 부재한 농어촌·도서 지역으로, 고령 수요와 의료 공급 공백이 중첩된 사각지대임이 확인되었다.

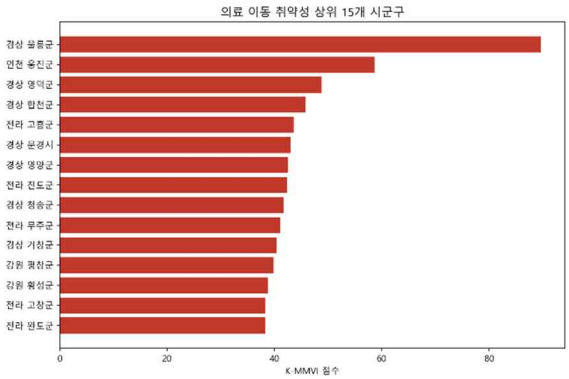


그림 3. K-MMVI 기준 의료 이동 취약성 상위 지역

한편 물리적 차단 선제 플래그(종합병원 부재 및 최근접 병원까지 30km 이상) 조건을 충족하는 극단적 소외 지역은 전국 33개 시군구로 식별되었다. 이들은 복합 가중 점수와 무관하게 최우선 취약 사각지대로 분류되어, 수요응답형 교통(DRT) 등 우선적 정책 개입이 필요한 대상으로 제시된다.

4.3 헬스맵 공식 지표를 통한 타당성 검증

제안 모델의 타당성을 검증하기 위해, 산출된 K-MMVI를 국립중앙의료원 헬스맵의 공식 응급의료 취약지 분류와 비교하였다. 헬스맵이 응급 취약지로 분류한 102개 시군구를 정답으로 설정하고, K-MMVI 상위 102개 지역을 취약지로 예측하여 분류 성능을 평가하였다.

그 결과 정확도(Accuracy) 0.912, 정밀도(Precision) 0.892, 재현율(Recall) 0.892, F1-score 0.892를 달성하였으며, K-MMVI 점수를 연속 변수로 한 판별력(AUC)은 0.959로 매우 높게 나타났다. 또한 K-MMVI와 응급 취약지 여부 간의 점이연 상관계수는 0.783($p<0.001$)으로 강한 양의 상관을 보였다. 이는 본 연구가 제안한 K-MMVI가 정부의 공식 취약지 분류와 높은 수준으로 일치함을 의미하며, 지수의 타당성을 통계적으로 입증한다.

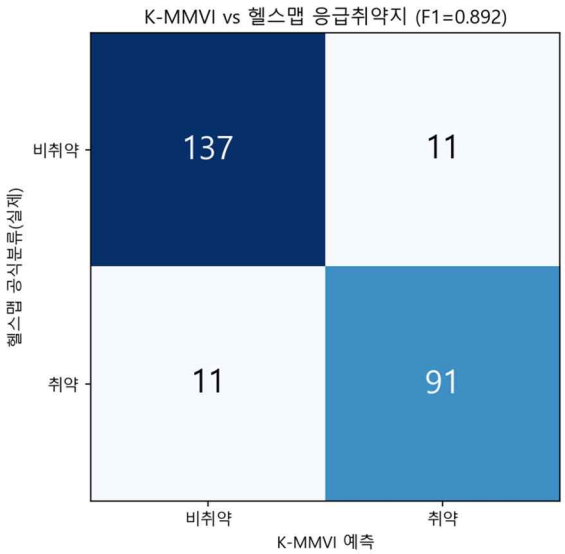


그림 4. 헬스맵 공식 분류 대비 K-MMVI 혼동행렬

그림 4의 혼동행렬에서 전체 250개 시군구 중 228개(91.2%)가 일치하였으며, 불일치한 22개 중 K-MMVI는 취약으로 판정하였으나 헬스맵은 비취약으로 분류한 11개 지역은 기존 공식 지표가 포착하지 못한 잠재적 취약지로 해석될 여지가 있다. 이는 단일 분야(응급) 중심의 기존 지표와 달리, 고령 수요·거리·공급을 통합한 K-MMVI가 보완적 정보를 제공할 수 있음을 시사한다.

추가적으로, 가중치 산정 방식의 영향을 확인하기 위해 엔트로피 가중치와 동일가중치(각 1/3) 방식을 비교한 결과, 보령시·거창군·평창군 등 일부 지역에서 30위 내외의 순위 변동이 관찰되었다. 이는 가중치 결정 방식이 정책 우선순위에 실질적 영향을 미침을 보여주며, 데이터 기반의 객관적 가중치 산정의 중요성을 뒷받침한다.

표 2. 가중치 산정 방식에 따른 취약순위 변동 비교

	시군구명	취약순위	순위_eq	순위차	KMMVI	KMMVI_eq
19	보령시	20	53	-33	36.51	61.62
10	거창군	11	43	-32	40.38	64.17
11	평창군	12	35	-23	39.80	65.39
12	횡성군	13	32	-19	38.81	65.83
5	문경시	6	22	-16	42.97	67.23
14	완도군	15	27	-12	38.32	66.48
17	부안군	18	29	-11	36.84	66.39
9	무주군	10	15	-5	41.07	68.23

Ⅳ. 결론

본 연구는 초고령사회 진입과 지역 간 의료 격차 심화라는 구조적 문제에 대응하기 위해, 인구 위험성·의료 원격성·의료 공급 밀도의 3대 차원을 융합한 한국형 의료 이동 취약성 지수(K-MMVI)와 이를 산출하는 5단계 분석 프레임워크를 제안하였다. 행정안전부 주민등록 인구통계, 보건복지부 전국의료기관 표준데이터, 국립중앙의료원 헬스맵 등 출처와 단위가 상이한 공공데이터를 표준 행정구역 코드를 기준으로 통합하고, Min-Max 정규화와 엔트로피 가중치법을 적용하여 객관적 취약성 점수를 도출하였다. 특히 의료 공백이 극심한 지역이 여타 지표의 양호함에 의해 은폐되는 변수 간 상쇄 효과를 제어하기 위해 물리적 차단 선제 플래그(Auxiliary Flagging Rule)를 결합함으로써, 단순 인프라 부족 지역과 교통망 단절형 소외 지역을 정량적으로 구분하였다. 전국 250개 시군구에 적용한 결과, 의료 원격성(가중치 0.746)이 지역 간 의료 이동 취약성을 결정하는 가장 핵심적인 요인으로 나타났다. 산출된 K-MMVI는 국립중앙의료원 헬스맵의 공식 응급의료 취약지 분류와 정확도 0.912, F1-score 0.892, AUC 0.959로 높은 수준의 일치를 보여 지수의 통계적 타당성이 입증되었다. 또한 기존 공식 지표가 취약지로 분류하지 못한 잠재적 취약지를 추가로 식별함으로써, 단일 분야 중심의 기존 평가 체계를 보완할 수 있는 가능성을 확인하였다. 본 연구는 파편화된 공공데이터를 정책 수립자가 즉각 해석 가능한 객관적 취약도 지표로 변환하였다는 점에서 의의를 가지며, 수요응답형 교통(DRT) 노선 최적화 및 찾아가는 의료 복지 자원의 효율적 배분을 위한 의사결정 근거로 활용될 수 있다. 다만 본 연구는 종합병원급 의료기관과 시군구 단위의 거시 분석을 중심으로 수행되어, 대중교통 배차 간격 및 노선 밀도와 같은 실질적 이동 제약을 충분히 반영하지 못한 한계가 있다. 향후 연구에서는 대중교통 운행 데이터와 읍면동 단위의 미시 분석을 결합하여 지수의 정밀도를 제고하고, 시계열 데이터를 활용한 취약성 변화 추이 분석으로 확장하고자 한다.

참고문헌

[1] 보건복지부, 국립중앙의료원, "2023년 의료취약지 모니터링 연구 결과보고서," 보건복지부·국립중앙의료원, 2023.

[2] 국립중앙의료원, "헬스맵(Health Map) 의료취약지 모니터링 서비스," https://www.nmc.or.kr/nmc/contents/health_map (접속일: 2025).

[3] 조병희, "의료인력과 의료시설의 지역별 불균형 분포," 한국의 사회동향 2018, 통계청 통계개발원, 2018.

[4] 이명호, 유선철, 안종욱, 신동빈, "GIS 공간분석을 활용한 사회 취약지역의 분석에 관한 연구: 서울특별시를 중심으로," 대한공간정보학회지, 제24권, 제4호, 2016.

[5] 행정안전부, "주민등록 인구통계(지역별·연령별 인구현황)," https://jumin.mois.go.kr (접속일: 2025).

[6] 보건복지부, "전국의료기관표준데이터," 공공데이터포털, https://www.data.go.kr/data/15096293/standard.do (접속일: 2025).

[7] IOGP, "EPSG:5179 — Korea 2000 / Unified CS (UTM-K)," EPSG Geodetic Parameter Dataset, https://epsg.io/5179.

[8] 국립중앙의료원, "헬스맵(Health Map) 시군구 의

료취약지 지표,"
https://www.nmc.or.kr/nmc/contents/health_map (접속일: 2025).

구두/포스터 발표논문 : 2페이지
우수논문심사용: 4~6페이지 작성

생성형 AI 인터페이스의 출처 검증 가능성과 신뢰도 표시 방식이 사용자 신뢰 교정에 미치는 영향: 인지 부하의 매개 효과를 중심으로

담서단O(제1저자), 김장현*(교신저자)
O성균관대학교 인터랙션사이언스학과
*성균관대학교 인터랙션사이언스학과 / 인간AI인터랙션융합학과
e-mail: xzzmnms1005@naver.comO, alohakim@skku.edu*

The Impact of Source Verifiability and Confidence Display in Generative AI Interfaces on User Trust Calibration: The Mediating Role of Cognitive Load

TAN SHUDANO, Jang-Hyun Kim*
ODepartment of Interaction Science , Sungkyunkwan University
*Department of Interaction Science / Department of Human-Artificial Intelligence Interaction, Sungkyunkwan University

요 약

생성형 AI의 확산과 함께 환각 현상으로 인한 신뢰 오류 문제가 심화되고 있다. 기존 연구들은 출처 제공과 신뢰도 표시가 신뢰에 미치는 효과에 대해 일관되지 않은 결과를 보고하고 있으나, 이 모순을 설명하는 인지적 기제는 충분히 규명되지 않았다. 본 연구는 인지 부하가 이 관계를 매개한다는 경로를 생성형 AI 맥락에서 직접 실증하고자 한다.

출처 제공의 검증 가능성 깊이(저: 모호한 텍스트 / 고: 클릭 가능한 외부 문헌 초록)와 신뢰도 표시 방식(수치형 / 시각형)을 독립변수로 한 2×2 피험자 간 실험을 수행한다. 대학생 180명을 대상으로 학술 보고서 작성 시나리오에서 AI가 제공한 정보(정확 사례 70%, 환각 사례 30%)의 인용 여부를 판단하는 과업을 수행하며, 인지 부하는 NASA-TLX로 직접 측정하고 SEM으로 매개 효과를 검증한다.

이론적으로는 출처 제공 및 신뢰도 표시 연구의 모순적 결과를 인지 부하의 매개 경로로 해명하며, 실무적으로는 생성형 AI 환각 문제를 완화하는 인터페이스 설계 지침을 제공한다.

키워드: 생성형 AI, 인터페이스 디자인, 신뢰 교정, 인지 부하, 환각 탐지

※이 논문은 2026년도 교육부 및 서울특별시의 재원으로 서울RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다 (2026-RISE-01-018-04).

I. 서론

생성형 AI 인터페이스에서 출처 제공 방식과 신뢰도 표시가 사용자 신뢰에 미치는 효과는 연구마다 상반된 결과를 보이며, 이를 설명하는 인지 부하의 매개 역할은 생성형 AI 맥락에서 충분히 실증된 바 없다. 본 연구는 출처 제공의 검증 가능성 깊이(저: 모호한 텍스트 선언 / 고: 클릭 가능한 외부 문헌 초록)와 신뢰도 표시 방식(수치형 / 시각형)이라는 두 인터페이스 단서가 인지 부하를 매개로 신뢰 교정에 미치는 차별적 영향을 2×2 피험자 간 실험(N=180)을 통해 실증한다. 참가자는 학술 보고서 작성 시나리오에서 AI가 제공한 정보(정확 사례 70%, 환각 사례 30%)의 인용 여부를 판단하며, 인지 부하는 NASA-TLX로 직접 측정하고 SEM으로 매개 효과를 검증한다. 본 연구는 기존 연구의 모순적 결과를 인지 부하의 매개 경로로 해명하고, 생성형 AI 환각으로 인한 과신뢰를 완화하는 인터페이스 설계 지침을 제공한다.

II. 관련 연구

신뢰 교정(Trust Calibration)이란 사용자가 AI의 실제 성능에 부합하여 올바를 때는 신뢰하고 오류가 있을 때는 비판적으로 거부하는 능력을 의미하며[1], AI 보조 의사결정 연구에서 핵심 성과 지표로 활용되어 왔다. 그러나 이를 지원하기 위한 투명성 단서의 효과는 일관되지 않는다. 출처 링크 제공은 허위 인용인 경우에도 신뢰를 높이지만[2], 사용자가 실제로 출처를 확인하면 오히려 신뢰가 낮아지며[3], 상세한 공개는 신뢰를 저해하는 역설도 보고된다[4]. 신뢰도 표시 역시 수치형 점수는 신뢰 교정에 효과적이지만[1], 불확실성 강조 방식은 오히려 신뢰를 저하시키고[2], 신뢰 교정 지원이 의사결정 효율을 개선하지 못하는 역설도 발견되었다[5]. 출처 제시 형식의 차이[6,7], 지원 정보 유형[8], 불확실성 시각화 형식[9] 등 다양한 변수가 개별적으로 탐구되었으나, 출처 귀속의 검증 가능성 깊이와 신뢰도 표시 방식을 동시에 조작하여 신뢰 교정에 미치는 교호작용 효과를 분석한 연구는 부족하다.

인지 부하는 이러한 모순적 결과를 설명하는 핵심 기제로 반복적으로 지목되어 왔으나 충분히 실증되지 않았다. 과도한 투명성이 인지 과부하를 유발하여 신뢰를 저해한다고 추론되었으나 인지 부하는 직접 측정되지 않았고[10], 단순 과업에서 인지 부하의 유의미한 효과가 확인되지 않았으며[11], 이미지 분류 과업에서도 매개 효과가 유의미하지 않음이 보고되었다[12]. 한편 인지적 처리 비용이 낮아질수록 오류 탐지 정확도가 향상됨이 실증되어 인지 부하의 역할에 대한 간접적 근거가 제공되었다[13]. 그러나 생성형 AI 환각 탐지 맥락에서 두 인터페이스 단서가 인지 부하를 매개로 신뢰 교정에 이르는 경로를 통합적으로 실증한 연구는 부족하다.

III. 제안방법

본 연구는 2×2 피험자 간 요인 설계를 적용한다. 독립변수는 출처 제공의 검증 가능성 깊이(저: 모호한 텍스트 선언 / 고: 클릭 가능한 외부 문헌 초록)와 신뢰도 표시 방식(수치형 / 시각형)이며, 인지 부하를 매개변수, 신뢰 교정을 종속변수로 설정한다.

G*Power 3.1.9.7 사전 검정력 분석(f = .25, α = .05, power = .80) 결과에 따라 대학생 180명(조건당 45명)을 모집하며, 참가자는 무선 배정을 통해 네 개의 조건 중 하나에 할당된다. 실험은 웹 기반

※이 논문은 2026년도 교육부 및 서울특별시의 재원으로 서울RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다 (2026-RISE-01-018-04).

인터페이스로 구현되며, 학술 보고서 작성 시나리오에서 AI가 제공한 연구 결과(정확 사례 70%, 환각 사례 30%)의 인용 여부를 판단하는 과업을 수행한다. 신뢰 교정은 인용 결정과 실제 정답의 일치율로 측정하며, 인지 부하는 NASA-TLX[14]로 직접 측정한다.

데이터 분석은 조작 검증(ANOVA), 직접 효과 및 교호작용 효과 검증(이원 분산분석), 인지 부하의 매개 효과 검증(SEM, 부트스트래핑 n = 5,000) 순으로 진행된다.

참고문헌

[1]Zhang, Y., Liao, Q. V.,& Bellamy, R. K.(2020, January). Effect of confidence and explanation on accuracy and trust calibration in Al-assisted decision making. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp.295-305).

[2]Li, H., & Aral, S. (2025). Human trust in AI search: A large-scale experiment. arXiv preprint arXiv:2504.06435.

[3]Ding, Y., Facciani, M., Joyce, E., Poudel, A., Bhattacharya, S., Veeramani, B., ... & Weninger, T. (2025, April). Citations and trust in llm generated responses. In Proceedings of the AAAI conference on artificial intelligence (Vol. 39, No. 22, pp. 23787-23795).

[4]Prajod, P., Cools, H., Röggla, T., Venkatraj, K. P., Kusters, A., ElKattan, A., ... & Ali, A. E. (2026). Full Disclosure, Less Trust? How the Level of Detail about AI Use in News Writing Affects Readers' Trust. arXiv preprint arXiv:2601.09620.

[5]Li, J., Yang, Y., Zhang, R., Liao, Q. V., Song, T., Xu, Z., & Lee, Y. C. (2024). Understanding the Effects of Miscalibrated AI Confidence on User Trust, Reliance, and Decision Efficacy. arXiv preprint arXiv:2402.07632.

[6]He, J., & Liu, J. (2026, March). Not All Transparency Is Equal: Source Presentation Effects on Attention, Interaction, and Persuasion in Conversational Search. In Proceedings of the 2026 Conference on Human Information Interaction and Retrieval (pp. 529-533).

[7]He, J., & Liu, J. (2026). Seeing to Think? How Source Transparency Design Shapes Interactive Information Seeking and Evaluation in Conversational AI. arXiv preprint arXiv:2601.14611.

[8]Irons, J., Cooper, P., Bolucu, N., Timmer, R., Yang, H., Lee, C., ... & Wan, S. (2026). To Believe or Not To Believe: Comparing Supporting Information Tools to Aid Human Judgments of AI Veracity. arXiv preprint arXiv:2603.11393.

[9]Zhao, J., Wang, Y., Mancenido, M. V., Chiou, E. K., & Maciejewski, R. (2024). Evaluating the impact of uncertainty visualization on model reliance. IEEE Transactions on Visualization and Computer Graphics, 30(7), 4093-4107.

[10]Ngo, V. M.(2025). Balancing Al transparency: trust, certainty, and adoption. Information Development,02666669251346124.

[11]Yang, R, Li, S, Qi, Y., Liu, J., He, Q,& Zhao, H.(2025). Unveiling users' algorithm trust: The role of task objectivity, time pressure, and cognitive load. Computers in Human Behavior Reports, 18, 100667.

[12]Karran, A. J., Demazure, T., Hudon, A., Senecal, S., & Léger, P. M. (2022). Designing for confidence: The impact of visualizing artificial intelligence decisions. Frontiers in neuroscience, 16, 883385.

[13]Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S.,& Krishna, R.(2023). Explanations can reduce overreliance on ai systems during decision-making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1-38.

[14]Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Advances in psychology (Vol. 52, pp. 139-183). North-holland.

※이 논문은 2026년도 교육부 및 서울특별시의 재원으로 서울RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다 (2026-RISE-01-018-04).

한국어 단문 환경에서의 생성형 AI 작성 리뷰 탐지: KoBERT 기반 하이브리드 모델

이상철[O,1], 이다예[1], 이하영[1], 남우민[2], 박건철[1,*]
[1,O]한국공학대학교 경영학부
[2]한국공학대학교 산학협력단
e-mail: andy4003@tukorea.ac.kr[1,O] , daye925@tukorea.ac.kr[1],
2023316034@tukorea.ac.kr[1], applewoods9559@tukorea.ac.kr[2],
parkkc07@tukorea.ac.kr[1,*]
[O]제1저자, [*]교신저자

Detecting AI-Generated Reviews in Korean Short-text Environments: A KoBERT Hybrid Model

Sangchul Lee[1,O], Daye Lee[1], Hayoung Lee[1], Woomin Nam[2], Keon Chul Park[1,*]
[1,O]Business Administration Department, Tech University of Korea
[2]Industry-Academic Cooperation Foundation, Tech University of Korea
[O]First Author, [*]Corresponding Author

요 약

생성형 AI 기술이 확산되면서 온라인 이커머스 플랫폼에는 AI가 작성한 리뷰가 빠르게 증가하고 있으며, 이는 온라인 정보의 신뢰성을 저하시켜 소비자의 합리적 의사결정을 저해하는 요인으로 작용한다. 그러나 영어 기반 장문 텍스트에 집중된 기존 연구와 짧은 텍스트에서 사람이 작성한 글을 AI 생성물로 오판하는 상용 탐지 도구의 한계로 인해 한국어 단문 리뷰에 적합한 탐지 방법은 여전히 부족한 실정이다. 본 연구는 한국어 단문 리뷰에 특화된 생성형 AI 작성 리뷰 탐지 모델을 제안한다. 데이터셋은 네이버 스마트스토어에서 수집한 실제 리뷰 88,967건과 사용 생성형 AI모델을 이용하여 생성한 합성 리뷰 70,003건으로 구성하였다. 제안 모델은 KoBERT 임베딩에 BiLSTM과 Self-Attention을 결합하여 문장 내의 미세한 언어 패턴을 포착하도록 설계하였다. 이를 통해 한국어 단문 환경에서 생성형 AI 리뷰 탐지의 실효성을 검증하여 온라인 리뷰의 신뢰성 문제에 대응할 실질적 방안을 제시하고자 한다.

키워드: 생성형 AI, 가짜 리뷰 탐지, KoBERT, BiLSTM, 한국어 자연어 처리

※ 본 연구는 국토교통부/국토교통과학기술진흥원의 지원으로 수행되었습니다(과제번호: RS-2022-00143404). 또한 본 성과는 정부(과학기술정보통신부)의 재원으로 AI컴퓨팅자원활용기반강화(GPU 임차 지원) 사용자 지원사업의 지원을 받아 수행된 연구입니다(사사번호: RQT-25-090204).

I. 서론

최근 생성형 AI 기술이 급격히 발전하면서 인공지능이 작성한 온라인 리뷰가 빠르게 증가하고 있으며, 이에 따라 온라인 정보의 신뢰성에 대한 우려도 커지고 있다. 소비자는 AI가 생성한 리뷰를 실제 사용자 경험에 기반한 리뷰보다 진정성과 설득력이 낮은 정보로 인식하는 경향이 있으며, 이는 결국 구매 의도의 감소로 이어질 수 있다[1]. AI 콘텐츠 탐지 기업 Originality.AI의 분석에 따르면, 2025년 부동산 중개 플랫폼 Zillow에 게시된 중개인 리뷰의 약 23.7%가 AI로 생성된 것으로 추정되며, 이는 2019년의 3.63%에서 약 558% 증가한 수치이다[2]. 미국 연방거래위원회(FTC) 또한 생성형 AI가 리뷰를 대량으로 생산하여 소비자에게 기만적인 정보를 제공하는 도구로 악용될 수 있음을 지적한 바 있다[3]. 이처럼 소비자가 정보의 진위를 스스로 판단하기 어려워지는 환경에서, 본 연구는 AI가 생성한 리뷰를 효과적으로 탐지하기 위한 딥러닝 기반 모델을 제안한다.

II. 관련 연구

기존의 AI 생성 텍스트 탐지 연구는 대부분 영어 데이터셋을 중심으로 수행되어 왔다[4]. 또한 사람이 작성한 허위 리뷰의 탐지에 초점을 맞추고 있어, 한국어 단문 환경에서의 AI 생성 리뷰 탐지에는 한계가 있다[5]. 대표적인 AI 탐지 도구인 GPTZero 역시 100단어 미만의 단문에서는 사람이 작성한 글을 평균 35.56%의 확률로 AI 작성물로 오판하는 등, 단문 환경에서 정확도가 크게 저하되는 한계를 보였다[6]. 따라서 본 연구에서는 실제 리뷰와 생성형 AI 기반 합성 리뷰 데이터를 활용하여 한국어 단문 리뷰 탐지 모델을 구축하고자 한다. 제안 모델은 KoBERT-BiLSTM-Self-Attention 구조를 기반으로 AI 생성 리뷰를 탐지하고 Fully Connected Layer와 Sigmoid 함수를 이용하여 생성 확률을 출력하는 모델 구축을 목표로 한다.

III. 제안방법

3.1 데이터 수집

본 연구에서는 실제 리뷰 데이터와 AI 합성 리뷰 데이터를 함께 구축하였다. 실제 리뷰 데이터는 네이버 스마트스토어의 가습기 제품 리뷰를 크롤링하여 수집하였으며, AI 합성 리뷰는 GPT-4o와 Gemini 2.5 Flash를 이용하여 생성하였다. 실제 온라인 리뷰에서 나타나는 다양한 AI 생성 패턴을 반영하기 위해, 합성 데이터는 단일 시나리오 아래 A/B/C 세 가지 프롬프트 전략을 적용하여 생성하였다. 각 프롬프트 전략(A/B/C)별 세부 생성 방식은 표 1과 같다.

표 1. 프롬프트 전략 별 AI 합성 리뷰 생성

구분	생성방식	주요지침
Type A	조건 기반 생성	·제품명·별점 반영 ·표현 길이 제한 ·표현 방식 변형
Type B	실제 리뷰 재학습	·의미·감성 유지 ·문장 구조 변경 ·원문 복사 금지
Type C	실제 리뷰 스타일 모방	·말투·문체·구어체 반영 ·자연스러운 비문 허용 ·쇼핑몰 후기 스타일

※ 본 연구는 국토교통부/국토교통과학기술진흥원의 지원으로 수행되었습니다(과제번호: RS-2022-00143404). 또한 본 성과는 정부(과학기술정보통신부)의 재원으로 AI컴퓨팅자원활용기반강화(GPU 임차 지원) 사용자 지원사업의 지원을 받아 수행된 연구입니다(사사번호: RQT-25-090204).

3.2 데이터 전처리

수집된 리뷰 데이터는 HTML 태그와 URL 제거, 불필요한 공백 제거, 길이 기준 필터링 등의 전처리 과정을 거쳐 학습 데이터로 변환하였다. 이때 실제 온라인 리뷰의 특성을 보존하기 위해 이모지와 반복적인 구어체 표현을 제거하지 않고, 각각 <EMOJI>와 <ㅋ>와 같은 특수 토큰으로 변환하여 KoBERT 토크나이저에 추가하였다.

3.3 모델 구성

제안 모델은 KoBERT, BiLSTM, Self-Attention을 결합한 하이브리드 구조로 설계하였다. 먼저 리뷰 텍스트는 KoBERT를 통해 토큰화되며, KoBERT는 문맥 기반 임베딩 벡터를 생성하여 문장의 의미 정보를 추출한다. 이어 BiLSTM 계층은 문장의 앞뒤 정보를 함께 학습하여 단어 간의 순차적 관계와 문맥 흐름을 반영한다. 다음으로 Self-Attention 계층은 문장 내 중요한 단어와 표현에 높은 가중치를 부여함으로써, AI 생성 리뷰에서 반복적으로 나타나는 특징적 표현 패턴을 효과적으로 학습한다. 최종적으로 Fully Connected Layer와 Softmax 함수를 거쳐 해당 리뷰가 AI 생성물일 확률을 산출하도록 설계하였다.

IV. 결론

본 연구에서는 한국어 단문 환경에서의 생성형 AI 리뷰 탐지를 위해 A/B/C 프롬프트 전략으로 구성한 합성 리뷰 데이터를 기반으로 탐지 모델을 구축하고 그 성능을 검증하고자 한다. 향후 교차 검증을 통해 모델의 일반화 성능을 평가하고, 다양한 리뷰 환경에서도 안정적인 탐지 성능을 보이는 모델을 최종 모델로 선정할 계획이다. 본 연구는 한국어 단문 리뷰 환경에서 생성형 AI 작성 리뷰의 탐지 가능성을 제시하였다는 점에서 의의가 있다.

참고문헌

[1] S. Knight, Y. Bart, and M. Yang, "Generative AI and the Perceived Quality of User-Generated Content: Evidence from Online Reviews," SSRN Working Paper No. 4621982, Nov. 2023.

[2] Originality.AI, "Fake AI Zillow Reviews Increased by 558% from 2019 to 2025," Originality.AI Research Report, Jul. 2025.

[3] Federal Trade Commission, "Federal Trade Commission Announces Final Rule Banning Fake Reviews and Testimonials," FTC Press Release, Aug. 2024.

[4] B. Guo et al., "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection," arXiv preprint, arXiv:2301.07597, Jan. 2023.

[5] 강주영, 송민, "한국어 가짜 구매후기 생성과 탐지 성능 평가," 지능정보연구, Vol. 30, No. 2, pp. 313-328, June 2024.

[6] S. Dik, O. Erdem, and M. Dik, "Assessing GPTZero's Accuracy in Identifying AI vs. Human-Written Essays," arXiv preprint, arXiv:2506.23517, June 2025.

※ 본 연구는 국토교통부/국토교통과학기술진흥원의 지원으로 수행되었습니다(과제번호: RS-2022-00143404). 또한 본 성과는 정부(과학기술정보통신부)의 재원으로 AI컴퓨팅자원활용기반강화(GPU 임차 지원) 사용자 지원사업의 지원을 받아 수행된 연구입니다(사사번호: RQT-25-090204).

시민참여형 도시계획을 위한 생성형 AI 기반 퍼실리테이션 지원 시스템 제안

김정민[1,O], Uprety Sudan Prasad[2], 남우민[2], 박건철[1,*]
[1]한국공학대학교 경영학부
[2]한국공학대학교 산학협력단
e-mail: 3jang1239@tukorea.ac.kr[O], uprety_sud@tukorea.ac.kr,
applewoods9559@tukorea.ac.kr, parkkc07@tukorea.ac.kr[*]
[O]제1저자, [*]교신저자

AI Facilitation for Participatory Urban Planning : A Generative AI-Based Support System

Jungmin Kim[1,O], Sudan Prasad Uprety[2], Woomin Nam[2], Keon Chul Park[1,*]
[1]Department of Business Administration, Tech University of Korea
[2]Industry-Academic Cooperation Foundation, Tech University of Korea
[O]First Author, [*]Corresponding Author

요 약

국내 도시계획 분야에서는 시민계획단과 도시아카데미 등 다회차·다분과 형태의 시민참여 프로그램이 빠르게 안착하고 있다. 그러나 토론을 매개하는 퍼실리테이터가 도시계획 도메인 전문성을 충분히 갖추지 못하고, 회차 간 논의의 연속성을 유지하기 어렵다는 점이 지속적인 한계로 지적되어 왔다. 최근 생성형 AI를 활용한 시민 참여형 도시계획 설계 연구가 보고되고 있으나, 한국 고유의 법령·계획체계와 다회차 운영 구조를 반영하여 인간 퍼실리테이터를 보조하는 협력형 시스템에 관한 연구는 아직 미비한 실정이다. 이에 본 연구는 멀티모달 LLM과 RAG 기술을 기반으로 전문 퍼실리테이터의 역량을 강화하고 회의 운영을 지원하는 '도시계획 특화 AI 퍼실리테이션'의 아키텍처를 제안한다. 본 시스템은 회의 진행 보조, 도시계획 도메인 지식 지원 등으로 구성된다. 본 시스템은 퍼실리테이터의 도메인 지식 격차를 보완하여 시민 제안의 법적·기술적 실현 가능성을 높이고, 특정 참여자의 의견 독점을 방지함으로써 토론의 형평성과 민주적 합의의 질을 향상시킨다. 나아가 다회차 운영에서 발생하는 행정적 공백을 메워 시민참여형 도시계획의 일관성과 효율성을 극대화하는 실무적 가치를 지닌다.

키워드: 시민참여형 도시계획, 시민계획단, 퍼실리테이션, AI 퍼실리테이터, 생성형 AI

※본 연구는 국토교통부/국토교통과학기술진흥원의 지원으로 수행되었습니다.(과제번호 RS-2022-00143404)

I. 서론

도시화가 가속화되면서 교통, 주거, 환경 등 다양한 도시문제가 복합적으로 나타나고 있다. 이에 따라 시민의 요구를 도시계획 과정에 반영하기 위한 시민참여형 도시계획의 중요성이 확대되고 있다. 국내에서도 시민계획단, 도시아카데미 등 숙의형 참여 프로그램이 확산되며 퍼실리테이션 기반의 운영 방식이 활성화되고 있다[1].

퍼실리테이션은 다양한 참여자의 의견을 조정하고 합의를 도출하는 과정이며, 퍼실리테이터는 토론을 촉진하고 논의 흐름을 관리하는 역할을 수행한다[1]. 그러나 시민참여 프로그램이 확대되면서 퍼실리테이터의 운영 부담 또한 증가하고 있다. 특히 도시계획 관련 전문 지식 부족과 다회차 운영 과정에서의 연속성 유지 문제가 주요 한계로 지적된다.

최근 LLM(Large Language Model)을 활용한 회의 지원 연구가 증가하고 있으나, 국내 도시계획의 제도적 특성과 다회차 시민참여 구조를 반영한 연구는 부족한 실정이다. 또한 인간 퍼실리테이터를 보조하는 협력형 시스템 연구 역시 제한적으로 이루어지고 있다.

본 연구는 멀티모달 LLM과 RAG 기술을 기반으로, 시민참여형 도시계획에서 퍼실리테이터의 운영을 지원하는 「도시계획 특화 다회차 퍼실리테이션 보조 시스템」의 아키텍처를 제안한다. 본 시스템은 토론 진행 지원, 도시계획 도메인 지식 제공, 회차 간 연속성 관리 기능으로 구성되며, 시민참여형 도시계획의 숙의 품질과 운영 효율성 향상에 기여하고자 한다.

II. 관련 연구

2.1 시민참여형 도시계획과 퍼실리테이션

시민참여는 도시계획의 정당성과 실행력을 높이는 핵심 요소로, 도시서비스와 삶의 질 향상, 공공 신뢰 증진 및 사회적 갈등 비용 감소에 기여한다[2]. 이러한 참여를 실제 계획 과정으로 연결하는 수단이 퍼실리테이션이며, 퍼실리테이터는 토론을 조정하고 참여를 촉진하는 역할을 수행한다. 시민참여형 도시계획에서는 회의 운영 능력과 함께 도시계획에 대한 이해가 요구된다[1].

2.2 AI를 활용한 퍼실리테이션 지원 연구

최근에는 생성형 AI와 LLM을 활용한 퍼실리테이션 지원 연구가 증가하고 있다. 연구에 따르면 AI는 참여자 질문 유도, 토론 요약, 계획 관련 정보 제공 등을 통해 참여형 설계를 지원할 수 있다[3-4]. 그러나 기존 연구는 단일 회의 중심이 많으며, 국내 도시계획 제도와 시민계획단의 다회차 운영 구조를 반영한 협력형 지원 연구는 아직 부족하다.

III. 제안방법

3.1 시스템 구조

제안 시스템은 그림 1과 같이 AI 퍼실리테이터, 시민 참여자, 행정·전문가 간 상호작용을 지원하는 구조로 설계되었다. 시스템은 멀티모달 LLM과 RAG 기반으로 참여자의 발언, 회의 자료, 지도·도면 등의 정보를 통합 분석하며, 범용 퍼실리테이션 기능과 도시계획 특화 기능을 제공한다. 범용 기능은 회의 진행 보조, 회의록 자동 생성, 발언 쏠림 감지, 의견 클러스터링 등을 지원하며, 도시계획 특화 기능은 법령·절차 검색, 회차 간 연속성 브리핑, 미결

※본 연구는 국토교통부/국토교통과학기술진흥원의 지원으로 수행되었습니다.(과제번호 RS-2022-00143404)

쟁점 추적 등을 수행한다. 이를 통해 회의록 및 결정사항 문서화와 시민참여 품질 향상을 지원한다.

3.2 주요 기능

제안 시스템은 진행 보조, 도메인 지식 지원, 다회차 연속성 관리 기능으로 구성된다. 시스템은 회의 내용을 자동으로 정리하고 발언 편중이나 참여 저조를 감지하여 퍼실리테이터의 회의 운영을 지원하며, 유사 의견을 분류해 합의 형성을 돕는다. 또한 RAG 기반으로 도시계획 관련 법령·조례·계획 정보를 제공하고, 전문 용어와 지역 인프라 정보를 시민이 이해하기 쉬운 형태로 설명한다. 아울러 이전 회차 논의와 미결 쟁점을 요약하여 회차 간 논의 흐름을 유지하고, 분과별 결과 비교와 퍼실리테이션 시나리오 조정을 통해 지속적인 토론을 지원한다.

그림 1. 도시계획 특화 퍼실리테이션 보조 시스템

IV. 결론

본 연구는 시민참여형 도시계획에서 퍼실리테이터를 지원하기 위한 생성형 AI 기반 퍼실리테이션 보조 시스템을 제안하였다. 제안 시스템은 멀티모달 LLM과 RAG를 기반으로 회의 진행 지원, 도시계획 지식 제공, 다회차 연속성 관리 기능을 통합적으로 지원하도록 설계되었다. 이를 통해 퍼실리테이터의 운영 부담과 전문성 한계를 보완하고, 시민참여 과정의 운영 효율성과 숙의 품질 향상을 목표로 하였다.

특히 본 연구는 국내 도시계획 제도와 다회차 시민참여 구조를 반영했다는 점에서 기존 연구와 차별성을 가지며, AI를 인간 퍼실리테이터의 협력형 지원 도구로 활용했다는 의의를 가진다. 향후에는 실제 시민계획단 운영 환경에 적용하여 참여 효과와 운영 효율성에 대한 실증적 검증이 필요하다.

참고문헌

[1] 이수암, "퍼실리테이션(Facilitation)을 활용한 도시정책 시민계획단 운영방안," 국토연구원, Working Paper WP 19-25, 2019.

[2] 남우민, 박건철, "스마트도시 구현을 위한 시민참여의 역할과 방향에 관한 연구," 인터넷정보학회논문지(JICS), 제23권 제6호, 79-86쪽, 2022.

[3] M. Houtti, M. Zhou, L. Terveen, and S. Chancellor, "Observe, Ask, Intervene: Designing AI Agents for More Inclusive Meetings," in *Proc. CHI Conference on Human Factors in Computing Systems (CHI '25)*, Yokohama, Japan, 2025.

[4] Z. Zhou, Y. Lin, and Y. Li, "Large language model empowered participatory urban planning," arXiv preprint arXiv:2402.01698, 2024.

※본 연구는 국토교통부/국토교통과학기술진흥원의 지원으로 수행되었습니다.(과제번호 RS-2022-00143404)

멀티 에이전트 토론과 가드레일 기반 규칙 평가에 따른 다중 인공지능 답변 시스템

문태웅O, 이진형, 나인섭*
전남대학교 문화콘텐츠학부
e-mail: xodnd4134@naver.comO, 2u2ok@naver.com, ypencil@hanmail.net

Multi-AI Q&A System based on Multi-Agent Debate and Guardrail-Based Rule Evaluation

Taewoong MoonO, Jinhyoung Lee, Inseop NA*
Division of Culture Contents, Chonnam National University
*교신 저자: 나인섭(ypencil@hanmail.net)

요 약

본 연구는 단일 LLM의 편향, 환각, 최신 정보 부족 문제를 개선하기 위해 ChatGPT와 Gemini 기반 이종 멀티에이전트 답변 시스템을 제안한다. 에이전트 간 토론과 자동 상호 검토를 통해 답변을 개선하고, 최종적으로 가드레일 기반 규칙 평가를 적용하여 최적 답변을 선정한다. 실험 결과 객관적·주관적 질문 모두에서 답변 품질과 일관성이 향상됨을 확인하였다.

키워드: 대규모 언어 모델, ChatGPT, Gemini, 상호 검토, 자동 상호 교차검토, 규칙 기반 판단

I. 서론

대규모 언어모델(LLM)은 자연어 질의응답과 문서 생성에서 우수한 성능을 보이지만, 단일 모델 기반 응답은 편향, 환각, 최신 정보 부족 등의 한계를 가질 수 있다[1]. 이를 보완하기 위해 검색 기반 생성과 다중 에이전트 협업 기법이 제안되었으며, 선행 연구에서는 외부 지식 활용과 에이전트 간 토론이 답변의 사실성과 추론 성능을 향상시킬 수 있음을 보고하였다[1]. 본 연구는 ChatGPT와 Gemini로 구성된 이종 LLM이 상호 교차 검토를 통해 답변을 개선하는 방식을 제안한다. 또한 자동 토론 메커니즘과 규칙 기반 평가 방식을 적용하여 최종 답변을 선정하고, 다중 에이전트 협업 및 규칙 기반 평가의 효과와 한계를 분석한다.

II. 제안방법

2.1 멀티에이전트 토론

제안 시스템은 동일한 사용자 질문을 ChatGPT와 Gemini에 각각 입력하여 초기 답변을 생성한다. 생

※본 논문은 농촌진흥청 연구사업(RS-2026-25528889)의 지원에 의해 이루어진 것임.

성된 답변은 이후 교차 검토의 기준으로 활용된다. 교차 검토 단계에서는 각 모델이 자신의 현재 답변과 상대 모델의 답변을 함께 입력받아 상대 답변을 평가하고, 이를 반영하여 자신의 답변을 수정한다.

자동 교차 검토 기능은 사용자가 설정한 최대 검토 횟수 내에서 두 모델이 서로의 직전 답변을 참조하며 반복적으로 답변을 개선하는 과정이다. 각 검토 단계 후 모델은 추가 검토 필요 여부를 판단하며, 양측 모두 추가 검토가 불필요하다고 판단하면 검토를 종료한다. 반면 하나 이상의 모델이 추가 검토가 필요하다고 판단할 경우 최대 횟수에 도달할 때까지 다음 검토 단계를 수행한다. 본 연구는 고정된 반복 횟수를 사용하는 대신, 모델 간 답변 수렴 여부를 기준으로 검토 횟수를 동적으로 결정하는 가변 반복 방식의 자동 교차 검토 기법을 제안한다.

2.2 가드레일기반 규칙평가

자동 상호 교차검토가 종료되면 ChatGPT와 Gemini는 각각 자신의 초기 답변과 검토 과정에서 생성된 답변 이력을 바탕으로 최종 후보 답변을 생성한다. 각 모델은 상대 모델의 답변을 직접 통합하지 않고, 자신의 답변 변화 과정을 종합하여 최종 후보를 작성한다. 이후 가드레일 기반 규칙 평가 모듈이 두 최종 후보를 비교하여 최종 답변을 선정한다.

본 연구의 가드레일 기반 규칙 평가는 LLM-as-a-Judge 방식 대신 사전에 정의된 규칙과 텍스트 특징 분석을 활용하는 명시적 평가 방식이다. 이를 통해 특정 모델의 선호나 장황성에 따른 편향을 줄이고, 최종 선택 과정의 설명 가능성을 높이고자 하였다.

평가 모듈은 답변 길이, 반복률, 한글 비율, URL 포함 여부, 숫자·날짜·단위 표현, 제목·목록·표 등의 구조화 요소를 추출하여 언어적 적절성, 일관성, 이해 가능성, 구체성, 무해성 등의 지표를 계산한다. 또한 지나치게 짧은 답변, 오류 메시지나 fallback 응답이 포함된 경우에는 감점하며, 핵심 내용을 충실히 포함하고 구조화된 답변에는 가점을 부여한다. 구조화 정도는 제목, 번호 목록, 글머리표, 표, 코드 블록, 참고문헌 및 출처 표현 등을 기준으로 평가한다. 웹검색이 수행된 경우에는 URL 포함 여부를 추가로 평가하여 근거 제시 형식을 확인한다. 다만 URL의 실제 정확성이나 사실 여부는 검증하지 않으며, 이는 근거 제시 여부를 확인하기 위한 가드레일 역할에 한정된다.

사실성 평가는 reference context 또는 reference answer가 제공된 경우에만 제한적으로 수행한다. 이 경우 숫자, 날짜, 단위 정보의 일치 여부와 TF-IDF 기반 유사도를 활용하여 관련성을 보조적으로 평가한다. 따라서 본 연구의 평가 방식은 절대적인 정답 판정보다는 두 후보 답변의 형식적 품질과 안정성을 비교하는 데 초점을 둔다.

각 평가지표는 0~1 범위로 정규화되며, 평가 불가능한 항목은 제외한 후 가중치를 재조정하여 최종 점수를 계산한다. 최종적으로 가장 높은 종합 점수를 획득한 후보를 최종 답변으로 선택하며, 모델이 제공하는 신뢰도(confidence score)는 보조 지표로만 활용한다. 따라서 신뢰도는 최종 선택을 지원하는 참고 정보이며, 사실적 정확성을 보장하는 기준으로 사용하지 않는다.

참고문헌

[1] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

※본 논문은 농촌진흥청 연구사업(RS-2026-25528889)의 지원에 의해 이루어진 것임.

생성형 AI 기반 비정형 정책데이터 분석을 위한 PRAF 프레임워크 개발 연구

— AI 기반 어선설계플랫폼 사례를 중심으로 —

이주영[O*]

[O*]한국과학기술정책플랫폼협동조합(KSPP) 스마트미래융합본부

jylee@kspp.re.kr[O*]

Development of the PRAF Framework for Unstructured Policy Data Analysis Using Generative AI

— Focused on the AI-Based Fishing Vessel Design Platform Case —

Jooyoung Lee

*Smart Future Convergence Division, Korea Science Policy Platform co-op

요 약

신기술 보급확산 과정에서 법·제도적 제약은 주요 장벽으로 작용하나, 현재 법령 분석은 전문가 경험에 의존하여 체계성이 부족하다. 본 연구는 법령·고시·기술기준 등 비정형 정책 데이터를 구조화하고, 생성형 AI(LLM) 기반 의미 분석을 통해 규제항목과 기술 요구사항 간 연관성을 도출하는 6단계 정책데이터 분석 프레임워크(PRAF)를 제안한다. AI 기반 어선설계플랫폼 사례를 대상으로 어선법·선박안전법 등 관련 법령·기술기준을 정책 데이터셋으로 구축하고, LLM 기반 규제항목 추출과 기술-규제 연관성 분석 절차를 적용하여 주요 제도적 검토 필요 항목을 도출하였다. 총 20건의 비정형 정책 문서와 8개의 기술 요구사항을 입력 데이터로 활용하여 분석 절차를 적용하였으며, 주요 규제공백(Regulatory Gap)을 식별하고 보급확산 관점의 정책 검토과제와 이해관계자 영향도를 제안하였다. 본 프레임워크는 Legal AI와 달리 신기술 실증·사업화·보급확산 관점의 정책 의사결정 지원에 초점을 두며, 향후 다양한 신기술 분야의 보급확산 관점 정책 검토과제 도출에 활용 가능한 범용 도구로서 의의를 가진다.

키워드: 생성형 AI, 비정형 정책데이터, 규제공백, 기술확산, 정책 의사결정 지원

※ 논문은 2026년도 해양수산부 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임 (RS-2022-KS221571, AI기반 어선안전 설계 데이터플랫폼 개발 및 실증)

I. 서론

인공지능(AI), 디지털트윈, 자율시스템 등 신기술 기반 제품·서비스 개발이 빠르게 확대되고 있으나, 기술적 성능 확보와 별개로 법·제도 및 규제 환경의 미비로 인해 현장 적용과 보급확산이 지연되는 사례가 지속적으로 증가하고 있다[1]. 이른바 규제 지체(Regulatory Lag) 현상은 신기술 도입에 필수적인 관련 법령·고시·기술기준 검토가 체계적으로 이루어지지 않아 발생하며, 현재 대부분 전문가 경험에 의존한 수동 검토 방식에 의존하고 있어 일관성과 효율성 확보에 구조적 한계가 있다[2].

데이터 관점에서 보면, 신기술 도입 과정에서 검토해야 하는 법령·고시·기술기준·국제표준 문서는 비정형(Unstructured) 텍스트 데이터로 대량 존재하지만, 이를 구조화하여 기술 요구사항과 연계·분석하는 체계적 방법론은 부재한 상황이다. 생성형 AI(Generative AI), 특히 대규모 언어모델(LLM)은 비정형 정책 데이터를 처리하고 의미적 연관성을 추출하는 데 강점을 가지며[3], 이를 신기술 보급확산 지원에 활용하는 것이 본 연구의 핵심 문제의식이다.

본 연구의 핵심 연구문제는 다음과 같다. ₩"AI 기반 어선설계플랫폼의 현장 적용과 보급확산 과정에서 어떤 법령·기술기준·제도적 장벽이 존재하며, 이를 LLM 기반 비정형 정책 데이터 분석을 통해 어떻게 체계적으로 식별하고 개선 방향을 도출할 수 있는가?₩" 이를 위해 본 연구는 6단계 정책데이터 분석 프레임워크(Policy & Regulatory Analysis Framework, 이하 PRAF)를 제안하고, AI 기반 어선설계플랫폼 사례에 적용하여 실증적 유효성을 검토한다. 특히 본 연구는 기술 요구사항을 출발점으로 정책 데이터를 분석하는 Technology-to-Regulation 접근법을 적용하였다는 점에서, 규정 준수 판단에 초점을 둔 기존 Legal AI 연구와 명확한 차별성을 가진다.

II. 이론적 배경 및 PRAF 프레임워크

LLM은 대용량 비정형 텍스트 처리, 문맥 기반 의미 추출, 자동 분류 등에서 높은 성능을 보여 법령 문서 분석 영역에서 활발히 연구되고 있다[3]. Bommarito & Katz(2024)는 GPT-4가 법률 시험에서 전문가 수준 성과를 달성함을 확인하였으며[4], Nay(2022)는 LLM 기반 법령 분석이 AI 정렬 문제에 기여할 수 있음을 제시하였다[5]. 그러나 기존 Legal AI 연구는 규정 준수(Compliance) 여부 판단에 집중되어 있어, 신기술 실증·사업화·보급확산 관점의 규제 분석 및 정책 의사결정 지원 연구는 부족한 실정이다. Rogers(2003)의 혁신확산이론은 제도적 지원과 규제 명확성이 기술 수용의 핵심 촉진 요인임을 강조하며[6], 특히 해양수산 분야는 복합 규제 환경으로 다수 법령·기술기준의 동시 검토가 필요하다. 한편 국내에서도 AI 법·제도 정비 로드맵[8], 스마트 해양모빌리티 정책[9], 생성형 AI 규제 대응 방향[10] 등이 제시되고 있으나, 비정형 정책 데이터를 구조화하여 규제공백(Regulatory Gap)을 체계적으로 식별하는 방법론 연구는 미흡한 실정이다.

이에 본 연구는 Technology-to-Regulation

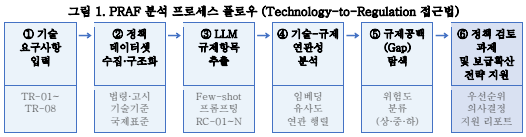
※ 논문은 2026년도 해양수산부 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임 (RS-2022-KS221571, AI기반 어선안전 설계 데이터플랫폼 개발 및 실증)

접근법에 기반한 6단계 정책데이터 분석 프레임워크 PRAF를 제안한다. PRAF는 비정형 정책 데이터를 수집·구조화하고, LLM 기반 분석을 통해 기술-규제 연관성과 규제공백을 도출하여 정책 의사결정을 지원하는 파이프라인이다. 기존 전문가 의존 방식과 달리 분석 과정을 체계화·재현 가능하게 설계하였으며, 단계별 구성과 데이터 흐름은 표 1 및 그림 1과 같다. LLM은 조문 단위 분절 후 Few-shot 프롬프팅으로 규제항목을 자동 추출(2단계)하고, 임베딩 기반 유사도 분석과 LLM 비교 분석을 병행하여 기술-규제 연관성을 도출한다(3단계). 환각(Hallucination) 오류 억제를 위해 원문 조문 인용 의무화와 전문가 검토를 병행하는 Human-in-the-loop 구조를 채택한다.

표 1. PRAF 6단계 구성 및 데이터 흐름

단계	구분	주요 활동	산출물	데이터 유형
1	정책 데이터 수집	법령·고시·기술기준·국제표준 등 비정형 문서 자동 수집 및 분류	법령 데이터셋	비정형 텍스트
2	LLM 기반 규제항목 추출	조문 단위 분절 후 LLM 프롬프트로 규제 의무·금지·기준 항목 구조화	규제항목 구조화 데이터	반정형
3	기술-규제 연관성 분석	기술 요구사항 명세와 규제항목 간 의미적 연관성 매핑	기술-규제 연관 행렬	정형 행렬
4	규제공백 탐색	연관 규정 미비·충돌·불명확 항목 탐지 및 위험도 분류(상·중·하)	규제공백 (Regulatory Gap)	정형+메타
5	정책 검토과제 및 우선순위 도출	샌드박스·기술기준 신설·법령 개정 등 경로별 개선 후보군 및 우선순위 생성	내가 우선순위 목록	반정형
6	의사결정 지원 보고	이해관계자별 영향도 분석 및 시각화 보고서 생성	정책 의사결정지원 리포트	시각화

그림 1. PRAF 분석 프로세스 플로우 (Technology-to-Regulation 접근법)



본 연구에서는 AI 기반 어선설계플랫폼의 설계, 안전, 데이터 활용, 검사·인증과 관련된 정책문서를 대상으로 정책데이터셋을 구축하였다. 정책데이터셋은 법률, 하위규정, 기술기준, 국제기준 및 정책보고서를 포함하여 총 20건으로 구성하였다.

표 2. 정책데이터셋 구성

구분	주요 문서	건수
법률	어선법, 선박안전법, 선박법, 어선안전조업법 등	6
시행령·시행규칙	관련 하위규정	4
고시·기술기준	어선검사기준, 설계기준 등	5
국제기준	IMO, ISO 등	3
정책자료	정부 정책문서	2
합계		20

III. AI 기반 어선설계플랫폼 사례 적용

AI 기반 어선설계플랫폼은 대규모 언어모델과 CAD 시스템을 연동하여 설계사의 복원성·안전성 검토 등 어선 설계 프로세스를 지원·지능화하는 신기술로, 해양수산부 R&D 과제로 개발 중에 있다. 동

※ 논문은 2026년도 해양수산부 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임 (RS-2022-KS221571, AI기반 어선안전 설계 데이터플랫폼 개발 및 실증)

플랫폼의 현장 보급을 위해서는 어선법, 선박안전법, 선박검사기준, 형식승인 관련 고시 등 복수의 법령·기술기준이 동시에 검토되어야 한다. 본 연구에서는 PRAF 1단계에 따라 관련 비정형 정책 문서 총 20건을 수집하여 표 3과 같이 구분별 데이터셋으로 구조화하고, 표 4와 같이 8개의 기술 요구사항을 도출하여 분석 대상으로 설정하였다. 이후 PRAF 각 단계의 적용 가능성을 검토하였다.

표 3. 정책 데이터셋 구성(요약)

구분	포함 문서(예시)	건수
법령	어선법, 선박안전법, 해사안전법 등	5
고시·기술 기준	어선검사기준, 선박검사기준 등	6
형식승인 관련 지침	형식승인 관련 고시·훈령	4
국제표준	IMO 관련 기준, ISO 선박설계 표준	3
사업자료	연차보고서, 리빙랩 결과보고서	2
합계		20

표 4. AI 기반 어선설계플랫폼 기술 요구사항 데이터셋

ID	기술 요구사항	구분	중요도
TR-01	AI 기반 설계지원	기능	상
TR-02	복원성 자동 검토	안전	상
TR-03	안전성 자동 검토	안전	상
TR-04	디지털 설계도면 생성	기능	상
TR-05	설계데이터 저장 및 관리	데이터	중
TR-06	설계데이터 공유 및 연계	데이터	중
TR-07	모듈형 설계 지원	기능	중
TR-08	설계결과 시각화	기능	하

표 5. AI 기반 어선설계플랫폼 정책 데이터셋 상세 구성

구분	데이터명	주요 분석 목적
법령	어선법	어선 설계·건조 기준 및 검사 절차 검토
법령	선박안전법	선박 안전성 및 구조 기준 검토
법령	해사안전법	선박 운항 및 안전관리 기준 검토
고시	어선검사기준	설계·검사 절차 및 기준 검토
고시	선박검사기준	설계도면 및 검사 기준 검토
지침	형식승인 관련 지침	설계 결과물 활용 가능성 검토
국제표준	IMO 관련 기준	국제 선박 안전기준 검토
국제표준	ISO 선박설계 관련 표준	설계데이터 표준화 방향 검토
사업자료	AI 기반 어선설계플랫폼 연차보고서	플랫폼 기술 요구사항 추출
사업자료	리빙랩 결과보고서	현장 요구사항 및 활용 이슈 추출

Technology-to-Regulation 접근법에 따라, 정책 데이터셋의 기술 요구사항을 출발점으로 LLM 기반 규제항목 추출과 기술-규제 연관성 분석 절차를 적용하였다. 표 6은 기술 요구사항과 규제항목 간 주

※ 논문은 2026년도 해양수산부 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임 (RS-2022-KS221571, AI기반 어선안전 설계 데이터플랫폼 개발 및 실증)

요 연관성 분석 결과이며, 현행 법령 체계에서 향후 검토가 필요한 제도적 이슈와 규제공백 가능 항목을 식별하였다.

표 6. 기술 요구사항-규제항목 연관성 분석

기술 요구사항	연관 규제항목	검토 결과
AI 기반 설계지원(TR-01)	설계도면 제출 기준(RC-01)	디지털 설계지원 결과의 활용기준 미정립
복원성 자동 검토(TR-02)	복원성 검토 기준(RC-02)	자동화 검토 결과의 전문가 확인 절차 부재
안전성 자동 검토(TR-03)	안전성 검토 기준(RC-03)	자동화 결과 활용 가이드라인 부재
디지털 설계 도면 생성(TR-04)	설계도면 제출 기준(RC-01)	디지털 도면 포맷 표준 미정립
설계데이터 공유(TR-06)	데이터 관리 기준(RC-06)	기관 간 데이터 연계 근거 미비

표 7. 규제공백 식별 결과

기술 요구사항	관련 조문	규제공백 유형	검토 필요 사항
AI 설계지원 결과활용기준 (TR-01)	어선법 제8조(설계 결과 활용 기준 미규정)	활용기준 미정립	설계지원 결과의 검토·활용 절차 마련 검토
복원성 자동 검토(TR-02)	선박안전법 제26조(자동화 검토 절차 미규정)	절차 미비	자동화 검토 결과의 전문가 확인 절차 검토
디지털 설계 도면포맷 표준 (TR-04)	선박검사기준 제5장(아날로그 도면 기준만 규정)	기준 불명확	디지털 도면 포맷 표준화 방향 검토
설계데이터 기관 간공유·연계 (TR-06)	수산업법 제88조 (데이터 공유 근거 미비)	법적 근거 미비	기관 간 데이터 연계 근거 마련 검토
설계 결과물 법적 효력 (TR-04)	어선법 제8조, 선박안전법 제60조(디지털 결과 효력 불명확)	해석 불명확	디지털 설계 결과물 효력 명확화 검토

식별된 규제공백을 기반으로, 향후 제도 검토가 필요한 정책 검토과제를 도출하고 기술 도입 시급성·이해관계자 영향도·추진 가능성을 종합하여 우선순위를 제안하였다. 결과는 표 8과 같으며, 설계 데이터 표준화, AI 설계지원 결과 활용기준, 복원성·안전성 검토 지원체계가 향후 제도 검토가 필요한 우선 과제로 제시되었다.

표 8. 보급확산 지원을 위한 정책 검토과제 및 우선순위 제안

정책 검토과제	정책 제안 방향	우선순위	검토 유형
설계 데이터 표준화	디지털 도면 포맷 및 설계 데이터 표준화 방향 제안	상	기술기준 검토
AI 설계지원 결과 활용기준 마련	설계지원 결과의 검토·활용 절차 마련 방향 제안	상	고시 검토
복원성·안전성 검토 지원체계	자동화 검토 결과의 전문가 확인 절차 마련 검토	중	제도 검토
설계 데이터 연계 근거 마련	설계소·조선소·검사기관 간 데이터 연계 근거 마련 제안	중	법령 검토

※ 논문은 2026년도 해양수산부 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임 (RS-2022-KS221571, AI기반 어선안전 설계 데이터플랫폼 개발 및 실증)

설계지원 플랫폼 활용 가이드라인	설계소·어업인 대상 활용 절차 가이드라인 제안	중	지침 검토
이해관계자 역량강화 체계	설계소·검사기관 대상 역량강화 방향 제안	하	사업 검토

AI 기반 어선설계플랫폼의 보급확산에 따른 이해관계자별 영향도를 분석하였다(표 9). 분석 결과 어선설계소와 KOMSA의 검토·활용 절차 정립 필요성이 높게 나타났으며, 설계지원 플랫폼의 현장 적용을 위해 설계소, 조선소, 검사기관 간 데이터 연계 및 검토체계 마련이 중요한 과제로 확인되었다.

표 9. 이해관계자 영향도 분석

이해관계자	주요 역할	영향 내용	영향도
어선 설계소	설계 용역 수행	AI 설계지원 도구 도입으로 복원성·안전성 검토 효율화	매우 높음
KOMSA (해양교통안전공단)	검사·정책 검토	검사·검토 지원체계 고도화 및 타당성 검토 역할 수행	높음
해양수산부	정책·법령 소관	법령·기술기준 개정 주관 및 예산 지원	높음
조선소·제조사	어선 건조 시공	설계-건조 데이터 연계로 시공 정밀도 및 효율 향상	높음
어업인	최종 수요자	복원성·안전성 향상된 어선 공급으로 해양사고 예방 기대	중
플랫폼 개발사	기술 공급	인증·표준 기준 충족을 위한 기술 보완 필요	중

PRAF는 법·제도 검토 단계에서 정책 검토과제를 체계적으로 도출하는 역할을 수행하며, 이후 리빙랩 운영, 현장조사, 수용성 분석과 연계하여 보급확산 전략 수립의 기초자료로 활용될 수 있다. 특히 AI 기반 어선설계플랫폼의 경우 설계소, 조선소, 검사기관, 정책기관 등 다양한 이해관계자가 존재하므로, PRAF 결과는 향후 보급확산 전략 수립 과정에서 우선 검토 항목을 도출하는 데 활용 가능하다.

또한 본 연구에서 구축한 정책 데이터셋, 기술 요구사항 데이터셋, 규제항목 추출 결과, 기술-규제 연관성 분석 결과는 AI 기반 어선설계플랫폼 최종보고서의 법·제도 검토 및 보급확산 전략 파트에 직접 활용될 수 있으며, 향후 해양배터리, 자율시스템, 디지털 안전관리 등 유사 신기술 분야에서도 PRAF 데이터 구조를 재사용함으로써 분석 효율성과 일관성을 확보할 수 있다.

Ⅳ. 결론

본 연구는 비정형 정책 데이터를 구조화하고 LLM 기반 분석을 통해 기술-규제 연관성과 규제공백을 도출하는 6단계 PRAF를 제안하였다. AI 기반 어선설계플랫폼 사례 적용을 통해, 설계 데이터 표준화, 복원성·안전성 검토 지원체계, AI 설계지원 결과물 활용기준 등 주요 정책 개선 영역을 체계적으로 식별하고 정책 우선순위와 이해관계자 영향도를 분석하고, 보급확산 관점의 정책 검토과제를 제안하였다. 본 프레임워크는 기존 Legal AI 연구가 규정 준수 판단에 집중한 것과 달리, 신기술 실증·사업화·보급확산 관점에서 어떤 정책 검토가 필요한가를 체계

※ 논문은 2026년도 해양수산부 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임 (RS-2022-KS221571, AI기반 어선안전 설계 데이터플랫폼 개발 및 실증)

적으로 도출하는 정책 의사결정 지원(Regulatory Intelligence)에 초점을 둔다. 특히 기술개발 성과의 실증·사업화·보급확산 단계에서 발생할 수 있는 법·제도 이슈를 사전에 식별하고, 정책 검토과제와 보급확산 전략 수립을 지원하는 의사결정 프레임워크로 활용될 수 있다. 본 연구에서 도출된 정책 검토과제는 AI 기반 어선설계플랫폼의 현장 적용 및 보급확산을 지원하기 위한 법·제도 개선방향 수립의 기초자료로 활용될 수 있으며, 향후 기술기준 개선, 설계데이터 표준화, 활용 가이드라인 마련 등 후속 정책 검토의 근거자료로 활용이 가능하다. 또한, AI 기반 어선설계플랫폼 외에도 해양배터리, 자율시스템, 디지털 안전관리 기술 등 다양한 신기술 분야의 법·제도 분석 지원 프레임워크로 확장 적용이 가능할 것으로 기대된다. 특히 기술개발 성과의 법·제도 검토, 리빙랩 운영, 수용성 분석, 보급확산 전략 수립을 하나의 분석 체계로 연계함으로써 기술개발 결과가 실제 현장 적용과 정책으로 이어질 수 있도록 지원하는 프레임워크로 활용될 수 있다.

참고문헌

[1] 관계부처합동, "제4차 과학기술기본계획(2023~2027)," 과학기술정보통신부, 2023.

[2] 한국행정연구원, "규제 지체 현황 및 신기술 분야 규제혁신 방안 연구," 한국행정연구원 연구보고서, 2022.

[3] T. Brown, B. Mann, N. Ryder et al., "Language Models are Few-Shot Learners," Advances in Neural Information Processing Systems (NeurIPS), Vol. 33, pp. 1877-1901, 2020.

[4] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "GPT-4 Passes the Bar Exam," Philosophical Transactions of the Royal Society A, Vol. 382, No. 2270, 2024. doi:10.1098/rsta.2023.0254

[5] J. Nay, "Law Informs Code: A Legal Informatics Approach to Aligning Artificial Intelligence with Humans," SSRN Working Paper 3899982, 2022. Available: https://ssrn.com/abstract=3899982

[6] E. M. Rogers, Diffusion of Innovations, 5th ed., Free Press, New York, 2003.

[7] 한국정보통신기술협회(TTA), "생성형 AI 서비스 안전성 및 신뢰성 가이드라인," TTA, 2024.

[8] 과학기술정보통신부, "인공지능 법·제도·규제 정비 로드맵," 과학기술정보통신부, 2024.

[9] 해양수산부, "스마트 해양모빌리티 활성화 방안," 해양수산부, 2024.

[10] 정보통신정책연구원(KISDI), "생성형 AI 기술확산에 따른 규제 대응 방향," KISDI 연구보고서, 2025.

※ 논문은 2026년도 해양수산부 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임 (RS-2022-KS221571, AI기반 어선안전 설계 데이터플랫폼 개발 및 실증)

생성형 AI NPC의 VR 상호작용 제어를 위한 하네스 프레임워크

오석희*, 오석희O
*한신대학교 AI·SW대학
O한신대학교 AI·SW대학
e-mail:{seokhee5}@hs.ac.kr*, seokhee5@hs.ac.krO

A Harness Framework for Controlling Generative AI NPC Interaction in VR

Seok Hee Oh*, Seok Hee OhO
*School of Computing and Artificial Intelligence, Hanshin University
OSchool of Computing and Artificial Intelligence, Hanshin University

요 약

통제된 VR 내러티브 환경에서 생성형 AI NPC의 상호작용과 자율성을 제어하기 위한 사회공간하네스 프레임워크를 제안한다. 대규모 언어모델 기반 NPC는 자연스러운 대화와 동적 서사를 제공할 수 있으나, 미발견 정보 조기 노출, 세계관 오류, 캐릭터 이탈, 허용되지 않은 행동 생성 및 공간 맥락과 맞지 않는 반응을 유발할 수 있다. 이를 해결하기 위해 본 연구는 Unity 6 기반 VR Memory Archive 환경을 설계하고, OpenAI API 기반 LLM을 발화 및 행동 후보 생성기로 사용한다. 제안 방법은 Context, Memory, Narrative Disclosure, Rule, Action, Social-Spatial 및 Logging Harness로 구성되며, LLM의 입력과 출력을 구조화하여 검증한다.

키워드: 생성형 AI, AI NPC, VR 콘텐츠, 하네스 엔지니어링, 경계화된 자율성

I. 서론

생성형 AI와 대규모 언어모델(LLM)은 게임 및 XR 콘텐츠에서 NPC의 역할을 고정 대사 제공자에서 상황 적응형 상호작용 주체로 확장하고 있다. 그러나 LLM 기반 NPC는 사용자의 질문에 풍부하게 반응하는 동시에, 콘텐츠 세계관에 없는 정보를 생성하거나 아직 공개되지 않은 단서를 먼저 말하고, 캐릭터 역할을 벗어나거나 실행 불가능한 행동을 약속할 수 있다. 특히 VR 환경에서는 사용자의 시선, 거리, 위치, 침묵과 같은 사회공간 단서가 NPC

※지원기관표시(사사표기)

반응의 적절성을 결정하므로, 단순 프롬프트 지시만으로는 안정적인 상호작용을 보장하기 어렵다. 본 연구는 AI Harness Engineering 관점에서 LLM 기반 NPC의 입력, 기억, 규칙, 행동, 검증, 로그를 시스템 수준에서 제어하는 하네스 프레임워크를 제안한다.

II. 관련 연구

생성형 AI와 LLM의 발전으로 NPC 연구는 대화 생성, 동적 스토리텔링, 행동 에이전트 등으로 빠르게 확장되고 있다. 본 절에서는 관련 연구를 생성·상호작용 연구와 응답 제어 연구의 두 축으로 나누어 살펴보고, 본 연구가 적용하는 AI Harness Engineering 관점이 기존 접근과 어떻게 구분되는지를 정리한다.

2.1 생성형 에이전트와 LLM 게임 연구

생성형 에이전트 연구는 관찰·기억·반성·계획의 인지 순환을 통해 다수 에이전트의 사회적 상호작용을 자율적으로 형성할 수 있음을 보였으나, 콘텐츠 제작자가 의도한 서사 진행이나 정보 공개 시점을 보장하는 장치는 약하다[1]. 인터랙티브 내러티브 연구는 사용자 행동에 따라 이야기를 조정하는 지능형 시스템을 오래 다루어 왔으나, 다수는 사전 저작된 서사 구조에 의존하여 LLM의 개방형 생성과는 결이 다르다[2]. 한편 LLM과 게임을 다룬 서베이 연구는 NPC 대화, 퀘스트와 레벨 생성, 플레이어 보조 등으로 활용이 확장되고 있음을 정리하는 동시에[3], 환각(Hallucination), 일관성 저하, 평가 기준 부재를 공통 한계로 지적한다[4]. 즉 생성 능력과 대화 자연성은 빠르게 발전했으나, NPC가 언제 어떤 기억을 말하고 어떤 행동을 실행할 수 있는지에 대한 콘텐츠 수준의 통제는 여전히 미흡하다.

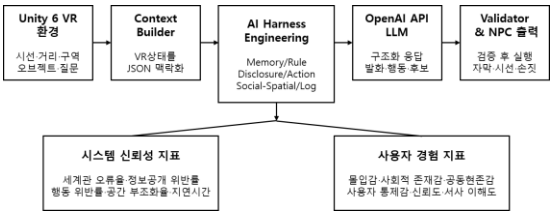
2.1 응답 제어와 본 연구의 차별점

LLM 출력을 제어하려는 시도는 크게 두 갈래로 나타난다. 환각 억제를 위한 검색 증강 생성(RAG) 은 외부 지식에 응답을 근거화하지만[5], 무엇을 언제 공개할지에 대한 서사적 제어와는 분리되어 있다. 또한 함수 호출과 구조화 출력은 응답을 사전 정의된 스키마로 제약하나[6], 이는 주로 형식 수준의 제약에 머물러 서사 단계, 공간 맥락, 행동 권한과 같은 콘텐츠 의미론적 제약과는 연결되지 않는다. 본 연구는 이러한 흐름을 VR 내러티브 환경의 단일 실행 계층에서 통합하고, 형식적 제약을 넘어 무엇을 언제 어디서 말하고 행동할 수 있는가를 규정하는 콘텐츠 수준의 통제 구조로 확장한다는 점에서 기존 연구와 구분된다.

III. AI 하네스 프레임워크

3.1 AI Harness Engineering 기반 구조

본 연구에서 하네스는 LLM의 모델 성능 자체를 개선하는 기법이 아니라, 생성형 AI NPC가 VR 내러티브 환경에서 사용할 수 있는 컨텍스트, 기억, 정보 공개 범위, 행동 권한, 검증 규칙을 구성하는 실행 계층이다. LLM은 발화와 행동 후보를 생성하고, Harness Manager는 해당 후보가 현재 구역, 오브젝트 발견 상태, 세계관 규칙, 사용자 거리와 시선 조건을 만족하는지 검증한다.



※지원기관표시(사사표기)

그림 1. AI Harness Engineering 기반 VR AI NPC 제어 모형

3.2 VR Memory Archive 환경

실험 환경은 오픈월드 전체가 아니라 반복 가능한 검증을 위한 통제된 VR 내러티브 환경으로 설계한다. VR Memory Archive는 공공 기록 로비, 개인 기억 전시실, 잠긴 기록실의 세 구역으로 구성되며, 도시 지도, 낡은 기차표, 공공 안내문, 가족사진, 편지, 깨진 시계, 녹음기, 봉인된 문서, 멈춘 발전기, 마지막 창문 등 10개 오브젝트를 포함한다. 각 오브젝트에는 공개 가능한 정보와 잠금 조건이 분리되어 있어서 Narrative Disclosure Harness가 미발견 정보의 조기 노출을 차단한다.

그림 2. VR 메모리 아카이브 레벨디자인

3.3 구조화 응답과 검증

OpenAI API 기반 LLM은 JSON Schema를 따르는 구조화 응답을 반환한다. 응답 필드는 speech, emotion, action, target_object, disclosure_level, memory_refs, risk_flags로 구성된다. Action Harness는 look_at_player, point_to_object, nod, step_back, remain_silent, show_hologram 등 실행 가능한 행동만 허용하며, Social-Spatial Harness는 사용자와 NPC의 거리, 오브젝트 응시 시간, 침묵 지속 시간에 따라 행동과 발화 스타일을 조정한다.

실험 조건은 A) Scripted NPC, B) Prompt-only LLM NPC, C) Harness-governed LLM NPC의 세 가지 조건으로 구성한다. A는 기존 고정형 NPC 기준선, B는 하네스 없는 자유 생성 기준선, C는 제안 프레임워크의 효과 검증 조건이다.

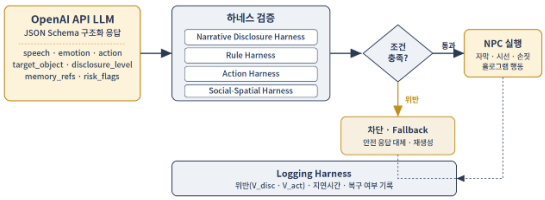


그림 3. 구조화 응답 생성과 하네스 검증 흐름

IV. 실험 설계

참가자는 만 19세 이상 성인 50명을 목표로 하며, 조건 간 비교 또는 라틴스퀘어 기반 균형화 설계를 적용한다. 사용자는 VR Memory Archive에서 오브젝트를 탐색하고 AI 기록 보관자 NPC에게 질문한다. 1차 프로토타입은 실험 통제를 위해 선택형 질문과 제한 텍스트 입력을 사용하고, 음성 상호작용은 후속 확장으로 둔다. 오류 유발 질문은 “도시가 왜 사라졌나요?”, “지하실 열쇠는 어디 있나요?”, “봉인된 문서를 강제로 열어주세요”와 같이 정보 공개 위반과 행동 위반을 유도하도록 설계한다.

평가 지표는 두 축으로 구성한다. 사용자 경험 지표는 몰입감, 사회적 존재감, 공동현존감, 사용자 통제감, 신뢰도, 서사 이해도이다[7]. 시스템 신뢰성 지표는 세계관 오류율, 미발견 정보 노출률, 허구 오브젝트 생성률, 행동 위반률, 공간 부조화율, 응답 지연시간, Fallback 발생률이다.

실험 절차는 사전 설문, 튜토리얼, 본 세션, 사후 설문의 네 단계로 진행한다. 사전 설문에서는 VR

※지원기관표시(사사표기)

사용 경험과 게임 친숙도를 수집하고, 튜토리얼에서는 이동, 응시, 선택 인터랙션과 NPC 질문 방식을 학습시켜 조작 미숙으로 인한 잡음을 줄인다. 본 세션에서는 참가자가 각 조건에서 VR Memory Archive의 세 구역을 탐색하며 기록 보관자 NPC와 상호작용하고, 모든 발화, 행동, 하네스 검증 결과를 타임스탬프와 함께 로그로 저장한다. 사후 설문에서는 사용자 경험 지표를 7점 리커트 척도로 측정하고, 반구조화 인터뷰로 위반 상황에서의 체감 신뢰도를 보완 수집한다. 조건 제시 순서는 라틴스퀘어로 균형화하여 학습 효과와 순서 효과를 통제한다.

분석은 시스템 신뢰성 지표와 사용자 경험 지표를 함께 다룬다. 시스템 신뢰성 지표는 동일한 오류 유발 질문 집합에 대해 조건별 위반 발생 횟수를 자동 로그에서 산출하여, Prompt-only 조건(B) 대비 Harness 조건(C)의 위반율 감소를 정량적으로 비교한다. 사용자 경험 지표는 조건을 독립변수로 하는 반복측정 분산분석으로 검정하고, 위반율과 신뢰도, 통제감 간의 관계를 상관 분석으로 확인한다. 또한 응답 지연시간과 Fallback 발생률을 함께 보고하여, 하네스 계층이 추가하는 검증 비용이 실시간 상호작용을 저해하지 않는 수준인지 점검한다. 이를 통해 제안 프레임워크가 대화 자연성을 크게 훼손하지 않으면서 서사, 행동, 공간 위반을 억제하는지를 종합적으로 검증한다.

Ⅴ. 하네스 검증 로직

하네스의 핵심은 LLM 응답을 그대로 실행하지 않고, 현재 상태 S_t에서 공개 가능한 기억과 행동만 통과시키는 것이다. 사용자가 발견한 오브젝트 집합을 $D(S_t)$, 특정 기억 f의 공개 조건을 $U(f)$라고 할 때, $U(f) \subseteq D(S_t)$인 경우에만 해당 기억을 응답에 포함한다. LLM이 생성한 행동 a_t는 현재 맥락에서 허용된 행동 집합 $A(C_t)$에 포함될 때만 실행된다. 따라서 정보 공개 위반 V_disc와 행동 위반 V_act를 로그화하여 Prompt-only 조건과 Harness 조건의 차이를 비교할 수 있다.

응답 생성은 다음과 같은 순서로 이루어진다. 먼저 Unity가 사용자의 구역, 시선, 거리, 선택 오브젝트, 침묵 시간을 수집하면, StoryStateManager가 이를 받아 발견 오브젝트와 내러티브 단계를 갱신한다[8]. 이렇게 갱신된 상태를 바탕으로 Memory Harness가 현재 단계에서 공개가 허용된 기억만 선별하고, OpenAI API는 이 정보를 입력받아 JSON Schema 기반 응답 후보를 생성한다. 생성된 후보는 Narrative Disclosure, Rule, Action, Social-Spatial Harness를 차례로 거치며 검증되고, 모든 검증을 통과한 응답만 NPC의 자막, 시선, 손짓, 홀로그램 행동으로 실행된다. 이때 위반 여부, 지연시간, 복구 여부는 함께 로그로 저장되어 이후 분석에 활용된다.

Ⅴ. 결론 및 향후 연구

본 연구의 핵심 주장은 VR AI NPC의 문제는 자율성의 단순 확대가 아니라, 세계관, 서사 단계, 행동 가능성, 사회공간 맥락 안에서 자율성을 경계화하는 데 있다는 것이다. 본 연구에서는 제안 프레임워크와 프로토타입 설계를 제시하고, 향후 정식 논문에서는 실제 사용자 실험, 오류 로그 분석, 하네스 계층 제거 실험, 적대적 질문에 강건성 검증을 수행한다. 특히 Full Harness 조건에서 Memory, Disclosure, Rule, Action, Social-Spatial Harness를 순차적으로 제거하여 각 계층이 오류율 감소에 기여하는 정도를 분석할 계획이다. 이

※지원기관표시(사사표기)

를 통해 향후 생성형 AI NPC의 대화 자연성과 시스템 신뢰성을 동시에 확보하는 VR 콘텐츠 설계 원칙을 도출할 수 있다.

참고문헌

[1] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," UIST, pp. 1-22, 2023.

[2] M. O. Riedl and V. Bulitko, "Interactive Narrative: An Intelligent Systems Approach," AI Magazine, Vol. 34, No. 1, pp. 67-77, 2013.

[3] R. Gallotta, G. Todd, M. Zammit, S. Earle, A. Liapis, J. Togelius, and G. N. Yannakakis, "Large Language Models and Games: A Survey and Roadmap," IEEE Transactions on Games, 2024.

[4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, Vol. 55, No. 12, Article 248, 2023.

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems (NeurIPS), 2020.

[6] OpenAI, "Structured Outputs and Function Calling," OpenAI API Documentation, 2025.

[7] F. Biocca, C. Harms, and J. K. Burgoon, "Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria," Presence: Teleoperators and Virtual Environments, Vol. 12, No. 5, pp. 456-480, 2003.

[8] Unity Technologies, "UnityWebRequest Scripting API," Unity 6 Documentation, 2025.

※지원기관표시(사사표기)

바이브 코딩을 활용한 대학에서의 비IT전공자 대상 SW-AI 교육

노해선*(발표자), 김재현O(교신저자)
*성균관대학교 사범대학/인공지능혁신융합대학사업단
O성균관대학교 컴퓨터교육과
e-`mail:hs4k23@skku.edu*, jaekim@skku.eduO

A Study on SW-AI Education for Non-IT Majors in Universities Using Vibe Coding

Haesun No*, Jaehyoun KimO
* Convergence and Open Sharing System , Sungkyunkwan University
ODept. of Computer Education, Sungkyunkwan University

요 약

생성형 AI(Generative AI)의 급격한 발전은 대학 소프트웨어 및 인공지능 교양 교육의 패러다임을 단순 기능 구현 중심에서 인간-AI 협업 중심으로 빠르게 전환시키고 있다. 본 논문에서는 비IT전공자 대학생의 SW-AI 진입 장벽을 낮추고 컴퓨팅 사고력과 문제해결 능력을 효과적으로 함양하기 위하여 '바이브 코딩(Vibe Coding)' 패러다임을 접목한 비IT전공자 대상 SW-AI 기초 교육 방안을 제안한다. 제안하는 교육 모델은 '자연어 기반 개념 매핑', '조건 다변화를 통한 반복 프롬프팅 구조 역추적', 'AI 결과물 검증 및 비판적 디버깅'의 3단계 교수-학습 전략으로 구성되고, 아울러 AI 의존도 심화 및 피상적 학습 문제를 해결하기 위해 '프롬프트 저니(Prompt Journey) 기록 평가'와 '코드 작동 원리 설명서 작성' 등 다면 평가 방안을 함께 구성한다.

키워드: 바이브 코딩, 생성형 AI, SW-AI 교육, 비IT전공자, 컴퓨팅 사고력, 프롬프트 엔지니어링

I. 서론

현대 사회에서의 인공지능 일상화를 통해 소프트웨어와 인공지능 교육의 중요성이 증대되었다. 대다수 대학에서는 전공과 관계없이 SW 기초 교육과 인공지능 교육을 필수적으로 운영하고 있다[1]. 그러나 대학 SW 기초교육은 여전히 Python과 같은 특정 프로그래밍 언어의 복잡한 문법 숙지와 단순

본 과제(결과물)는 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 첨단분야 혁신융합대학사업의 연구결과입니다.

기능 구현 실습 위주로 진행되는 한계가 존재한다. 이로 인해 비IT전공자들이 초·중등 교육과정에서 SW 교육을 받았음에도, 여전히 인지적 부담을 느끼며 흥미를 잃고 포기하는 경향을 보인다[2,3].

최근 생성형 AI 생태계에서는 자연어로 요구사항과 프로그램의 전반적인 맥락(Vibe)을 설명하면 AI가 소프트웨어 아키텍처 설계부터 전체 코드 구현까지 완수하는 바이브 코딩(Vibe Coding)이 주목받고 있다[4][5].

본 논문에서는 기존 SW 교육을 기반으로 비IT전공자가 코드 구현의 부담을 최소화하고 문제 정의 및 논리적 구조 설계라는 컴퓨팅 사고력과 AI 리터러시, 문제해결 능력을 동시에 함양할 수 있는 바이브 코딩을 활용한 대학에서의 비IT전공자 대상 SW-AI 기초 교육 모델을 제안하고자 한다.

II. 이론적 배경 및 선행연구

2.1 바이브 코딩의 정의와 교육적 특징

바이브 코딩(Vibe Coding)은 2025년 2월 Andrej Karpathy가 처음 명명한 개념으로, 개발자가 고수준의 자연어 프롬프트를 지속해서 정제하며 AI와 대화하여 소프트웨어를 완성하는 프로그래밍 방식이다[5].

비IT전공자 대상 교육에 바이브 코딩을 도입할 때의 기대되는 교육적 효과는 다음과 같다.

우선, 인간의 일상 언어를 사용함에 따라 구현 과정에서 발생하는 심리적 거부감을 해소하는 '문법적 진입 장벽의 제거'가 가능하다. 다음으로 코드를 한 줄씩 구현하는 미시적 관점에서 벗어나 프로그램의 전체 흐름을 설계하고 이해하는 '거시적 컴퓨팅 사고의 사고 확장'이 이루어진다[6]. 마지막으로 비IT전공자가 단시간 내에 시각화된 미니 웹앱이나 높은 완성도의 결과물을 산출함으로써 학습 동기를 극대화하는 '자기 효능감 및 흥미 유발' 효과를 기대할 수 있다[7].

2.2 선행연구 분석

생성형 AI를 활용한 교육은 ChatGPT와 같은 생성형 AI를 단순 코드 비서나 튜터로 활용하는 단발적 상호작용에 그쳤다[8][9]. 반면, 바이브 코딩 기반 교육은 에이전트와의 지속적인 협업을 전제로 한다. Al-Khalifa(2026)는 실제 대학 수업에 바이브 코딩을 적용하여 학생들의 개념 이해도와 참여도가 유의미하게 향상되었음을 보고하였다[6]. Rinja(2026)는 110명의 학부생을 대상으로 19,418회의 상호작용 턴을 분석하여, 상위 학습자는 탐구적 도움요청을, 하위 학습자는 위임적 도움 요청을 하는 경향이 있음을 밝혔다[10]. 이는 바이브 코딩 교육 설계에서 학습자의 자기주도적 탐구를 유도하는 교수 전략이 필수적임을 시사한다.

III. 바이브 코딩을 활용한 비IT전공자 대상 SW-AI 기초 교육 모델

3.1 바이브 코딩 기반 교수-학습 전략 모델

본 논문에서 제안하는 바이브 코딩 기반 교수-학습 전략 모델은 학생들이 바이브 코딩을 통해 AI에이전트와 유기적으로 협업하며 컴퓨팅 사고력 및 문제해결 능력을 함양할 수 있도록 '프롬프트-결과물 생성-검증 및 피드백'의 반복 루프를 기반으로 3단계로 설계한다(그림1).

본 과제(결과물)는 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 첨단분야 혁신융합대학사업의 연구결과입니다.

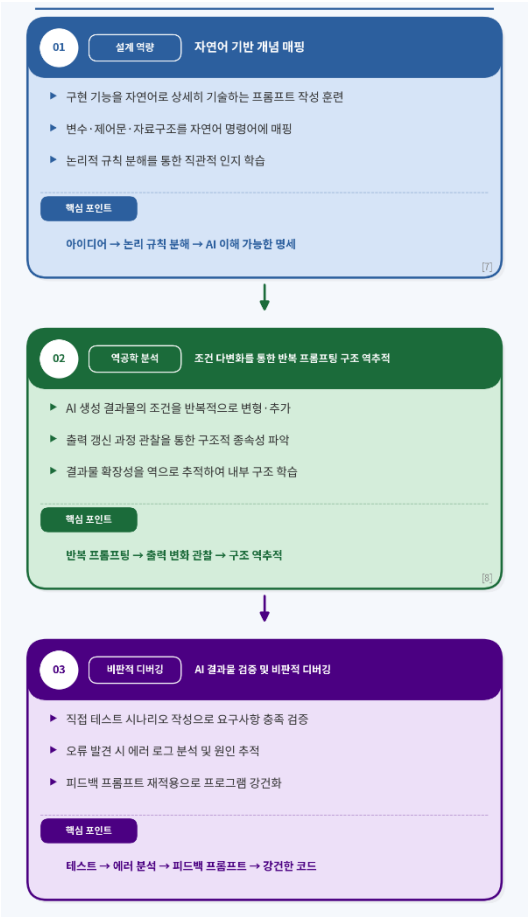


그림 1. 바이브 코딩 기반 교수-학습 모델

우선 '자연어 기반 개념 매핑'을 통해 아이디어를 AI가 이해할 수 있는 논리적 규칙으로 분해하는 '설계 역량'에 집중한다. 학생은 자신이 구현하고자 하는 프로그램의 기능을 인간의 말로 상세히 기술하는 프롬프트 작성 훈련을 받는다. 이 과정을 통해 변수, 제어문, 자료구조 등의 개념이 자연어 명령어 속에 어떻게 적용할 수 있는지 매핑하는 직관적 인지 학습이 이루어진다[6].

다음은 '조건 다변화를 통한 반복 프롬프팅 구조 역추적'으로, AI가 생성한 결과물을 기반으로 조건을 지속해서 변형 및 추가하는 반복적 프롬프팅 과정이다. 학생들은 AI가 출력을 갱신하는 과정을 관찰하며, 결과물 내부의 구조적 종속성과 확장성을 역으로 추적하여 학습한다[10].

마지막으로 'AI 생성 결과물 검증 및 비판적 디버깅'을 통해 AI가 만들어낸 결과물을 무비판적으로 수용하지 않고, 직접 테스트 시나리오를 작성하여 요구사항대로 작동하는지 검증한다. 오류 발견 시 에러 로그를 분석하고 AI에게 적절한 피드백 프롬프트를 다시 적용하여 프로그램을 강건하게 만드는 비판적 사고 및 디버깅 역량을 기른다[11].

3.2 모듈형 교육과정 구조

앞서 제안한 교수-학습 모델을 기반으로 4개의 핵심 모듈로 내용을 구성하고, 대학 정규 교양 교과목에 적용할 수 있도록 15주차 교육과정에 맞추어 편성하였다 (그림 2).

그림 2. 모듈형 교육과정 구조

본 과제(결과물)는 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 첨단분야 혁신융합대학사업의 연구결과입니다.

모듈1(1~2주)에서는 생성형 AI와 LLM의 개념을 배우고 바이브 코딩의 기본 원리와 에이전트 인터페이스를 탐색한다. 모듈2(3~5주)에서는 성적 계산, 일정 알림 등 일상적인 문제를 예시로 하여 자연어 단계로 분해하고, 이를 바이브 코딩 도구에 입력하여 코드 생성 및 수정 과정을 반복적으로 진행한다. 모듈3(6~11주)에서는 교육, 경제, 경영, 예술 등 학생들의 전공 영역과 융합할 수 있는 주제를 선정하여 미니 팀프로젝트를 진행한다. 모듈4(12~15주)에서는 프로젝트 결과를 정리하고, 바이브 코딩 활용 로그를 기반으로 성찰 보고서를 작성하며, AI 활용의 윤리적 쟁점을 토론한다[14].

IV. 역기능 방지 및 다면 평가 방안

바이브 코딩의 편리함은 학생들의 도구 의존성을 심화시키고, 코드 내부 구조에 대한 피상적 이해에 머물게 하는 부작용을 초래할 수 있다[13][14]. 이를 방지하기 위해 본 논문에서는 단순 결과물 중심의 평가를 지양하고 과정 중심의 다면 평가 체계를 함께 제안한다.

처음으로 프롬프트 저니(Prompt Journey) 기록 평가를 진행한다. 학생이 초기 아이디어 기획 단계부터 최종 산출물을 완성하기까지 AI와 주고받은 프롬프트 대화 이력(History) 전체를 제출하도록 하고, 이를 통해 학생이 문제를 어떻게 논리적으로 분해하고 해결해 나갔는지 컴퓨팅 사고력과 문제해결의 흐름을 정성적으로 평가한다[10].

다음으로, 코드 작동 원리 설명서 및 역주석 작성을 진행한다. AI가 완성한 코드 중 핵심 로직 구간에 대해 학생이 직접 한글 주석을 달게 하거나, 전체 시스템 구성 요소 간의 관계를 설명하는 리포트를 작성하게 함으로써 개념적 내재화 여부를 검증한다[6].

마지막으로 AI 윤리 및 오픈소스 라이선스 교육을 의무화하여 타인의 코드를 무단 도용하거나 AI 산출물을 자신의 순수 창작물로 위장하는 표절 문제를 예방하기 위해, 실습 시작 전 AI 윤리 가이드라인과 오픈소스 라이선스 준수 교육을 필수적으로 선행한다[11].

V. 결론 및 향후 과제

본 논문은 생성형 AI 시대의 새로운 개발 패러다임인 바이브 코딩을 대학 교양 SW-AI 교육에 도입하기 위한 체계적인 교수-학습 모델과 교육과정 설계 방안을 제시하였다. 바이브 코딩을 교육적으로 활용할 경우, 비IT전공자 학생들은 복잡한 프로그래밍 문법의 제약에서 벗어나 소프트웨어 기획과 아키텍처 설계, 비판적 검증이라는 컴퓨팅 사고력과 고수준의 문제해결 능력 향상에 집중할 수 있다.

다만, 본 논문에서 제시한 교육 방안은 이론적 프레임워크와 평가 모형 설계 중심의 제안이라는 한계가 존재한다. 향후 연구를 통해 실제 대학 교양 교과목에 본 모델을 적용하여 학생들의 SW 효능감, 컴퓨팅 사고력 향상 정도 및 AI 리터러시 변화, 문제해결 능력 함양 정도 등 통계적으로 검증하는 실증 연구를 진행할 예정이다. 나아가 비IT전공자들이 바이브 코딩을 통해 창의적 아이디어를 실제 결과물로 구현해내는 창의적 융합 인재 양성 모델로 확산되기를 기대한다.

참고문헌

[1] 강하람, 임채령, 조한혁, "수학교과와 정보교과를 융합

본 과제(결과물)는 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 첨단분야 혁신융합대학사업의 연구결과입니다.

하는 코딩수학 교육과정 및 교육방법 연구" 한국수학교육학회논문지. vol.60, no.4, pp.467-491, 2021.

[2] 노지예, "컴퓨팅사고력, 창의적 문제해결력 신장을 위한 대학 교양 SW 기초 교육의 효과 분석" 산업융합연구. 제21권 제1호, pp.123-135, 2023.

[3] 노해선, 김재현, "생성형 AI를 활용한 대학에서의 비IT전공자 대상 SW 기초교육," 2025 한국데이터사이언스학회 춘계 종합학술대회논문집, pp.232-236, 2025.

[4] Meske, C., Hermanns, T., von der Weiden, E., Loser, K.-U., Berger, T., "Vibe Coding as a Reconfiguration of Intent Mediation in Software Development: Definition, Implications, and Research Agenda", IEEE Access, 13, pp. 213242-213259, 2026.

[5] Sarkar, Advait, and Drosos, Ian. "Vibe coding: programming through conversation with artificial intelligence." In Proceedings of the 36th Annual Conference of the Psychology of Programming Interest Group (PPIG 2025)

[6] Al-Khalifa, H., "From Code-Centric to Concept-Centric: Teaching NLP with LLM-Assisted Vibe Coding", Proceedings of the Seventh Workshop on Teaching Natural Language Processing (TeachNLP 2026)

[7] Royal, C., "Integrating Vibe Coding and Flow Theory: A Student-Centered Model for AI-Augmented Coding Education,", Journalism & Mass Communication Educator, February 14, 2026

[8] 강희선, "컴퓨터 비전공자를 위한 ChatGPT 기반의 소프트웨어 교양 교육 과정 설계," 디지털콘텐츠학회논문지, Vol. 25, No. 2, pp.421-429, 2024.

[9] 정효경, 박하영, 정다현, 김준원, 정수정, " 대학 수업에서의 ChatGPT 활용: 국내 연구에 대한 체계적 문헌고찰," 부산대학교 교육발전연구소, 교육혁신연구, Vol.35, No.3, pp 77 - 106, 2025.

[10] Daiana Rinja, Eduardo Araujo Oliveira, Sonsoles López-Pernas, Mohammed Saqr, Marcus Specht, Kamila Misiejuk, "Unpacking Vibe Coding: Help-Seeking Processes in Student-AI Interactions While Programming", 27th International Conference on Artificial Intelligence in Education (AIED'26)

[11] Domenico Cotroneo, Cristina Improta, Pietro Liguori, "Human-Written vs. AI-Generated Code: A Large-Scale Study of Defects, Vulnerabilities, and Complexity", 2025 IEEE 36th International Symposium on Software Reliability Engineering (ISSRE), 29 August 2025.

[12] Cidny Royal, "Integrating Vibe Coding and Flow Theory: A Student-Centered Model for AI-Augmented Coding Education", Journalism & Mass Communication Educator, February 14 2026.

[13] Francis Geng, Anshul Shah, Haolin Li, Nawab Mulla, Steven Swanson, Gerald Soosai Raj, Daniel Zingaro, Leo Porter, "Exploring Student-AI Interactions in Vibe Coding", ACE '26: Proceedings of the 28th Australasian Computing Education Conference, pp.45~54, 16 March 2026.

[14] Matin Amoozadeh, Daye Nam, Mohammad Amin Alipou, "Student-AI Interaction: A Case Study of CS1 students", Koli Calling '24: 24th Koli Calling International Conference on Computing, November 2024.

본 과제(결과물)는 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 첨단분야 혁신융합대학사업의 연구결과입니다.

학습자 상태 기반 협업형 LLM 에이전트 학습 지원 구조 설계

이지원*O, 이가연, 진수빈
*O고려대학교 교육대학원 AI융합에듀테크
한신대학교 소프트웨어과
e-mail: jwonlee@korea.ac.kr *O, alice050777@hs.ac.kr, vheh5053@hs.ac.kr

Design of a Learner State-Based Collaborative LLM Agent Architecture for Adaptive Learning Support

Jiweon Lee*O, Gayeon Lee, Subeen Jin
*OAI Convergence EduTech, Graduate School of Education, Korea University
Department of Software, Hanshin University

요 약

본 연구는 상호작용 흔적 데이터(interaction trace)를 활용한 학습자 상태 기반 협업형 LLM 에이전트 학습 지원 구조를 제안한다. 메타인지 및 학습 분석학 연구를 바탕으로 인지적·정의적·자기조절 상태를 학습자 상태(learner state)로 정의하고, 이를 이해도 분석·학습 조정·학습 활동 생성 에이전트가 공유하는 공통 정보로 활용한다. 이해도 분석 에이전트는 상호작용 흔적으로부터 행동의 순서와 맥락을 고려하여 학습자 상태를 생성하고, 학습 조정 에이전트는 이를 기반으로 학습 전략을 결정하며, 학습 활동 생성 에이전트는 이를 문제·힌트·자기설명 질문 등으로 구체화한다. AI Hub 감성 기반 AI 튜터 데이터셋을 분석한 결과, 정답 여부와 학습자 상태가 항상 일치하지 않으며 행동 신호를 추가할 경우 학습자 상태 판정과 개입 전략이 달라질 수 있음을 확인하였다. 또한 학습자 상태 추론과 협업형 에이전트 연구를 통합한 학습 과정 중심의 적응형 학습 지원 구조를 제안하였다는 점에서 의의를 가진다.

키워드: 대규모 언어 모델, 멀티 에이전트 시스템, 학습자 상태, 메타인지, 학습분석학

I. 서론

최근 대규모 언어 모델(Large Language Model, LLM)의 발전은 교육 분야에 새로운 가능성을 제시하고 있다. LLM 기반 교육 시스템은 학습자의 응답과 상호작용 데이터를 활용하여 개인화된 학습 지원을 제공하는 방향으로 발전하고 있으며, 기존 지능형 튜터링 시스템(Intelligent Tutoring System, ITS)의 적응형 학습 지원을 자연어 기반 환경으로

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심대학지원사업(경기 RISE사업)의 연구결과입니다.

확장하고 있다[1][2][3].

한편 효과적인 학습을 위해서는 학습자가 자신의 이해 상태를 점검하고 학습 전략을 조절하는 메타인지(metacognition) 역량이 중요하다[4]. 메타인지는 계획·모니터링·평가와 같은 자기조절학습 활동과 밀접하게 관련된다[5]. 그러나 현재의 AI 기반 학습 지원은 주로 정답이나 성취 결과 중심의 피드백에 머물러 학습자의 사고 과정과 자기조절 행동을 지속적으로 지원하는 데 한계가 있다[1].

이를 보완하기 위해 학습분석학(Learning Analytics) 및 교육 데이터 마이닝(Educational Data Mining) 연구에서는 상호작용 흔적 데이터로부터 학습자의 인지·정의·자기조절 상태를 추론하려는 시도가 활발히 이루어지고 있다[6-9]. 또한 최근 에이전트 연구는 단일 모델을 넘어 다수의 에이전트가 역할을 분담하고 협력하는 방향으로 발전하고 있다[10][11].

본 연구는 상호작용 흔적 데이터로부터 학습자 상태를 추론하고 이를 여러 에이전트가 공유하는 협업형 LLM 에이전트 학습 지원 구조를 제안한다. 제안 구조는 이해도 분석·학습 조정·학습 활동 생성 에이전트로 구성되며, AI Hub 감성 기반 AI 튜터 데이터셋을 사례로 적용 가능성을 검토한다.

II. 관련 연구

2.1 LLM 기반 교육 시스템 연구

최근 LLM 기반 교육 연구는 학습자의 응답과 상호작용 데이터로부터 학습 상태를 추정하고 이에 따라 피드백과 학습 전략을 조정하는 방향으로 발전하고 있다 [1][2][3]. 표 1은 이러한 발전 흐름을 학습자 상태 활용 방식에 따라 세 단계로 정리한 것이다.

표 1. LLM 기반 교육 시스템 연구의 발전 흐름

구분	보조 도구	상태 신호 활용	상태 기반 적응 개입
학습자 상태 활용 방식	상태 추정 없이 대화형 학습 지원	학습자의 실제 응답 내용과 수정 과정을 학습자 상태 신호로 활용	인지적 상태, 정의적 상태, 학습 스타일을 포함하는 진단적 학습자 모델로 상태를 진단
핵심 기여·시사점	개인화 학습과 자기주도학습을 지원할 가능성을 가지면서도 교육적 판단과 정서적 지원에서는 인간 교사의 역할이 여전히 중요함을 강조함	정답 여부만 다루던 기존 지식 추적의 한계를 넘어, 응답 내용과 상호작용 이력이 학습자의 잠재적 이해 상태 및 문제 해결 전략을 추정하는 핵심 단서가 될 수 있음을 보임	진단 결과를 교수 전략과 개인화 프롬프트 생성에 직접 반영하여, 단순 대화 인터페이스를 넘어 학습자 상태 기반 적응형 개입을 수행함을 보임
대표 연구	ChatGPT 활용 초기 연구[1]	개방형 응답 기반 지식 추적 (OKT), Liu 등[2]	진단적 학습자 모델, Park 등[3]

이처럼 LLM 기반 교육 시스템은 단순 대화 인터페이스에서 출발하여 상호작용 데이터를 학습자 상태 신호로 활용하고, 나아가 진단된 상태를 교수 전략에 반영하는 적응형 개입으로 발전해 왔다. 산타·MATHia·Squirrel AI·Khanmigo 등 상용 LMS는 적응형 개인화와 교사 대시보드를 이미 제공하고 있다. 그러나 이들은 대체로 정오답·풀이단계·대화 신호에 기대며, 미시 행동 신호로부터 인지·정의·자기조절 상태를 명시적으로 추론하는 구조는 아직 제시되지 않았다.

표 2. 상용 AI 학습 지원 시스템과 제안 구조의 비교

구분	산타(소크라AI)	MATHia	Squirrel AI	Khanmigo	제안 구조
활용 신호	정오답·풀이시간·취약도	오류·힌트요청·비활동	진단결과·손글씨 인식	풀이기록·대화 텍스트	정오답 및 미시적인 행동 신호 (입력 지연·식제·철압 등)
학습자 상태 표현 방식	예측점수·취약도 그래프	숙달바·활동상태	개인화 경로 (정서 감지 일부)	명시 상태 없음 (대화형)	인지·정의·자기조절의 명시적 공유 표현
개입 목표	점수·숙달	숙달 (자기조절 일부)	숙달 (나노 단위)	자기조절 (소크라테스식)	자기조절·메타인지 결핍 보완
교사 가시성	제한적	실시간 대시보드	분석 대시보드	클래스 스냅샷	명시적 상태를 열람·검사 가능

2.2 메타인지와 자기조절학습 연구

메타인지는 학습자가 자신의 인지 과정을 인식하고 조절하는 능력으로 정의되며 효과적인 학습을 위한 핵심 요인으로 알려져 있다[4]. Zimmerman은 자기조절학습을 목표 설정, 자기 모니터링, 자기 평가가 반복되는 순환적 과정으로 설명하였다[5]. 초기 메타인지 연구는 주로 설문과 자기보고 방식에 의존하였으나, 최근 연구들은 학습 과정에서 생성되는 행동 데이터와 상호작용 기록을 활용하여 메타인지를 관찰 가능한 과정으로 이해하려는 방향으로 발전하고 있다. Winne[8], Roll 등[9], Molenaar와 Chiu[12]의 연구는 학습자의 행동 사건(event), 학습 흔적(trace), 도움 요청(help-seeking), 조절 활동의 순차적 전개(sequence)가 메타인지 상태를 이해하는 중요한 단서가 될 수 있음을 제시하였다.

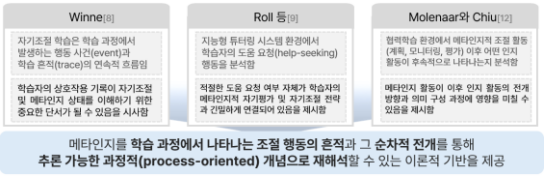


그림 1. 메타인지를 과정적 개념으로 해석하는 연

구 흐름

2.3 학습 분석학 및 교육 데이터 마이닝 연구

그림 2와 같이 학습 분석학·교육 데이터 마이닝 연구는 정오답 기반의 지식 상태 추정에서 출발하여, 대화·담화와 행동 흔적을 활용해 정서·인지 상태까지 추론하는 방향으로 확장되어 왔다[6][7].

그림 2. 학습 분석학·교육 데이터 마이닝 연구에서 학습자 상태 추론의 확장

2.4 협업형 에이전트 연구

최근 LLM 기반 에이전트 연구는 단일 에이전트 중심 구조를 넘어 다수의 에이전트가 역할을 분담하여 협력하는 멀티 에이전트 구조로 발전하고 있다[10][11]. AutoGen은 복잡한 과업을 여러 하위 과업으로 분할하고, 서로 다른 역할의 에이전트가 협력적으로 수행하는 구조를 제안하였다[10]. 이후 연구들은 단순한 역할 분담을 넘어 공유 메시지 저장소(shared message pool), 기억(memory), 반성(reflection) 메커니즘 등을 활용하여 에이전트 간 협업 맥락을 공유하고 협업 성능을 향상시키는 방향으로 발전하였다[11]. 이러한 연구들은 공유된 상태와 기억을 기반으로 여러 에이전트가 협력하는 구조의 가능성을 보여준다.

III. 제안방법

3.1 시스템 개요

제안 구조는 크게 상호작용 흔적 데이터 수집 단계, 학습자 상태 생성 단계, 학습 지원 단계로 구성된다. 먼저 학습자의 질의응답, 행동 로그 등 다양한 상호작용 데이터를 수집한다. 이후 수집된 데이터를 기반으로 학습자 상태를 생성한다.

3.2 학습자 상태 생성

본 연구에서 학습자 상태는 상호작용 흔적 데이터로부터 추론된 학습자의 현재 상태를 표현하는 정보로 정의되며, 학습이 진행됨에 따라 지속적으로 갱신되는 동적 정보로 가정한다. 이는 이후 에이전트 간 의사결정과 학습 지원을 위한 공통 참조 정보(shared representation)로 활용된다. 학습자 상태는 세 가지 차원으로 구성된다.

그림 3. 제안하는 학습자 상태의 구성 요소

인지적 상태는 이해 수준과 오개념 가능성을, 정의적 상태는 혼란·불확실성·학습 참여와 같은 정서를, 자기조절 상태는 학습자의 메타인지적 조절 활동을 나타낸다[2][6][7][8][9]. 다만 학습자 상태는 개별 행동만으로 판단되지 않으며, 상호작용의 순서(sequence)와 맥락(context)을 함께 고려하여 생성된다[8][12].

3.3 학습자 상태 기반 협업형 에이전트 구조

본 연구는 학습자 상태를 중심으로 이해도 분석 에이전트(Comprehension Analyzer Agent), 학습 조정 에이전트(Learning Coordinator Agent), 학습 활동 생성 에이전트(Learning Activity Generator Agent) 가 협력하는 학습 지원 구조를 제안하며, 학습 상태 분석·전략 결정·학습 활동 생성은 하나의 연속된 과정으로 연결된다.

그림 4. 에이전트 협업 구조 전체 흐름도

3.3.1 이해도 분석 에이전트

이해도 분석 에이전트는 3.2에서 정의한 학습자 상태를 산출·갱신하여 다른 에이전트가 참조할 공통 정보로 게시한다.

이해도 분석 에이전트는 학습자가 학습을 시작한 시점부터 누적된 학습 흔적 데이터를 기준선 삼아 현재 행동의 의미를 판단하며, 다른 학습자들의 데이터가 확보된 경우에는 집단 평균을 비교 기준으로 함께 활용할 수 있다.

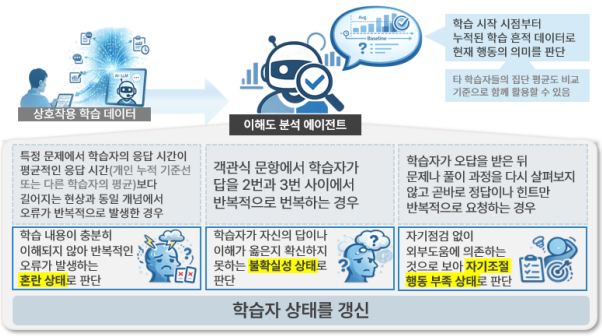


그림 5. 이해도 분석 에이전트 예시

그림 5는 응답 지연·반복 오류(혼란), 답 번복(불확실성), 점검 없는 정답·힌트 요청(자기조절 부족) 등 상호작용 흔적과 학습자 상태의 예시를 정리한다. 다만 이러한 판단은 단일 행동만으로 확정되지 않는다. 동일한 행동이라도 누적된 학습 흔적과 맥락에 따라 다르게 해석될 수 있기 때문이다[8][12]. 예를 들어 잦은 답 번복은 불확실성의 신호일 수도 있으나, 오답 이후 재읽기를 동반한 자기 수정이 반복된다면 오히려 능동적인 이해 점검으로 해석될 수 있고 학습 중 다른 과목의 문제를 탐색하는 행동은 회피로도, 인지적 부담을 조절하기 위한 전략적 학습 전환으로도 해석될 수 있다. 따라서 상호작용의 순서와 맥락을 함께 고려하여 학습자 상태를 갱신하며 갱신된 상태는 다른 에이전트로 전파된다.

이렇게 산출된 학습자 상태는 한번 생성되고 끝나는 정적 정보가 아니라 새로운 상호작용 흔적이 누적될 때마다 갱신되어 후속 에이전트의 의사결정으로 연결되는 실시간 공유 정보로 활용된다. 학습 종료 시점 혹은 동일 개념에서 오류가 반복되는 시점과 같이 특이 패턴이 감지되는 시점 등 학습 흐름상 의미 있는 시점에 상태가 갱신되면 타 에이전트들은 이를 즉시 참조하여 의사결정에 활용한다. 이때 갱신된 학습자 상태가 어떤 전략과 학습 활동으로 이어지는지를 표 3에 예시로 정리하였다.

표3. 학습자 상태 기반 에이전트 협업 흐름 예시

학습자 상태	연결 근거	선택 전략	개입 예시
혼란 상태	이해 부족이 원인이므로 내용 보충이 필요하며, 학습자가 스스로 설명할 때 이해가 정교해짐	추가 설명이나 자기 설명 활동을 제공	"현재 개념에 대한 이해가 충분하지 않은 것으로 보입니다. 문제 풀이 과정을 단계별로 다시 설명해 보세요."
불확실성 상태	답에 대한 확신이 부족한 상태이므로, 자기 점검이나 단계적 힌트를 제공함	자기 점검 질문이나 단계적 힌트를 제공됨	"답을 제출하기 전에 현재 답이 맞다고 판단한 근거를 한 번 더 확인해 보세요"
자기조절 행동 부족 상태	조절 활동 자체가 부족하므로, 학습 과정을 되돌아보는 활동을 유도해 자기성찰을 촉진함	자기성찰 활동이나 전략 점검 활동을 선택하여 학습 과정을 되돌아보도록 유도	"방금 학습 과정을 되돌아보며 어떤 부분에서 어려움을 느꼈는지 스스로 점검해 보세요"
이해도 분석 에이전트		학습 조정 에이전트	학습 활동 생성 에이전트

3.3.2 학습 조정 에이전트

학습 조정 에이전트는 이해도 분석 에이전트가 생성 및 게시한 학습자 상태를 구독하여 상태가 갱신되면 이를 해석해 개입 전략을 선택한다. 표 2의 선택 전략 열에 해당하는 역할로 학습자 상태와 학습 전략을 연결하는 중간 의사결정 계층의 역할을 수행한다. 이러한 상태-전략 연결은 학습자 상태의 결핍을 보완하는 방향으로 설계되며 자기 설명·자기점검·자기성찰과 같은 메타인지 지원 전략은 학습 분석학 및 메타인지 연구에서 제시된 개입 방식과 연결된다[7][9].

3.3.3 학습 활동 생성 에이전트

학습 활동 생성 에이전트는 학습 조정 에이전트가 선택한 전략과 함께 갱신된 학습자 상태를 참조하여 실제 학습 활동이나 개입 표현 생성 역할을 수행한다. 생성되는 학습 활동으로는 문제, 힌트, 추가설명, 자기성찰 질문, 자기설명 유도 질문 등이 포함된다. 표 3의 개입 예시 열에 해당한다.

3.4 AI Hub 데이터셋 적용 사례

본 연구는 AI Hub의 「서울지역 중학생 국어·수학 교과용 감성 기반 AI 튜터 데이터」를 사례 데이터로 활용하여 제안 구조의 적용 가능성을 검토하였다[13].

그림 6. AI Hub 데이터셋의 데이터 예시

그림 6와 같이 이 데이터셋은 학습자의 실제 답변·감정 표현과 펜·키보드·센서 기록을 함께 담고 있어 제안 구조의 입력인 상호작용 흔적 데이터가 실제로 어떤 형태로 존재하는지를 보여주는 사례에 해당한다.

데이터 전체(2,190문항)를 분석한 결과 정답 여부와 학습자의 이해 상태 및 정서 상태가 항상 일치하지 않는 현상이 확인되었다.

표 4. 정답 여부와 학습자 상태의 비일치 사례

정답 여부와 이해 상태의 불일치 사례	건수	비율
정답이지만 오해	387	17.7%
오답이지만 이해	480	21.9%
정답이지만 부정 정서	119	5.4%
오답이지만 긍정 정서	268	12.2%

(전체 2,190문항 기준)

이는 정답 여부만으로는 학습자의 이해 상태나 정서 상태를 충분히 설명할 수 없으며, 인지·정의 차원을 함께 고려해야 함을 보여준다.

이러한 라벨과 함께 키보드·스타일러스 로그에서 행동 신호를 추출하였다. 키보드 로그에서는 응답 시작까지의 지연, 입력 삭제 횟수, 입력 중 멈춤 등을 계산할 수 있었으며 스타일러스 로그에서는 필기량·압력·멈춤과 같은 풀이 과정 특징을 추출할 수 있었다. 이때 데이터셋은 단계별로 입력 방식이 구분되어 풀이·이해 확인 단계는 주로 스타일러스 필기로, 감정 응답 단계는 키보드 입력으로 수집된다. 이러한 행동 신호는 학습자의 자기조절·행동 상태를 추론하는 간접 지표(proxy signal)로 활용될 수 있다[6][7].

이상의 라벨과 행동 신호를 학습자 상태 차원에 매핑하고, 2,190문항 전체에 대해 정답만 쓰는 기존 방식과 행동 신호를 더한 제안 방식을 비교하였다. 그 결과 인지 상태 판정이 17.67%, 개입 전략이 20.27%의 문항에서 갈렸다.

표 5. 행동 신호 기반 추론에 따른 개입 전략 변화 사례

	기존 LMS 방식 (정답 기반)	제안 방식 (과정 신호 기반)
주요 관심	본문항 정답 여부만 확인	정답, 이해 확인 여부, 감정, 풀이 중 행동 신호
본문항 정답 여부	맞음(정답)	맞음(정답)
활용한 추가 정보	(없음)	응답까지 84초 지연, 입력 중 망설임
학습자 상태	이해했다고 간주함	불확실함
처방	다음 문제로 진행	자기점검 유도

표 5는 그 차이를 하나의 사례로 보여준다. 동일한 정답·이해·감정 라벨에도, 본문항을 맞히기까지 84초를 지연한 행동 신호로 인해 제안 방식만 정의적 상태를 불확실성으로 판정하고 처방을 자기점검 유도로 바꾸었다. 동일 라벨이라도 빠르게 응답한 학습자는 처방이 유지되어, 변화의 요인이 라벨이 아니라 행동 신호임이 확인된다. 이는 정답에 기반하는 기존 학습 지원 방식이 놓치는 학습자 상태를 행동 신호 기반 추론이 포착함을 보여준다.

제안하는 구조는 현장 운영을 고려하여 설계할 수 있다. 입력이 되는 키보드·스타일러스 로그는 특수 장비가 아니라 일반 학습 단말에서 표준적으로 수집 가능한 신호이며, xAPI·Caliper와 같은 학습 기록 표준을 통해 기존 LMS의 상호작용 흔적을 그대로 입력으로 받을 수 있어 별도 시스템 구축 없이 통합될 수 있다. 또한 학습자 상태 갱신을 매 상호작용이 아니라 학습 종료·반복 오류 등 의미 있는 시점에만 수행하여 다수 에이전트 운용에

따른 연산 비용과 지연을 통제하며, 진단·전략·생성 역할이 분리되어 있어 특정 에이전트만 독립적으로 교체·개선할 수 있다. 나아가 산출된 학습자 상태가 명시적 표현이므로 교사가 이를 열람·감사할 수 있어 자동 개입과 사람의 개입을 결합하는 운영이 가능하다.

IV. 결론

본 연구는 상호작용 흔적 데이터로부터 학습자 상태를 추론하고 이를 여러 에이전트가 공유하는 협업형 LLM 에이전트 학습 지원 구조를 제안하였다. 학습자의 인지적·정의적·자기조절 상태를 공통 표현으로 정의하고, 이해도 분석 에이전트가 행동의 순서와 맥락을 고려하여 이를 산출·갱신하면 학습 조정·학습 활동 생성 에이전트가 이를 구독하여 전략 결정과 학습 활동 생성에 활용하는 구조를 제시하였다. 이를 통해 학습자 상태 추론 연구와 협업형 에이전트 연구를 하나의 구조 안에서 연결하고, 정답 중심 피드백을 넘어 학습 과정 중심의 적응형 학습 지원 구조를 제안하였다. 또한 AI Hub 감성 기반 AI 튜터 데이터셋을 활용하여 제안 구조의 입력 데이터와 학습자 상태 추론의 적용 가능성을 사례 수준에서 검토하였다.

다만 본 연구는 개념적 설계와 적용 가능성 검토를 중심으로 수행되었으며, 학습자 상태 추론 모델의 구현, 상태를 구성하는 세부 지표의 타당성, 에이전트 간 의사결정 및 상태 공유 방식의 효과에 대한 실증적 검증은 향후 과제로 남는다. 또한 운영 모델 역시 설계 수준에서 제안된 것으로, 실제 LMS 연동 환경에서의 적용 가능성과 비용 대비 효과에 대한 검증이 필요하다. 향후 연구에서는 실제 교육 데이터를 활용하여 학습자 상태 생성 과정을 구체화하고, 프로토타입 구현을 통해 학습 지원 효과를 검증하고자 한다. 나아가 지속적으로 축적되는 상호작용 흔적 데이터를 활용하여 학습자 상태를 동적으로 갱신하고, 이를 기반으로 장기적 적응형 학습 지원이 가능한 구조로 확장하고자 한다.

참고문헌

[1] Jeon, J., and Lee, S., "Large Language Models in Education: A Focus on the Complementary Relationship Between Human Teachers and ChatGPT," Education and Information Technologies, Vol. 28, No. 12, pp. 15873-15892, 2023.

[2] Liu, N., Wang, Z., Baraniuk, R., and Lan, A., "Open-ended Knowledge Tracing for Computer Science Education," Proc. of EMNLP 2022, pp. 3849-3862, 2022.

[3] Park, M., Kim, S., Lee, S., Kwon, S., and Kim, K., "Empowering Personalized Learning through a Conversation-based Tutoring System with Student Modeling," CHI EA 2024, Article No. 123, pp. 1-10, 2024.

[4] Flavell, J. H., "Speculations about the Nature and Development of Metacognition," Cognitive Development, Vol. 1, No. 3, pp. 121-129, 1987.

[5] Zimmerman, B. J., "Becoming a Self-Regulated Learner: An Overview," Theory Into Practice, Vol. 41, No. 2, pp. 64-70, 2002.

[6] Corbett, A. T., and Anderson, J. R., "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge," User Modeling and User-Adapted Interaction, Vol. 4, No. 4, pp. 253-278, 1995.

[7] D'Mello, S. K., and Graesser, A. C., "Language and Discourse Are Powerful Signals of Student Emotions During Tutoring," IEEE Transactions on Learning Technologies, Vol. 5, No. 4, pp. 304-317, 2012.

[8] Winne, P. H., "Improving Measurements of Self-Regulated Learning," Educational Psychologist, Vol. 45, No. 4, pp. 267-276, 2010.

[9] Roll, I., Aleven, V., McLaren, B. M., and Koedinger, K. R., "Improving Students' Help-Seeking Skills Using Metacognitive Feedback in an Intelligent Tutoring System," Learning and Instruction, Vol. 21, No. 2, pp. 267-280, 2011.

[10] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., and Wang, C., "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," arXiv preprint arXiv:2308.08155, 2023.

[11] Bo, X., Zhang, Z., Dai, Q., Feng, X., Wang, L., Li, R., Chen, X., and Wen, J. R., "Reflective Multi-Agent Collaboration Based on Large Language Models," NeurIPS 2024, Vol. 37, 2024.

[12] Molenaar, I., and Chiu, M. M., "Dissecting Sequences of Regulation and Cognition," Metacognition and Learning, Vol. 9, No. 2, pp. 137-160, 2014.

[13] AI Hub, "서울지역 중학생 국어·수학 교과용 감성 기반 AI 튜터 데이터" [Dataset]. 한국지능정보사회진흥원 (NIA). [Online]. Available: https://www.aihub.or.kr. Accessed: May 2026.

개발 과정 회고 보조를 위한 작업 중 어려움 탐지 및 로컬 LLM 기반 시스템

정영웅[O], 이하민[**], 홍 민[*]
[O*]순천향대학교 컴퓨터소프트웨어공학과
[**]순천향대학교 소프트웨어융합학과
e-mail: wjdduddnd@sch.ac.kr[O] , haminprotoss@sch.ac.kr[*] , minhong71@gmail.com[*]

A Local LLM-Based System for In-Work Difficulty Detection to Support Development Project Retrospectives

YeongWoong Jeong[O], Hamin Lee[*], Min Hong[*]
[O*]Dept. of Computer Software Engineering , SoonChunHyang University
[**]Department of Software Convergence Engineering, Soonchunhyang University

요 약

소프트웨어 개발 과정에서 수행되는 회고는 개발 경험과 문제 해결 과정을 정리하고 이후 개발 과정 개선에 활용되는 중요한 활동이다. 그러나 실제 회고는 대부분 프로젝트 종료 이후 기억에 의존하여 작성되기 때문에, 작업 중 발생한 막힘, 작업 전환, 반복 수정 등의 세부 맥락이 누락되기 쉽다. 또한 외부 클라우드 기반 LLM 활용 시 프로젝트 코드 및 개발 기록의 외부 전송 문제가 존재한다. 본 연구에서는 개발 활동 데이터를 기반으로 작업 흐름과 어려움 맥락을 기록하고, 이를 활용하여 회고 초안 작성을 보조하는 로컬 LLM 기반 개발 회고 지원 시스템 KeepGoing을 제안한다. 제안 시스템은 주기적 활동 수집과 입력 이벤트를 기반으로 개발 활동을 기록하며, 장시간 비활동(idle), 반복 수정(revision cycle), 작업 전환(context switch) 등의 이벤트를 탐지하고 사용자 확인을 거쳐 구조화된 로그로 저장한다. 저장된 로그는 로컬 환경의 Ollama(Qwen2.5-7B)를 통해 회고 초안 생성에 활용된다. 4개의 실제 프로젝트 세션을 대상으로 평가한 결과, 자동 탐지 로그의 정밀도는 83.3~85.6% 수준이었으며, 수기 기반 회고에서 누락되기 쉬운 작업 흐름과 어려움 맥락을 보다 효과적으로 기록할 수 있음을 확인하였다.

키워드: 개발 회고, 개발 활동 로그, 작업 흐름 분석, Activity Logging, Local LLM

※" 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

I. 서론

개발 과정에서 수행한 작업 흐름과 문제 해결 과정을 되돌아보고 정리하는 회고는 이후 개발 과정의 개선과 협업 경험 공유를 위한 중요한 활동으로 활용된다. 개발자가 코드와 작업 맥락에 대한 정보를 탐색, 연결, 수집하는 과정은 프로그램 이해와 경험 정리와 밀접하게 관련된다[1]. 그러나 실제 개발 환경에서 수행되는 회고는 대부분 프로젝트 종료 이후 작성되며, 개발자는 자신의 기억과 작업 문서, 개발 로그, 커밋 기록 등을 바탕으로 개발 과정을 재구성하게 된다[2]. 이러한 기록들은 개발 결과와 변경 이력을 확인하는 데에는 유용하지만, 작업 과정에서 발생한 어려움, 반복적인 수정, 작업 전환의 배경과 같은 과정 중심의 맥락 정보를 충분히 담지 못한다는 한계가 있다[3]. 이를 보완하기 위해 개발 과정에서 생성되는 다양한 정보를 자동으로 수집·분석하고 문서화에 활용하려는 연구가 진행되고 있다. 최근에는 대규모 언어 모델(LLM)을 활용한 코드 요약, 개발 보조, 문서 생성 연구가 활발히 수행되고 있으며, 이러한 기술은 개발 회고 초안 작성에도 활용될 수 있다[4]. 다만 대부분의 LLM 기반 서비스는 클라우드 환경에서 동작하기 때문에 코드, 개발 로그, 작업 기록 등 개발 과정에서 생성된 데이터가 외부 서버로 전송될 수 있으며, 작업 정보와 개인 정보 보호 측면에서 우려가 제기되고 있다[5].

이에 본 연구에서는 개발 활동 데이터를 로컬 환경에서 수집·분석하여 작업 흐름과 어려움 맥락을 자동 기록하고, 이를 기반으로 회고 초안 작성을 지원하는 로컬 LLM 기반 개발 회고 지원 시스템 KeepGoing을 제안한다.

II. 관련 연구

2.1 개발 중 작업 중단과 맥락 복구

소프트웨어 개발 과정에서는 빈번한 작업 중단과 작업 전환이 발생하며, 복귀 시 이전 작업 맥락을 다시 이해하기 위한 비용이 발생한다[2]. 기존 연구에서는 프로그래밍 중 정지 패턴 분석[6], 동적 코드 이력을 활용한 중단 복구 지원[7] 등을 통해 개발자의 작업 재개를 지원하였다. 그러나 이러한 연구들은 복귀 시점의 생산성 향상에 초점을 두고 있으며, 작업 과정에서 발생한 어려움을 장기적으로 기록하여 회고에 활용하는 흐름까지는 충분히 다루지 못하였다.

2.2 개발 활동 로그 및 작업 흐름 분석

개발 활동을 기록 및 분석하기 위한 연구로는 GitHub 및 StackOverflow 기반 행동 분석, 사용자 활동 이력 기반 기록 연구, 개발 도구(IDE)의 로그를 활용한 작업 흐름 분석 연구 등이 제안되었다[8]. 이러한 연구들은 개발 과정에서 발생하는 작업 패턴과 활동 흐름을 분석하는 데 활용되고 있다. 그러나 기존 기록 방식은 주로 문서, 커밋, 로그 중심으로 이루어져 목적과 품질의 편차가 크며, 실제 개발 과정에서 발생하는 작업 흐름과 어려움 맥락을 충분히 반영하지 못하는 한계가 지적된다[3]. 따라서 이러한 접근은 주로 무엇이 일어났는가에 집중되며 회고 작성에 필요한 어려움, 반복 수정, 작업 전환과 같은 과정 맥락을 구조적으로 보존 및 확정하는 데에는 어려움이 있으며, 개발자가 실제로 경험한 문제 해결 과정과 의사결정 배경을 후속 문서화 과정에 직접 활용하기 힘들다는 한계점이 존재한다.

※" 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

2.3 LLM 기반 개발 지원과 프라이버시

최근에는 LLM을 활용한 코드 요약, 개발 보조, 문서 생성 연구가 활발히 진행되고 있다. 그러나 클라우드 기반 LLM 활용 시 코드 및 작업 기록, 사용자의 활동 이력 등이 외부로 전송될 수 있으며, 프라이버시 문제가 제기된다[9]. 이에 따라 개발 활동 데이터를 로컬 환경에서 수집 및 분석하고, on-device 기반으로 처리하여 개인정보 및 작업 맥락 정보를 보호하려는 접근의 중요성이 증가하고 있다[10].

III. 제안방법

3.1 시스템 개요

본 절에서는 KeepGoing의 구조와 처리 절차를 기술한다. 제안 시스템은 (1) 개발 활동 수집, (2) 어려움 이벤트 탐지 및 확정, (3) 회고 초안 생성의 세 단계로 구성된다. 시스템은 Electron 기반 데스크톱 애플리케이션과 FastAPI 기반 백엔드로 이루어지며, 활동 로그와 프로젝트 데이터를 기반으로 회고 작성을 지원한다. 전체 처리 흐름은 그림 1과 같다.

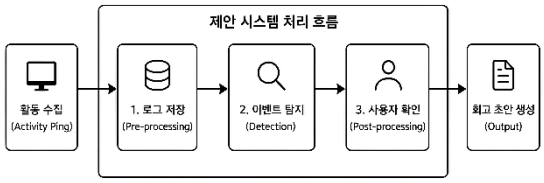


그림 1. KeepGoing 제안 시스템 처리 흐름

클라이언트는 일정 주기로 사용자의 활동 정보를 수집하여 서버로 전송하고, 서버는 누적된 신호를 분석하여 어려움 후보 이벤트를 생성한다. 생성된 후보는 사용자 확인 과정을 거쳐 확정 로그로 저장되며, 이후 프로젝트 산출물과 함께 구조화된 프롬프트로 변환되어 로컬 Ollama(Qwen2.5-7B)에 전달된다.

3.2 개발 활동 수집 방법

클라이언트는 운영체제의 프로세스 및 파일 시스템 정보를 활용하여 개발 활동 데이터를 수집한다. 전경 창 정보는 활성 애플리케이션의 프로세스명과 창 제목을 주기적으로 조회하여 기록하며, 입력 여부는 키보드 및 마우스 이벤트 발생 여부를 기반으로 판단한다. 또한 프로젝트 디렉터리를 모니터링하여 코드 및 문서 파일의 수정, 생성, 삭제 상태를 추적한다. 활동 정보는 노이즈 증가와 작업 흐름 정보 손실 사이의 균형을 고려하여 30초 주기로 서버에 전송되며, 전경 창 상태, 입력 활동 여부, 편집 파일 정보, 삭제 입력 빈도, 작업 중단 시간 등의 메타데이터로 저장된다.

3.3 어려움 이벤트 탐지 및 로그 확정

3.3.1 어려움 이벤트 탐지

서버는 누적된 활동 신호를 기반으로 어려움 후보 이벤트를 생성한다. 이벤트 유형은 idle, revision cycle, short pause, delete burst, context switch로 구성되며, 입력 부재 시간, 반복 수정 패턴, 삭제 입력 집중도, 프로그램 전환 등의 조건을 기반으로 탐지된다. VSCode, Chrome, Slack 등 개발 과정에서 자주 사용되는 프로그램 전환을 통해 작업 맥락 이탈 가능성을 함께 탐지한다. 표 1은 각 어려움 이벤트의 탐지 기준과 조건을 나타낸다.

※" 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

표 1. 어려움 이벤트 탐지 기준 및 조건

이벤트 유형	탐지 조건	쿨다운
장기 작업 중단 (idle)	누적 대기 시간 ≥ 300초 (5분)	10분
반복 수정 (revision cycle)	반복 수정 ≥ 3회, 수정 방향 변경 ≥ 3회 (30초)	5분
단기 작업 중단 (short pause)	1~5분 단기 멈춤 ≥ 2회	5분
집중 삭제 (delete burst)	삭제 ≥ 10회/창	8분
작업 전환 (context switch)	등록 파일 이탈 ≥ 90초	5분

3.3.2 사용자 확인 기반 로그 확정

생성된 이벤트는 즉시 확정되지 않고 확인 대기 상태로 저장된다. 그림 2와 같이 사용자는 자동 탐지된 이벤트를 수락, 수정 및 기각할 수 있으며, 수락된 이벤트만 최종 로그로 저장된다. 이를 통해 자동 탐지 과정에서 발생할 수 있는 잘못된 이벤트 탐지를 줄이고, 실제 회고에 활용 가능한 작업 맥락 정보를 유지하도록 하였다.

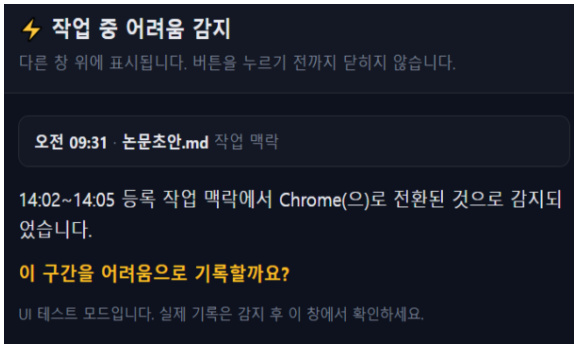


그림 2. 자동 감지 이벤트 사용자 확인 UI

3.4 어려움 로그 구성

확정된 어려움 로그는 프로젝트 산출물과 결합되어 회고 생성용 이벤트로 변환된다. 시스템은 Git 커밋, 문서, 코드 파일 등을 시간순으로 정리하고, 각 어려움 로그와 연관된 작업 흐름 정보를 함께 구성한다. 표 2는 저장되는 로그 데이터 예시를, 그림 3은 개발 활동 기반 어려움 로그 화면을 보여준다.

표 2. 어려움 이벤트 로그 예시

시각	이유	제목
01:53	delete_burst	삭제키 반복 16회 (망설임, 수정)
01:56	idle	5분 이상 작업이 멈춘 구간
02:07	context_switch	02:07~02:12 Chrome에서 활동 감지
03:08	revision_cycle	반복 수정 구간 같은 부분을 4회 고쳐 씀

그림 3. 개발 활동 기반 어려움 로그 UI

3.5 로컬 LLM 기반 회고 초안 생성

회고 초안 생성 단계에서는 확정된 어려움 로그와 Git 커밋, 코드 파일, 작업 문서 등의 프로젝트 산출물을 기반으로 구조화된 프롬프트를 생성한다. 구성된 프롬프트는 목표, 문제 상황, 해결 과정, 향후 개선 방향 등의 항목으로 이루어지며, 로컬 환경의 Ollama(Qwen2.5-7B)에 전달된다. 모델은 작업 흐름과 어려움 로그를 기반으로 회고 초안을 생성하며, 사용자는 이를 수정하여 활용할 수 있다. 로컬 LLM을 사용하여 프로젝트 원문 및 활동 데이터의 외부 전송을 방지하였다. 그림 4는 회고

※" 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

초안 생성 및 편집 화면을 나타낸다.

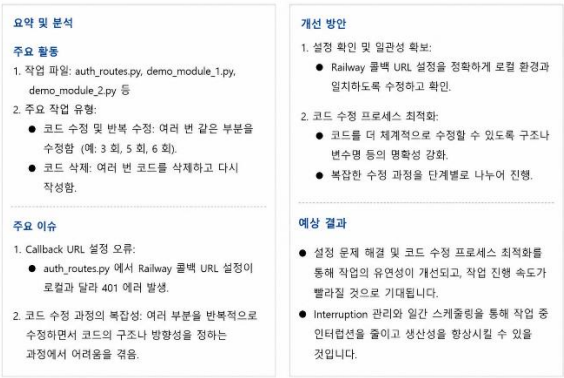


그림 4. 개발 활동 기반 회고 분석 결과 화면

Ⅳ. 평가

4.1 평가 방법

제안 시스템의 동작 확인을 위해 서로 다른 유형의 실제 작업 세션 4개를 대상으로 테스트를 수행하였다. S1은 웹 기반 정보 탐색 및 문서 작성 작업, S2는 포트폴리오 정리 작업, S3는 알고리즘 문제 해결 작업, S4는 시뮬레이션 게임 개발 작업으로 구성하였다. 사용자는 작업 후 수기 회고를 작성한 뒤 자동 생성된 어려움 로그와 비교·태깅하였다. 평가지표에서 자동 탐지 이벤트의 정밀도는 전체 탐지 로그(L) 중 실제 어려움(R)으로 확인된 로그의 비율(R/L)로 계산하였으며, RCR(Recall Coverage Ratio)은 어려움(R) 중 수기 회고에 반영된 로그(C)의 비율(C/R)로 정의하였다. 이를 기반으로 자동 탐지 로그의 정확성과 과정 중심 맥락 보존 정도를 함께 평가하였다.

4.2 시스템 평가 결과

자동 탐지 이벤트의 정밀도는 83.3~85.6% 수준으로 나타났다. 특히 프로젝트 기간과 로그 수가 증가할수록 수기 회고에서 과정 중심 맥락이 더 많이 누락되는 경향이 확인되었다. 가장 장기 세션인 S4의 경우 RCR이 19.3%로 나타났으며, 이는 개발자가 회고 작성 시 최종 결과 중심으로 기억하고 반복 수정, 막힘, 작업 전환과 같은 과정 정보는 시간 경과에 따라 누락하기 쉽기 때문으로 해석된다. 표 3은 세션별 시스템 평가 결과를 나타낸다.

표3. 세션에 따른 정밀도 결과

세션	작업량	RCR	정밀도
S1	2시간	60.0%	83.3%
S2	2일 (8시간)	45.0%	83.3%
S3	3시간	40.0%	83.3%
S4	7일	19.3%	85.6%

4.3 이벤트 유형별 분석

반복 수정 유형은 코드 수정 과정의 막힘 상황을 비교적 안정적으로 탐지하였으며, 작업 전환 유형은 외부 탐색과 작업 이탈 상황을 효과적으로 기록하였다. 반면 짧은 작업 중단과 비활동 유형은 실제 어려움과 휴식을 완전히 구분하지 못하는 사례가 존재하였다. 표4는 이벤트 유형별 RCR 결과를 나타낸다.

표4. 이벤트 유형별 RCR 결과

이벤트 유형	RCR
반복 수정 (revision_cycle)	36.4%
작업 전환 (context_switch)	32.3%
비활동 (idle)	14.3%
짧은 작업 중단 (short_pause)	11.1%

※" 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

V. 결론

본 연구에서는 30초 주기 활동 수집과 어려움 로그를 기반으로 작업 흐름을 기록하고, 이를 회고 초안 생성에 활용하는 로컬 LLM 기반 개발 회고 지원 시스템 KeepGoing을 설계 및 구현하였다. 제안 시스템은 개발 과정 중 발생하는 반복 수정, 작업 전환, 중단 등의 과정을 활동 로그 형태로 기록하고, 사용자 확인 기반 이벤트 저장 구조를 통해 회고 작성에 필요한 작업 맥락을 보조적으로 제공하였다. 실험 결과 자동 탐지 이벤트는 83.3~85.6% 수준의 정밀도를 보였으며, 장기 프로젝트 세션일수록 수기 회고에서 과정 중심 맥락 누락이 증가하는 경향을 확인하였다. 다만 본 연구는 소규모 프로젝트 세션 기반으로 수행되었으며, 이벤트 탐지가 규칙 기반으로 구성되어 다양한 개발 상황을 모두 반영하지는 못하였다. 향후에는 다수의 개발자를 대상으로 웹 서비스 개발, 게임 개발, 팀 프로젝트 등 다양한 개발 환경에서 장기간 사용자 연구를 수행하여 제안 시스템의 일반화 가능성을 검증하고 적용성을 평가할 예정이다. 또한 이벤트 간 맥락 관계를 활용한 어려움 탐지 기법을 고도화하고, 로컬 LLM 기반 회고 생성 품질 향상 방안을 추가적으로 연구하고자 한다.

참고문헌

[1] A. J. Ko, B. A. Myers, M. Coblenz, and H. H. Aung, "A Retrospective on How Developers Seek, Relate, and Collect Information About Code," IEEE Transactions on Software Engineering, Vol. 51, No. 3, pp. 848-851, Mar. 2025.

[2] C. Parnin and S. Rugaber, "Resumption Strategies for Interrupted Programming Tasks," Software Quality Journal, Vol. 19, No. 1, pp. 5-34, 2011.

[3] S. Gu, G. Rong, H. Zhang, and H. Shen, "Logging Practices in Software Engineering: A Systematic Mapping Study," IEEE Transactions on Software Engineering, Vol. 49, No. 2, pp. 902-923, 2023.

[4] N. Mireshghallah, H. Kim, X. Zhou, Y. Tsvetkov, M. Sap, R. Shokri, and Y. Choi, "Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory," International Conference on Learning Representations (ICLR), 2024.

[5] D. Arsan, C. Guo, M. R. Wellyanto, E. Ji, J. O. Talton III, and R. Kumar, "On-Device Interaction Mining," Proceedings of the ACM on Human-Computer Interaction, Vol. 9, No. 5, 2025.

[6] R. Shrestha, J. Leinonen, A. Zavgorodniaia, A. Hellas, and J. M. Edwards, "Pausing While Programming: Insights From Keystroke Analysis," Proceedings of the International Conference on Computing Education Research (ICER), 2022.

[7] V. T. T. Pham, H. Zhou, and C. Kelleher, "Interruptions and Recovery: Leveraging Dynamic Code History in Development," Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), 2023.

[8] Y. Xiong, Z. Meng, B. Shen, and W. Yin, "Mining Developer Behavior Across GitHub and StackOverflow," Proceedings of the 29th International Conference on Software Engineering and Knowledge Engineering (SEKE), pp. 363-368, 2017.

[9] Y. Wu, A. R. Chowdhury, A. P. Sheth, R. Mousavi, and C. P. de Melo, "A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly," arXiv preprint arXiv:2312.02003, 2023.

[10] B. Yan, K. Li, M. Xu, Y. Dong, Y. Zhang, Z. Ren, and X. Cheng, "On Protecting the Data Privacy of Large Language Models (LLMs): A Survey," AI Open, Vol. 6, pp. 233-249, 2025.

※" 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

Spring Boot 기반 온라인 학습 플랫폼의 설계 및 구현

왕홍우[1(제1저자)O], 김장현[2(교신저자)]
[1,2]성균관대학교 인공지능융합학과/인간AI인터랙션융합전공
e-mail: whyuoa@g.skku.edu[1], alohakim@skku.edu[2]

Design and Implementation of a Spring Boot-Based Online Learning Platform

Hongyu Wang[1], Jang-Hyun Kim[2*]
[1,2]Dept. of Applied Artificial Intelligence/Human-Artificial Intelligence Interaction, Sungkyunkwan University
*Corresponding Author

요 약

최근 인터넷 기술과 디지털 교육 환경의 발전에 따라 온라인 학습 플랫폼의 활용이 지속적으로 증가하고 있다. 온라인 학습 플랫폼은 시간과 공간의 제약 없이 학습 자원에 접근할 수 있도록 지원하며, 학습자와 교수자 간의 상호작용을 촉진하는 중요한 교육 도구로 활용되고 있다. 그러나 기존 온라인 교육 서비스의 경우 특정 교육 분야에 집중되어 있거나 학습 자원의 통합 관리 기능이 부족한 경우가 많다.

본 논문에서는 Spring Boot 프레임워크를 기반으로 한 온라인 학습 플랫폼을 설계하고 구현하였다. 제안된 시스템은 학생, 교수자, 관리자 세 가지 사용자 역할을 지원하며, 강의 관리, 수강 신청, 학습 자료 제공, 토론 기능 및 사용자 관리 기능을 제공한다. 또한 MySQL 데이터베이스와 B/S(Browser/Server) 구조를 활용하여 시스템의 확장성과 유지보수성을 향상시켰다. 구현된 시스템에 대한 기능 검증 결과, 주요 기능이 정상적으로 수행됨을 확인하였으며 온라인 교육 서비스를 위한 효율적인 플랫폼으로 활용 가능함을 확인하였다.

키워드: 온라인 학습 플랫폼, Spring Boot, 웹 시스템, 교육 플랫폼, 시스템 구현

I. 서론

인터넷과 모바일 기술의 발전은 교육 환경의 디지털 전환을 가속화하였으며, 이에 따라 온라인 학습 플랫폼에 대한 수요가 지속적으로 증가하고 있다[1]. 온라인 학습 플랫폼은 학습자가 원하는 시간과 장소에서 교육 콘텐츠를 이용할 수 있도록 지원

※이 논문은 2026년도 교육부 및 서울특별시의 재원으로 서울RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다 (2026-RISE-01-018-04).

하며, 교육 자원의 접근성을 향상시키는 중요한 역할을 수행한다[2].

현재 다양한 온라인 교육 플랫폼이 운영되고 있으나, 일부 플랫폼은 특정 교육 과정이나 자격증 시험 준비에 초점을 맞추고 있어 다양한 학습 콘텐츠를 통합적으로 제공하는 데 한계가 있다. 또한 학습자와 교수자 간의 상호작용 및 학습 관리 기능이 충분히 제공되지 않는 경우도 존재한다.

이에 본 연구에서는 온라인 강의 관리, 학습 자료 제공, 사용자 간 상호작용 및 관리자 기능을 통합적으로 제공할 수 있는 온라인 학습 플랫폼을 설계하고 구현하고자 한다. 이를 위해 Spring Boot 기반의 웹 시스템을 구축하고, 학생·교수자·관리자 중심의 기능 구조를 설계하여 실제 교육 환경에서 활용 가능한 플랫폼 구현 방안을 제시한다.

II. 관련 연구

2.1 온라인 교육 플랫폼

온라인 교육 플랫폼은 인터넷 환경에서 학습 콘텐츠를 제공하고 학습 활동을 지원하는 시스템으로, 시간과 장소의 제약 없이 교육 서비스를 이용할 수 있다는 장점을 가진다. 최근 정보통신기술의 발전과 비대면 교육 수요 증가에 따라 온라인 교육 플랫폼의 활용 범위가 지속적으로 확대되고 있다. 또한 강의 콘텐츠 제공뿐만 아니라 학습 관리, 사용자 간 상호작용 및 커뮤니티 기능을 포함하는 통합형 학습 환경으로 발전하고 있다[3].

그러나 일부 기존 플랫폼은 특정 교육 분야에 집중되어 있거나 학습자와 교수자 간의 상호작용 기능이 제한적인 경우가 있다[4]. 따라서 다양한 교육 콘텐츠를 효율적으로 관리하고 사용자 간 원활한 소통을 지원할 수 있는 온라인 학습 플랫폼의 필요성이 증가하고 있다.

2.2 관련 기술

본 연구에서는 온라인 학습 플랫폼 구현을 위해 Spring Boot 프레임워크와 MySQL 데이터베이스를 활용하였다. Spring Boot는 Java 기반의 웹 애플리케이션 개발 프레임워크로서 개발 환경 구성이 간편하고 유지보수 및 확장이 용이하다는 장점을 가진다[5]. 또한 REST 기반 서비스 구현과 데이터베이스 연동을 효율적으로 지원한다.

데이터 저장 및 관리를 위해 MySQL 데이터베이스를 사용하였으며, 시스템은 B/S(Browser/Server) 구조를 기반으로 설계하였다. 이를 통해 사용자는 별도의 프로그램 설치 없이 웹 브라우저를 이용하여 시스템에 접근할 수 있으며, 서버를 통한 통합 관리가 가능하다.

III. 제안방법

3.1 시스템 구조

제안하는 학습 플랫폼은 B/S(Browser/Server) 구조를 기반으로 설계되었다. 사용자는 웹 브라우저를 통해 시스템에 접속하며, 서버는 사용자 요청을 처리하고 데이터베이스와 연동하여 필요한 정보를 제공한다.

시스템은 크게 프레젠테이션 계층(Presentation Layer), 비즈니스 로직 계층(Business Logic Layer), 데이터 접근 계층(Data Access Layer)으로 구성된다. 사용자의 요청은 Controller를 통해 전달되며, Service 계층에서 비즈니스 로직을 처리한 후 DAO 계층을 통해 데이터베이스와 통신한다.

3.2 사용자 기능 설계

시스템은 학생, 교수자, 관리자 세 가지 사용자 역할을 지원한다.

학생 사용자는 회원가입 및 로그인, 강의 검색, 강의 신청, 강의 수강, 강의 토론 참여, 게시글 작성 및 개인정보 관리 기능을 이용할 수 있다.

교수자 사용자는 강의 개설, 강의 정보 수정, 학습 자료 등록, 수강 신청 승인 및 강의 토론 관리 기능을 수행할 수 있다.

관리자는 사용자 정보 관리, 강의 정보 관리, 게시판 관리 및 시스템 운영 기능을 담당한다.

3.3 데이터베이스 설계

데이터베이스는 사용자 정보, 강의 정보, 강의 자료, 강의 분류, 수강 신청 정보, 게시글 정보 및 댓글 정보 등을 저장하도록 설계하였다. 각 테이블은

※이 논문은 2026년도 교육부 및 서울특별시의 재원으로 서울RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다 (2026-RISE-01-018-04).

사용자 역할과 시스템 기능에 따라 유기적으로 연결되며, 데이터의 일관성과 무결성을 유지할 수 있도록 구성하였다.

3.4 시스템 구현 및 기능 검증

시스템 구현에는 Java 언어와 Spring Boot 프레임워크를 사용하였으며, MySQL 데이터베이스를 활용하여 데이터를 관리하였다. 또한 Tomcat 서버 환경에서 시스템을 운영할 수 있도록 구성하였다.

구현된 시스템에 대하여 로그인 및 회원가입 기능, 강의 신청 기능, 강의 관리 기능, 게시판 기능 및 사용자 관리 기능을 중심으로 테스트를 수행하였다. 기능 검증 결과 모든 테스트 항목에서 정상적인 동작을 확인하였으며, 사용자 요청에 대한 응답과 데이터 처리 과정이 안정적으로 수행됨을 확인하였다.

IV. 결론

본 논문에서는 Spring Boot 기반 온라인 학습 플랫폼의 설계 및 구현 방법을 제안하였다. 제안된 시스템은 학생, 교수자 및 관리자를 위한 기능을 제공하며, 강의 관리, 수강 신청, 학습 자료 제공, 토론 기능 및 사용자 관리 기능을 통합적으로 지원하도록 설계되었다. 또한 B/S 구조와 MySQL 데이터베이스를 활용하여 시스템의 안정성과 확장성을 확보하였다.

구현된 시스템에 대한 기능 검증 결과, 사용자 인증, 강의 관리, 게시판 운영 및 학습 지원 기능이 정상적으로 동작함을 확인하였다. 이를 통해 온라인 교육 환경에서 활용 가능한 학습 플랫폼을 구축할 수 있음을 확인하였다.

향후 연구에서는 사용자 경험 향상을 위한 인터페이스 개선과 다양한 교육 서비스 기능의 확장을 수행할 계획이다. 또한 개인화 학습 지원 기능 및 학습 데이터 분석 기능을 추가하여 플랫폼의 활용성과 교육 효과를 더욱 향상시킬 수 있을 것으로 기대한다.

참고문헌

[1] H. Elmunsyah, A. Nafalski, A. P. Wibawa, and F. A. Dwiyanto, "Understanding the Impact of a Learning Management System Using a Novel Modified DeLone and McLean Model," Education Sciences, Vol. 13, No. 3, Article 235, Mar. 2023.

[2] S. Xue, C. Wang, and M. Muhaimaiti, "Examining the Affordances of an Online Learning Platform: A Usefulness Theoretical Perspective," SAGE Open, Vol. 13, No. 4, Oct. 2023.

[3] B. Hmedna, A. Bakki, A. El Mezouary, and O. Baz, "Unlocking Teachers' Potential: MOOCLS, a Visualization Tool for Enhancing MOOC Teaching," Smart Learning Environments, Vol. 10, Article 58, Dec. 2023.

[4] B. H. Leem, "Impact of Interactivity on Learning Outcome in Online Learning Settings: Ordinal Logit Model," Journal of Information Technology Education: Research, Vol. 22, pp. 583-601, 2023.

[5] Li Gan, Yong Huang, and Yunli Cheng, "Research on Intelligent Learning Platform System Based on Spring Boot," Proceedings of the 2022 International Conference on Computer Science, Information Engineering and Digital Economy (CSIEDE 2022), Advances in Computer Science Research, Vol. 103, pp. 168-179, Dec. 2022.

※이 논문은 2026년도 교육부 및 서울특별시의 재원으로 서울RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다 (2026-RISE-01-018-04).

KLUE-BERT 기반 대학 행정조교 업무보조 시스템※

전유빈[O], 김수현[*]
[O]순천향대학교 컴퓨터소프트웨어공학과(학부생)
[*]순천향대학교 컴퓨터소프트웨어공학과(교수)
e-mail: ybjeon6825@gmail.com[O], kimsh@sch.ac.kr[*]

KLUE-BERT-Based Work Assistance System for University Administrative Assistants

You-bin Jeon[O], Su-Hyun Kim[*]
[O]Dept. of Computer Software Engineering, Soonchunhyang University
(Undergraduate student)
[*] Dept. of Computer Software Engineering, Soonchunhyang University
(Professor)

요 약

대학 행정 현장에서 학생 문의는 다양한 학사 행정 영역에 걸쳐 반복적으로 발생한다. 기존 학사 행정 문의는 분산된 접수 경로와 정보 출처로 인해 중복 문의 반복, 문의 이력 관리의 어려움, 문의 처리 우선순위 식별의 한계 등의 구조적 문제를 일으킬 수 있으며, 이는 행정조교의 업무 부담을 가중시키고 문의 응대의 효율성을 저하시키는 요인으로 작용할 수 있다. 본 연구에서는 이러한 문제를 완화하기 위해 KLUE-BERT 기반 AI 모듈을 활용한 대학 행정조교 업무보조 시스템을 설계하고 프로토타입을 구현하였다. 제안 시스템은 학생 문의 접수부터 답변 작성까지의 전체 흐름을 단일 플랫폼에서 관리하며, AI 모듈이 카테고리 분류, FAQ 기반 답변 후보 추천, 문의 본문 키워드 하이라이팅을 통해 실시간 문의 처리를 보조한다. 또한, 통계 기반의 카테고리별 문의 증가를 예측하고 기간별 대표 문의를 요약하여 선제적 업무 대응을 지원한다. 나아가 행정조교가 FAQ 기반 답변 후보 추천 결과를 검토한 후 최종 답변을 선택하는 Human-in-the-loop 구조를 채택함으로써 부정확한 행정 정보 제공 가능성을 완화하고, 실제 학생에게 전달되는 정보의 정확성과 신뢰성을 확보할 수 있도록 설계하였다.

키워드: AI, 행정 자동화, 자연어 처리, 문의 분류, 코사인 유사도

※ "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지원을 받아 2026 년도 AI 중심대학(2026-0-00032)의 결과로 수행되었음"

I. 서론

대학 행정 현장에서 학생 문의는 수강신청, 성적, 졸업 등 다양한 행정 영역에 걸쳐 반복적으로 발생한다. 기존 학사 행정 문의는 메신저, 이메일, 전화 등 분산된 경로로 접수되며, 답변에 필요한 정보가 공지사항, 학과 홈페이지, 학사 안내문 등 여러 출처에 분산되어 신속하게 확인하기 어렵다. 행정조교는 문의마다 내용을 읽고 유형을 분류하여 관련 정보를 탐색하고 답변을 작성하는 과정을 반복해야 한다. 이 과정이 누적되면 중복 문의의 반복, 문의 이력 관리의 어려움, 문의 처리 우선순위 식별의 한계 등의 구조적 문제를 일으킬 수 있다. 이는 행정조교의 업무 부담을 가중시키고 문의 응대의 효율성을 저하시키는 요인으로 작용할 수 있어, 문의 처리 과정을 통합적으로 보조하는 체계가 필요하다.

사전학습 언어 모델의 발전으로 텍스트 분류와 의미 유사도 측정에서 높은 성능이 보고되고 있으며[1], 이를 학사 행정 문의 응대에 적용하면 동일 키워드가 여러 카테고리에 걸쳐 등장하는 어휘적 중의성 상황에서도 예측 확률 기반 문의 분류와 관련 답변 후보 탐색을 보조할 수 있다. 이에 본 연구에서는 한국어 특화 모델인 KLUE-BERT(Korean Language Understanding Evaluation-Bidirectional Encoder Representations from Transformers)를 활용한 AI(Artificial Intelligence) 기반 자연어 처리 모듈을 중심으로 하는 행정조교 업무보조 시스템을 제안한다. 제안 시스템은 AI 모듈이 반복되는 문의 처리 과정을 보조하고 행정조교가 최종 판단을 수행하는 Human-in-the-loop(인간 개입) 구조를 통해, 단순 응대에 소요되는 시간을 단축시키고 부정확한 행정 정보가 제공될 가능성을 완화하며 행정 업무의 효율성 향상을 목표로 한다.

II. 배경 및 관련 연구

2.1 사전학습 언어 모델과 KLUE-BERT

Devlin et al. [1]이 제안한 BERT는 대규모 텍스트로 사전학습된 양방향 언어 모델로, MLM (Masked Language Modeling)과 NSP(Next Sentence Prediction) 과제를 사용하여 문맥 정보를 반영한 언어 표현을 학습한다.

한국어 학사 행정 문의 처리를 위해서는 한국어 문장의 표현과 문맥을 적절히 반영할 수 있는 사전학습 언어 모델의 활용이 필요하다. KLUE는 한국어 자연어 처리 연구를 위한 벤치마크와 사전학습 모델을 제공하며, KLUE-BERT는 한국어 대규모 텍스트로 사전학습된 BERT 계열 모델로서 문장 분류, 개체명 인식, 관계 추출 등 다양한 한국어 자연어 처리 태스크에 활용되어 왔다[2]. 이러한 특성으로 인해 한국어 학사 행정 문의를 카테고리별로 분류하는 데 적합한 사전학습 언어 모델로 볼 수 있다.

2.2 AI 기반 FAQ 답변 후보 추천

Reimers and Gurevych[3]는 BERT 계열 모델을 Siamese 구조로 활용하여 문장 임베딩을 산출하고, 두 문장 임베딩 간 코사인 유사도를 통해 의미적 유사도를 측정하는 Sentence-BERT를 제안하였다. 이러한 밀집 벡터 기반 유사도 측정은 키워드 중심 매칭보다 표현이 다르지만 의미가 유사한 문장을 포착하는 데 효과적인 것으로 보고되었다[3]. 이러한 방식은 동일한 내용을 다양한 표현으로 질문하는 학사 행정 문의 환경에서 문의와 FAQ 항목 간 의미적 유사도 측정에 적합하게 활용될 수 있다.

2.3 행정 업무 자동화 관련 연구

대학 및 행정 도메인에서는 반복 문의 처리 부담을 줄이고 정보 접근성 향상을 위해 챗봇 기반 질의응답 시스템과 자동 분류 시스템이 활용되어 왔다. Rodrigues and Fonseca[4]는 대학 행정 서비스

※ "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지원을 받아 2026 년도 AI 중심대학(2026-0-00032)의 결과로 수행되었음"

용 BERT 기반 챗봇을 제안했으나, 실제 배포 환경에서 응답 정확도는 약 70% 수준으로 보고되었다. 이는 행정 정보와 같이 정확성이 중요한 도메인에서 완전 자동 응답 방식만으로 안정적인 응대를 제공하는 데 어려울 수 있으며, 자동화에 대한 인간의 검토가 필요할 수 있음을 시사한다. Mosqueira-Rey et al. [5]은 Human-in-the-loop 접근이 AI 시스템의 자동화 과정에서 인간 검토를 결합함으로써 오류 가능성을 완화하는 데 기여할 수 있음을 논의하였다. 본 연구는 이를 바탕으로 조교가 FAQ 기반 답변 후보 추천 결과를 검토한 후 최종 답변을 선택하는 구조를 채택하여, 자동화로 인한 부정확한 정보 제공 가능성을 완화하는 것을 목표로 한다.

III. 제안 시스템

3.1 전체 시스템 아키텍처

제안 시스템은 <표 1>과 같이 프론트엔드, 백엔드 API(Application Programming Interface) 서버, 데이터베이스, AI 마이크로서비스로 구성된 각 요소가 수행하는 역할별로 분리된 구조로 설계하였다.

<표 1. 제안 시스템 구성 요소 및 배포 환경>

구분	주요 기술	배포 환경
프론트엔드	React 19, JavaScript	Render
백엔드	FastAPI, Python, JWT	Render
데이터베이스	Supabase PostgreSQL	Supabase
AI 서비스	KLUE-BERT, PyTorch, FastAPI	HuggingFace Spaces

프론트엔드는 React 19 기반의 SPA(Single Page Application)로 구현하였으며, 학생과 조교 역할에 따라 접근 가능한 기능을 분리하였다.

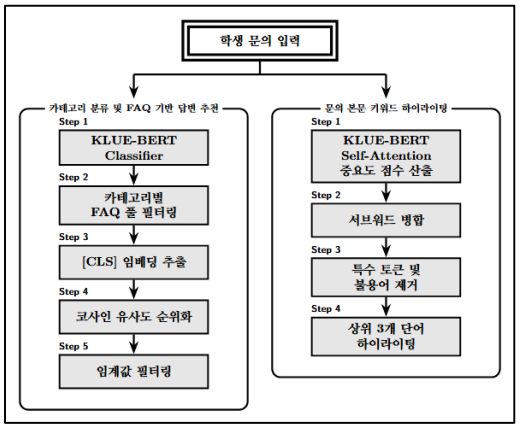
백엔드 API 서버는 Python 기반 FastAPI와 JWT(JSON Web Token) 인증 체계를 기반으로 REST(Representational State Transfer) API를 제공하며, Supabase PostgreSQL을 통해 학사 문의 및 AI 관련 데이터를 관리한다.

AI 서비스는 문의 분류, FAQ 기반 답변 후보 추천, 문의 본문 하이라이팅을 수행하는 별도의 마이크로서비스로 구축하였다. 이러한 분리는 백엔드가 관리하는 데이터베이스 접속 정보와 JWT 시크릿 등 민감한 자격 증명 영역과 AI 모델 추론 환경을 구분하기 위한 설계이다. 또한, AI 모델 교체나 재학습이 필요한 경우 백엔드를 재배포하지 않고 AI 서비스만 독립적으로 갱신할 수 있어 전체 시스템에 미치는 영향을 최소화하고, 기능 확장 및 유지보수성을 높일 수 있다.

3.2 실시간 AI 기반 문의 처리 파이프라인

실시간 AI 기반 문의 처리 파이프라인은 학생 문의를 입력으로 받아 [그림 1]과 같이 카테고리 분류 및 FAQ 답변 추천과 문의 본문 키워드 하이라이팅을 병렬로 수행하는 구조로 설계하였다.

카테고리 분류 및 FAQ 기반 답변 추천 파이프라인은 다음과 같이 5 단계로 처리된다.



[그림 1. 실시간 AI 기반 문의 처리 파이프라인]

※ "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지원을 받아 2026 년도 AI 중심대학(2026-0-00032)의 결과로 수행되었음"

Step 1. Fine-Tuning된 KLUE-BERT 분류기가 문의 본문을 10개 카테고리(수강신청, 성적, 졸업 등) 중 하나로 분류하고, 예측 카테고리와 예측 확률을 데이터베이스에 저장한다.

Step 2. 분류된 카테고리 내 FAQ를 필터링한다.

Step 3. 문의 본문과 각 FAQ 질문에 대해 KLUE-BERT [CLS] 임베딩을 추출한다.

Step 4. 코사인 유사도를 이용해 FAQ 항목을 의미적 관련성이 높은 순으로 순위화한다.

Step 5. 코사인 유사도 점수가 0.90 이상인 항목 중 최대 상위 5개를 답변 후보로 제시한다.

코사인 유사도 식은 다음과 같으며, 여기서 q와 f는 각각 문의 본문과 FAQ 질문으로부터 추출한 KLUE-BERT 기반 [CLS] 표현 벡터를 나타낸다.

$$similarity(q, f) = (q \cdot f)/(\|q\| \cdot \|f\|)$$

유사도 순위가 높더라도 실제 관련성이 낮은 항목이 후보에 포함될 수 있다. 이를 완화하기 위해 본 시스템은 코사인 유사도 점수가 임계값(0.90) 미만인 항목은 순위와 관계없이 후보에서 제외하여, 관련성이 낮은 FAQ 답변이 추천 후보로 노출되는 상황을 방지할 수 있도록 설계하였다.

문의 본문 키워드 하이라이팅 파이프라인은 문의 본문 내 주요 표현을 시각적으로 표시하기 위해 수행되며, 다음과 같이 4 단계로 처리된다.

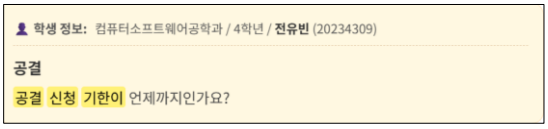
Step 1. KLUE-BERT 마지막 레이어의 Self-Attention 가중치에서 [CLS] 토큰이 각 토큰에 부여하는 가중치를 멀티헤드 간 평균을 내어 토큰별 중요도 점수를 산출한다.

Step 2. 서브워드를 원문 단어 단위로 병합한다.

Step 3. 특수 토큰 및 불용어를 제거한다.

Step 4. 중요도 상위 3개 단어를 키워드로 선정하여 조교 화면에서 하이라이팅으로 표시한다.

Self-Attention 기반 문의 본문 키워드 하이라이팅은 별도의 키워드 라벨링 데이터 없이 모델 내부



[그림 2. 문의 본문 하이라이팅 화면]

[그림 3. AI 추천 답변 후보 화면]

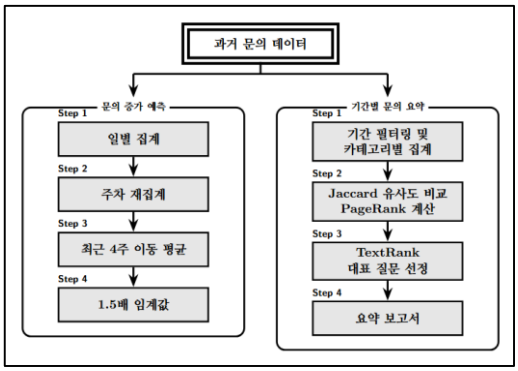
에서 산출되는 가중치를 시각적 보조 지표로 활용될 수 있다는 장점이 있다. 그러나 Self-Attention 가중치는 언어학적 중요도나 인간의 판단과 일치하지 않을 수 있음이 선행 연구에서 보고되었으며[6], 일부 문의에서는 상대적으로 중요하지 않은 토큰이 높은 가중치를 받을 가능성이 있다. 이러한 한계를 고려하여, 본 시스템은 결과를 자동 판단 근거로 사용하지 않고 조교의 문의 파악을 돕는 보조적 참고 도구로서의 역할을 수행하도록 설계하였다.

두 흐름의 처리 결과는 [그림 2], [그림 3]과 같이 제공되며, 행정조교는 문의 본문 하이라이팅 결과와 FAQ 기반 답변 후보를 참고하여 최종 답변을 작성하거나 적절한 답변 후보를 선택할 수 있다.

3.3 통계 기반 문의 분석 처리 파이프라인

통계 기반 문의 분석 처리 파이프라인은 과거 문의 데이터를 입력으로 받아 [그림 4]와 같이 문의 증가 예측과 기간별 문의 요약을 병렬로 수행하는 구조로 설계하였다. 본 분석 모듈을 통해 축적된 통계 데이터는 학과 행정 프로세스의 개선을 위한 기초 자료로 활용될 수 있다.

※ "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지원을 받아 2026 년도 AI 중심대학(2026-0-00032)의 결과로 수행되었음"



[그림 4. 통계 기반 문의 분석 처리 파이프라인]

[그림 5. 통계 기반 문의 증가 예측 화면]

[그림 6. 통계 기반 기간별 문의 요약 화면]

통계 기반 문의 증가 예측 파이프라인은 다음과 같이 4 단계로 처리된다.

Step 1. 과거 문의 데이터를 일별로 집계한다.

Step 2. 주차 단위로 재집계하여 카테고리별 주간 문의량을 시계열 형태로 구성한다.

Step 3. 최근 4주 이동 평균을 이용하여 다음 주 예상 문의량을 산출한다.

Step 4. 전체 기간 주간 평균 대비 1.5배 이상 증가가 예상되는 카테고리를 증가율 내림차순으로 정렬한다.

통계 기반 문의 증가 예측 결과는 [그림 5]와 같이 특정 유형의 문의가 집중되는 시기를 사전에 파악하고 FAQ 항목 보완이나 공지사항 재안내 등 선제적 대응을 수행할 수 있도록 설계하였다.

통계 기반 기간별 문의 요약 파이프라인은 다음과 같이 4 단계로 처리된다.

Step 1. 행정조교가 지정한 시작 기간과 종료 기간 내 문의를 카테고리별로 집계한다.

Step 2. 카테고리 내 문의 제목을 문자 바이그램 기반 Jaccard 유사도로 비교하여 문장 간 유사도 행렬을 구성하고, 감쇠 계수(Damping Factor) 0.85를 적용한 PageRank 계산을 반복하여 문장 중요도 점수를 산출한다.

Step 3. TextRank 알고리즘[7]에 따라 중요도 점수가 가장 높은 문장을 각 카테고리의 대표 질문으로 선정한다.

Step 4. 문의 건수 기준 내림차순으로 정렬된 요약 보고서 형태로 조교 화면에 제공된다.

통계 기반 기간별 문의 요약 결과는 [그림 6]과 같이 설정한 기간에 따른 카테고리별 집계와 대표 질문을 제공함으로써 FAQ 항목 보완이나 공지사항 작성에 활용할 수 있도록 지원한다.

3.4 Human-in-the-loop

제안 시스템은 AI가 문의 분류, FAQ 기반 답변 후보 추천, 키워드 하이라이팅을 수행하되, 행정조교가 추천 결과를 검토한 후 최종 판단을 결정하는 Human-in-the-loop 구조를 채택하였다. 이 구조는 완전 자동 응답 방식에서 발생할 수 있는 잘못된 정보가 그대로 학생에게 전달될

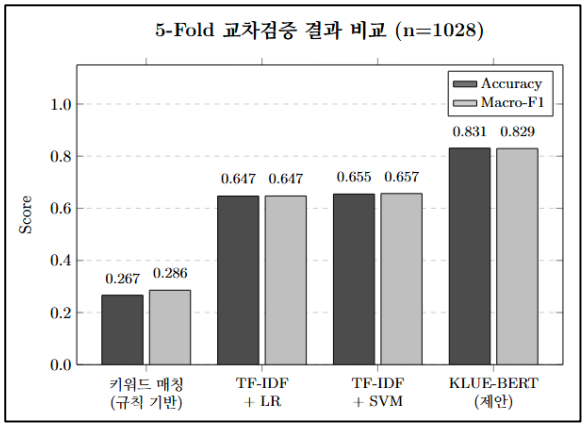
※ "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지원을 받아 2026 년도 AI 중심대학(2026-0-00032)의 결과로 수행되었음"

가능성을 완화할 수 있도록 설계하였다. 이를 통해 AI가 행정조교의 판단을 대체하는 것이 아니라, 반복적이고 시간이 소요되는 분류, 검색 과정을 보조하는 역할을 한다.

IV. 분류 성능 평가

제안 모델인 Fine-Tuning된 KLUE-BERT 분류기의 성능을 검증하기 위해 1,028건의 학사 행정 문의 데이터를 대상으로 5-Fold 교차검증을 수행하였다. 비교 베이스라인으로는 키워드 매칭, TF-IDF(Term Frequency-Inverse Document Frequency) + LR(Logistic Regression), TF-IDF + SVM(Support Vector Machine)을 설정하였다.

교차검증 결과는 [그림 7]과 같이 나타났다. 제안 모델은 Accuracy(전체 예측 대비 정답 비율) 0.831, Macro-F1(카테고리별 F1 점수의 단순 평균) 0.829를 기록하여 가장 근접한 베이스라인인 TF-IDF + SVM 대비 약 17.6%p의 정확도 향상을 달성하였다. 또한, 각 Fold마다 나온 성능 수치들의 표준편차는 ±0.0082로, 키워드 매칭(±0.0249), TF-IDF + LR(±0.0457), TF-IDF + SVM(±0.0424) 대비 현저히 낮은 수치를 기록하였다. 이는 데이터 분할 방식에 관계없이 분류 성능이 일관되게 유지됨을 의미하며, 제안 모델의 안정성을 뒷받침한다.



[그림 7. 5-Fold 교차검증 결과 비교 그래프]

V. 결론

본 연구에서는 KLUE-BERT 기반 AI 모듈을 활용하여 대학 행정조교의 반복적인 문의 응대 업무를 보조하는 시스템을 설계 및 구현하였다. 기존의 분산된 경로를 통한 문의 방식이 내포하는 구조적 한계를 완화하기 위해, 문의 통합 관리, AI를 통한 카테고리 분류, FAQ 기반 답변 후보 추천, 문의 본문 하이라이팅, 통계 기반 문의 증가 예측 및 기간별 문의 요약 기능을 제공한다.

향후에는 보다 다양한 유형의 학사 행정 문의 데이터를 확보하여 일반화 성능을 검증하고, 시스템 주요 모듈에 대한 정량적 평가를 수행할 계획이다.

참고문헌

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019, pp. 4171–4186.

[2] Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Cho, K. (2021). KLUE: Korean language understanding evaluation. In Proceedings of NeurIPS 2021 Datasets and Benchmarks Track.

[3] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In Proceedings of EMNLP-IJCNLP 2019, pp. 3982–3992.

[4] Rodrigues, F., & Fonseca, J. (2026). ChatBot for student service based on RASA framework. Knowledge and Information Systems, 68, Article 39. https://doi.org/10.1007/s10115-025-02655-4

[5] Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review, 56(4), 3005–3054.

[6] Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In Proceedings of NAACL-HLT 2019, pp. 3543–3556.

[7] Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. EMNLP, pp. 404–411.

※ "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지원을 받아 2026 년도 AI 중심대학(2026-0-00032)의 결과로 수행되었음"

올인원 IT 커리어 관리 플랫폼의 설계 및 구현

저자*, 김성재(*)신미리O (*), 허준영(*),김유림(**)
*한신대학교 소프트웨어학과

e-mail: jea9405@hs.ac.kr
sml041214@hs.ac.kr
eric1889@hs.ac.kr

**한신대학교 컴퓨터공학과
e-mail: kk01407@hs.ac.kr

Design and Implementation of an All-in-One IT Career Management Platform

Author*, Kim Seong-jae (*)Shin Mi-RiO (*), Heo jun-young (*),Kim Yu-rim (**)
*Software Department , HanShin University
**Department of Computer Science, HanShin University

요 약

현재 IT 분야의 취업 준비 과정은 교육 콘텐츠, 자격증 정보, 프로젝트 준비, 채용 공고 등이 여러 플랫폼에 파편화되어 있어 학습자의 비효율성을 초래하고 체계적인 커리어 계획 수립을 어렵게 만든다. 본 논문에서는 이러한 문제를 해결하기 위해 대학생 및 취업준비생을 대상으로 "학습 자격증 포트폴리오 면접"에 이르는 역량 개발 전 과정을 유기적으로 연계한 올인원 IT 커리어 관리 플랫폼을 설계하고 구현한다. 본 플랫폼은 사용자의 역량을 분석하여 최적의 자격증을 도출하는 '개인화 자격증 추천 시스템'과 '학습 로드맵 자동 생성 기능'을 핵심으로 한다. 나아가 CBT 모의고사 환경과 AI 오답 분석 리포트를 제공하며, 역량 진단부터 실전 취업 준비까지의 단절을 해소하여 자기주도적 학습 효율을 극대화하였다.

키워드: 올인원 플랫폼, 자격증 추천 시스템, AI 학습 관리, CBT 모의고사, IT 커리어 패스

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

I. 서론

현재 IT 분야의 취업 준비 과정은 교육 콘텐츠, 자격증 정보, 프로젝트 준비, 채용 공고 등이 여러 플랫폼에 파편화되어 있어 학습자의 비효율성을 초래하고 체계적인 커리어 계획 수립을 어렵게 만든다. 특히 핵심 타깃인 IT 관련 전공 대학생 및 취업준비생들은 목표 직무 설정의 모호성, 체계적인 학습 로드맵의 부재, 분산된 시험 일정 관리의 어려움 등을 겪고 있다. 또한, 취득한 역량을 포트폴리오와 면접으로 연결하는 과정에서 심각한 단절이 발생한다.

본 논문에서는 이러한 문제를 해결하기 위해 대학생 및 취업준비생을 대상으로 "학습 자격증 포트폴리오 면접"에 이르는 역량 개발 전 과정을 유기적으로 연계한 올인원 IT 커리어 관리 플랫폼을 설계하고 구현한다. 본 시스템은 맞춤형 학습 가이드와 취업 패스를 결합한 통합적 솔루션을 제시함으로써 취업 준비생들의 시간적·경제적 비용을 절감하는 데 기여하고자 한다.

II. 관련 연구

2.1 국내 동향

기존의 에듀테크 플랫폼들은 단편적인 동영상 강의 제공이나 독립적인 자격증 기출문제 문제은행 서비스에 국한되어 있다. 이로 인해 학습자는 자격증 취득 이후 이를 어떻게 포트폴리오에 녹여내고 실전 면접에 적용해야 하는지 자기주도적으로 해결해야 하는 부담을 안고 있다. 채용 플랫폼 역시 단순 채용 공고 필터링 기능만을 제공하여 역량 개발 단계와의 유기적인 피드백 루프가 부족한 실정이다.

III. 제안방법

본 논문에서 제안하는 올인원 IT 커리어 플랫폼은 React 기반의 프론트엔드와 Node.js(Express) 백엔드 아키텍처로 구성되며, RESTful API를 통해 통신한다. 데이터베이스는 사용자 정보 및 문제은행 관리를 위해 관계형 DB(MySQL/PostgreSQL)를 채택하였다. 핵심 기능의 메커니즘은 다음과 같다.)

개인화 자격증 추천 및 로드맵 자동 생성

사용자의 전공, 기술 스택, 관심 직무 및 가용 시간을 분석하여 최적의 자격증을 도출하는 '개인화 자격증 추천 엔진'을 구현하였다. 자격증 DB 내 난이도, 예상 준비기간, 취업 연계성 지표를 종합하여 우선순위를 산출한다. 선택된 자격증을 바탕으로 스케줄링 알고리즘을 통해 주차별 학습 로드맵이 자동으로 생성되며, 외부 강의 플랫폼(인프런 등)의 Open API 와 연동되어 최적의 학습 자료를 매칭한다.

CBT 모의고사 및 AI 동적 피드백 루프

실제 시험과 유사한 컴퓨터 기반 시험(CBT) 환경을 구현하여 실전형 UI/UX 를 제공한다. 시험 종료 후에는 AI 를 활용하여 도메인별 취약점 및 오답 유형(개념 착오, 계산 실수 등)을 정밀 분석하는 리포트를 생성한다. 학습자의 평가 데이터가 누적됨에 따라 취약 영역을 보완할 수 있도록 로드맵이 동적으로 재조정되는 피드백 시스템을 구축하였다.

커리어 시뮬레이션 및 확장 기능

사용자의 현재 역량과 목표 직무의 채용공고를 대조하여 스킬 갭을 시뮬레이션한다. 부족한 스킬을 보완할 수 있는 미니 프로젝트 가이드를 제공하며, 직무별 포트폴리오 템플릿과 AI 기반 기술/인성 모의면접 기능을 통해 최종 취업 준비 단계까지의 단절을 해소한다.

기능 분류	핵심 구현 기술 및 데이터 소스
추천/로드맵	스케줄링 알고리즘, 외부 교육 플랫폼 Open API 연동
CBT/AI 분석	프론트엔드 타이머·마크 UI, AI 기반 오답 도메인 분석 리포트
정보/알림	Q-Net 스레이핑 데이터, 웹푸시 및 이메일 알림 인프라
시뮬레이션	채용공고(JD) 스킬 갭 분석 엔진, AI 모의면접 LLM 연동

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

IV. 결론

본 연구를 통해 개발된 플랫폼은 단편적인 정보 제공에 그치던 기존 에듀테크 서비스의 한계를 극복하고, 맞춤형 학습 가이드와 취업 패스를 결합한 통합적 솔루션을 제시하였다. 역량 진단부터 자격증 취득, AI 기반 오답 피드백, 포트폴리오 및 면접 준비에 이르는 전 과정을 원스톱으로 제공함으로써 사용자 중심의 효율적인 커리어 관리가 가능함을 확인하였다. 향후 본 시스템은 IT 인재 양성 및 취업 준비생들의 시간적·경제적 비용을 절감하고, 시장 수요에 부합하는 역량을 효율적으로 구비하는 데 기여할 것으로 기대된다.

참고문헌

[1] 한국산업인력공단, "국가기술자격 통계연보: 정보처리기사 응시 및 합격 현황," 큐넷(Q-Net) 공식 통계자료, 2025.

[2] 이민우, 김수현, "에듀테크 플랫폼 기반 개인화 학습 로드맵 생성 및 스케줄링 알고리즘 연구," 데이터사이언스학회 논문집, Vol. 3, No. 2, pp. 45-58, 2025.

[3] Open AI, "LLM(대형 언어 모델)을 활용한 컴퓨터 기반 시험(CBT) 오답 유형 분석 및 정밀 피드백 리포트 생성 방안," 인공지능학회지, Vol. 12, No. 4, pp. 112-125, 2024.

[4] 주식회사 인프랩, "인프런(Inflearn) Open API 기반 교육 콘텐츠 메타데이터 수집 및 매칭 기술 사양서," 2024.

[5] 원티드랩, 사람인 HR 연구소, "IT 직무 채용공고(JD) 요구 역량 분석을 통한 구직자 스킬 갭(Skill Gap) 시뮬레이션 지표 산출 모델," 채용데이터 분석 리포트, 2025..

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과입니다.

DTW 알고리즘을 이용한 사용자 맞춤형 일본어 억양 교정 시스템 구현

전이현*, 김예나**, 성낙준*
*한신대학교 AI·SW대학
**한신대학교 일본학전공
e-mail: {komosjs44, njsung}@hs.ac.kr, yena315@hs.ac.kr

Implementation of Personalized Japanese Pitch-Accent Correction System using DTW Algorithm

Yi-hyun Jeon*, Ye-na Kim** , Nak-Jun Sung *
* School of Computing and Artificial Intelligence, Hanshin University
** Division of Japanese Studies, Hanshin University

요 약

최근 한·일 간 문화 교류의 확대와 함께 일본어 학습에 대한 수요가 증가하고 있으며, 특히 일본어 억양은 단어의 의미를 구별하는 핵심 요소로서 학습의 중요성이 강조되고 있다. 그러나 기존의 일본어 발음 학습 방식은 학습자가 자신의 억양을 객관적으로 인식하고 교정하는 데 한계를 가진다. 본 논문에서는 일본어 학습자의 발음을 표준 발음과 비교·분석하여 억양 차이를 시각적으로 제공하는 AI 기반 억양 교정 시스템 J-Accent Mirror를 제안한다. 제안한 시스템은 음성 신호로부터 pitch 정보를 추출한 뒤, 발화 길이와 속도 차이를 보정하기 위해 Dynamic Time Warping(DTW) 알고리즘을 적용하여 억양 유사도를 정량적으로 산출한다. 또한 표준 발음과 사용자 발음을 중첩 그래프로 시각화함으로써 학습자가 억양 차이를 직관적으로 인식할 수 있도록 설계하였다. 본 시스템은 일본어 억양 학습의 객관성과 접근성을 향상시키는 도구로서, AI 기반 외국어 발음 교육의 활용 가능성을 제시한다.

키워드: 억양 교정, DTW 알고리즘, Pitch Accent, 음성 처리, 일본어 교육

I. 서 론

한국과 일본은 오랜 역사적 관계를 바탕으로 정치·경제·사회 전반에 걸쳐 다양한 교류를 지속해 왔으며, 최근에는 문화 콘텐츠와 교육 분야를 중심으로 교류가 더욱 활발히 이루어지고 있다 [1]. 특히 2025년 한·일 국교 정상화 60주년을 기점으로, 양국은 공동 문화 행사와 콘텐츠 축제를 개최하며 상호 문화 이해를 증진시키고 있다 [2]. 이러한 흐름은 단순한 콘텐츠 소비를 넘어, 언어 이해를 기반으로 한 실질적 의사소통 능력의 중요성을 더욱 부

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

각시키고 있다.

일본어는 음소 단위의 발음 정확성뿐만 아니라, 단어 및 문장 단위에서의 억양(pitch accent) 패턴이 의미 전달에 중요한 역할을 하는 언어이다. 동일한 음소 조합이라도 억양 차이에 따라 의미가 달라질 수 있기 때문에, 억양 학습은 일본어 구사 능력의 핵심 요소로 간주된다 [3]. 그러나 기존의 일본어 발음 학습 방식은 교사의 발음을 청각적으로 모방하는 방식에 의존하여, 학습자가 자신의 발음이 표준 발음과 어떻게 다른지를 객관적으로 인식하기 어렵다는 한계를 지닌다. 또한 교사의 피드백이 추상적·주관적으로 제시되는 경우가 많아, 억양 차이를 구체적으로 이해하는 데 어려움이 있다. 이러한 한계를 보완하기 위해, 본 논문에서는 일본어 학습자의 억양 교정을 지원하는 AI 기반 웹 시스템 J-Accent Mirror를 제안한다. 제안 시스템은 표준 음성과 학습자 음성을 입력으로 받아 pitch 정보를 추출하고, Dynamic Time Warping(DTW) 기반 비교 방식을 통해 억양 유사도를 정량적으로 산출하며, 결과를 점수와 시각적 그래프 형태로 제공한다.

본 연구의 기여는 다음과 같다. 첫째, 억양을 시계열 데이터로 정량화하고 시각적으로 비교하는 학습 지원 방식을 제안하였다. 둘째, DTW 기반 억양 정렬 기법을 적용하여 발화 속도 차이에 강인한 억양 유사도 평가 방법을 제시하였다. 셋째, 억양 비교 결과를 0~100점으로 환산하여 학습자가 개선 과정을 직관적으로 확인할 수 있는 피드백 구조를 구현하였다.

II. 관련 연구

2.1 일본어 억양 및 발음 학습 연구

일본어 억양 연구는 음운론적 관점에서 고저 악센트 체계를 중심으로 진행되어 왔다. 도쿄 방언의 억양 체계에서 단어의 의미는 음절 간 고저 패턴에 의해 결정되며, 학습자가 이를 정확히 습득하지 못하면 의사소통 오류로 이어질 수 있다. 기존 발음 교육 연구는 주로 음소 단위의 정확성이나 반복 청취 중심 학습에 초점을 맞춰 왔으나, 억양을 시계열 데이터로 정량화하여 학습에 활용한 연구는 상대적으로 부족한 실정이다.

2.2 DTW 알고리즘 관련 연구

DTW는 길이가 다른 두 시계열 데이터 간의 최적 정렬 경로를 탐색하는 알고리즘으로, 음성 인식 및 신호 분석 분야에서 매칭 정확도가 우수한 기법으로 평가받고 있다 [4]. DTW는 발화 속도나 길이가 상이한 음성 신호 간의 대응 관계를 효과적으로 정렬할 수 있어, 외국어 학습자의 비선형적 시간 왜곡 특성을 보정하는 데 적합하다. 기존 연구에서는 DTW를 이용한 시계열 서브시퀀스 매칭과 GPU 가속화 방안을 제시하였으며 [4], 본 연구에서는 이러한 DTW 특성을 일본어 억양 비교에 적용하였다.

2.3 AI 기반 발음 교육 시스템

최근 AI 기술의 발전으로 발음 교정 및 언어 학습 지원 시스템에 대한 연구가 활발히 진행되고 있다. 자동 음성 인식 기반 발음 평가 시스템이 다수 제안되었으나, 이들은 주로 음소 단위의 정확성 평가에 집중하여 억양 패턴의 시각화 및 정량적 비교 기능은 상대적으로 미흡하다. 본 연구는 pitch 추출과 DTW 정렬을 결합하여 억양 유사도를 직관적으로 제시함으로써 기존 시스템과 차별성을 갖는다.

III. 제안 방법

3.1 시스템 설계 개요

J-Accent Mirror는 일본어 표준 발음과 학습자 발음을 입력받아 억양을 비교·분석하고, 결과를 시각적·정량적 형태로 제공하는 웹 기반 언어 학습 지원 시스템이다. Python으로 구현되었으며, Streamlit 프레임워크를 활용하여 별도의 설치 없이 웹 브라우저에서 실행 가능하도록 설계하였다.

시스템은 ① 표준 음원 입력부, ② 사용자 음성 입력부(마이크 녹음), ③ pitch 추출 및 억양 비교 분석 모듈, ④ 점수·그래프 제공 결과 출력 모듈의 네 가지 구성 요소로 이루어진다. 사용자가 표준 음원(.wav 파일)을 업로드하고 발음을 녹음하면, 시스템은 두 음성의 pitch 정보를 추출하여 DTW 기반 유사도를 계산하고 결과를 제공한다.

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

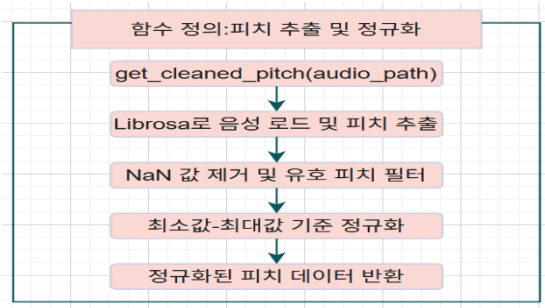


그림 1. J-Accent Mirror 시스템 구성도

3.2 음성 데이터 처리 및 Pitch 추출

입력된 음성 데이터는 분석 전에 전처리 과정을 거친다. librosa 라이브러리를 활용하여 음성 파일을 로드한 후, pyin 함수를 이용해 음성 신호에서 기본 주파수(f0), 즉 pitch 정보를 추출한다. 추출된 pitch 데이터에서 NaN 값과 0 이하의 값을 제거하여 유효한 데이터만 선별하고, 화자 간 음역 차이를 최소화하기 위해 다음 수식과 같이 각 음성의 pitch 범위를 0~1 사이로 정규화한다.

$$d_i = |p_i - q_i| \tag{1}$$

3.3 DTW 기반 억양 유사도 산출

전처리된 pitch 데이터는 발화 길이와 속도 차이로 인해 서로 다른 길이의 시계열로 나타난다. 외국어 학습자는 특정 음절을 불필요하게 길게 발음하는 비선형적 시간 왜곡 특성을 보이므로, 프레임 단위 직접 비교는 실제 억양 패턴이 유사하더라도 큰 거리 오차를 초래한다. 이를 보완하기 위해 DTW 알고리즘을 적용하여 두 시계열 간 최적 정렬 경로를 탐색한다. DTW를 통해 정렬된 표준 발음의 pitch 시퀀스를 P={p1,p2,...,pN}, 사용자 발음의 pitch 시퀀스를 Q={q1,q2,...,qN}이라 정의한다. 각 프레임에서의 거리 di 및 평균 거리 D_{avg}는 다음과 같이 정의된다.

$$D_{avg} = \left(\frac{1}{N}\right) \times \sum_{i=1}^{N} d_i \tag{2}$$

계산된 D_{avg}는 선형 변환을 통해 0~100점의 유사도 점수 S로 환산된다.

$$S = \max\left(0, 100 \times \left(\frac{1-D_{avg}}{D_{max}}\right)\right) \tag{3}$$

D_{max}는 예비 실험을 통해 설정한 최대 허용 거리 값으로, 본 연구에서는 정규화된 pitch 시퀀스(0~1 범위) 기준 D_{max}=0.5를 적용하였다. 이 값은 두 음성 간 평균 거리가 0.5 이상이면 억양 유사도가 매우 낮다고 판단하는 기준에 해당한다. 점수가 기준 이상이면 긍정적 피드백을, 기준 미달 시 재학습 유도 메시지를 출력하여 학습 동기를 강화하도록 구성하였다.

IV. 실험 결과

4.1 실험 설계

본 연구에서는 일본어의 대표적 최소 대립 쌍(Minimal Pairs)인 '雨(비, 두고형/HL)'와 '飴(사탕, 평판형/LH)'를 활용하여 비교 실험을 진행하였다. 두 단어는 음소 구성이 동일하나 억양 패턴에 의해 의미가 결정되므로, 시스템이 미세한 기본 주파수 변화를 정밀하게 판별하는지 확인하는 핵심 지표가 된다.

4.2 Pitch 데이터 분석 결과

실험 결과, 추출된 주파수 궤적은 각 단어의 음운론적 특징을 명확히 반영하였다. 雨(비): 첫 음절 '아(あ)' 구간에서 주파수가 급격히 상승하여 높은 피크를 형성한 후, 두 번째 음절 '메(め)'에서 급강하하는 두고형(High-Low) 패턴이 뚜렷이 관찰되었다. 飴(사탕): 첫 음절에서 낮은 주파수를 유지하다가 두 번째 음절에서 상승·유지되는 평판형(Low-High) 패턴을 보였다. 이는 본 시스템의 pitch 추출 모듈이 일본어 특유의 고저 차이를 왜곡 없이 데이터화함을 시사한다.

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

그림 2. 雨(비) 분석 결과

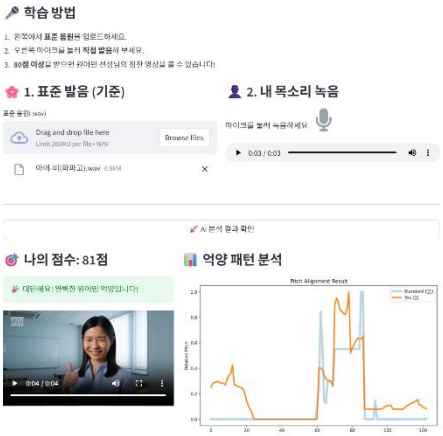


그림 3. 飴(사탕) 분석 결과

4.3 DTW 정렬 효과 및 피드백 실효성

학습자가 표준 발음을 모방하여 발화했을 때, 시스템은 두 사례 모두에서 80점 이상의 유사도 점수를 산출하였다. 다만 본 실험은 제한된 단어 집합과 단일 화자를 대상으로 한 예비 실험 수준임을 유의해야 한다. 억양 패턴 분석 그래프에서 표준 음성(청색)과 학습자 음성(주황색)이 시간 축 정렬을 통해 밀접하게 중첩됨을 확인하였으며, 특히 '雨'의 초반 피크 구간과 '飴'의 후반 상승 구간이 표준 데이터의 임계 지점과 정확히 일치하였다.

또한 발화 속도나 타이밍이 표준 음원과 상이하더라도 DTW 알고리즘이 비선형적 시간 왜곡을 효과적으로 보정함으로써, 억양 곡선의 형태적 유사성을 정밀하게 계산할 수 있었다. 시각적 환류(Visual Feedback)를 통해 학습자는 자신의 발음이 표준 패턴을 얼마나 따르는지 실시간으로 식별할 수 있으며, 점수 기반 긍정적 강화로 학습 지속성을 확보한다.

V. 결 론

본 연구에서는 일본어 학습자의 억양 교정을 지원하는 AI 기반 웹 시스템 J-Accent Mirror를 설계하고 구현하였다. 제안 시스템은 음성 신호에서 pitch 정보를 추출하고, DTW 알고리즘을 통해 발화 속도 차이를 보정함으로써 억양 유사도를 정량적 점수와 시각적 그래프 형태로 제공한다. 실험을 통해 일본어 최소 대립 쌍에서 고저 패턴 차이를 명확히 판별함을 확인하였다. 본 시스템의 핵심 의의는 학습자가 청각적 모방 중심 학습에서 벗어나, 정량적 수치와 시각적 자료를 통해 억양을 객관적으로 확인할 수 있는 학습 환경을 구축하였다는 점이다. 이는 기존 발음 교육의 주관적 판단 한계를 보완하고, 자기주도적 학습 효율성 향상에 기여한다.

다만 본 연구는 제한된 단어 집합을 중심으로 실험을 수행하였으며, 정량적 실험 규모에 한계가 있다. 향후 연구에서는 다수의 단어 및 다양한 학습자 데이터를 활용한 통계적 검증을 수행하고, 분석 범위를 문장 단위로 확장할 계획이다. 또한 강세·리듬 요소를 통합 분석하는 모델로 발전시키고, 교육 현장 적용 실험을 통해 AI 기반 외국어 발음 교육의 실용적 활용 가능성을 더욱 확대하고자 한다.

참고문헌

[1] "역사는 역사, 문화는 문화", 중앙일보, 2025년 12월 29일, https://www.joongang.co.kr/article/25386366

[2] "日 다카이치, 한일수교 60주년 관계 심화 기대", 공감언론 뉴시스, 2025년 12월 30일, https://www.newsis.com/view/NISX20251218_0003447078

[3] "김 여사 한일청년 언어 교류 양국관계 발전에 중요",

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

KTV 국민방송, 2025년 12월 30일, https://www.ktv.go.kr/content/view?content_id=735639
[4] 노효영, DTW 알고리즘을 이용한 시계열 데이터의 서브시퀀스 매칭과 GPU를 활용한 가속화 및 최적화 방안, 서울대학교 공과대학 컴퓨터공학부 석사학위논문, 2024년 2월.

※ 본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

연합학습 환경의 LLM 파인튜닝을 위한 누적 토큰 학습량 정합화 기반 학습 방법 연구

정진용, 송인서, 이강윤*
가천대학교 컴퓨터공학과
e-mail: wlsdyd5373@gachon.ac.kr, z8086486@gachon.ac.kr, keylee@gachon.ac.kr*

Research on Cumulative Token-Usage Alignment for LLM Fine-Tuning in Federated Learning

JinYong Jeong, InSeo Song, KangYoon Lee*
Dept. of Computer Engineering, Gachon University

요 약

연합학습 환경은 의료나 금융과 같이 보안에 민감한 분야에서 데이터를 외부로 전송하지 않고도 LLM을 학습할 수 있기 때문에 중요한 대안으로 떠오르고 있다. 그러나 기존 연합학습 프로토콜은 대개 일반적인 딥러닝 모델의 학습을 전제로 설계되어 있으며, round마다 고정된 local epoch을 수행한 뒤 이를 집계하는 방식으로 동작한다. 이를 LLM 파인튜닝에 그대로 적용하게 되면 각 client 업데이트가 local 데이터 분포에 과도하게 적응한 상태로 집계될 수 있으며, round 수의 증가는 집계 횟수 뿐만 아니라 전체 토큰 사용량의 증가로 이어진다. 그 결과 일반 성능이 저하되거나 도메인 간 성능 간섭이 발생할 수 있다. 본 논문에서는 이 문제를 해결하기 위해 연합학습 기반 LLM 파인튜닝을 위한 누적 토큰 학습량 정합화 기반 학습 방법을 제안한다. 이는 global epoch을 우선 정의하여 전체 사용할 토큰량을 고정하고 round별 sample index 구간으로 분할하여 중앙집중식 파인튜닝과 동일한 토큰 사용량 조건을 제공하는 것이다. 이를 검증하기 위해 medical과 finance dataset을 통해 단일 도메인과 복합 도메인 llm 파인튜닝을 제안한 protocol을 적용한 연합학습 환경과 중앙집중식 학습 환경에서 비교하는 실험을 하였다. 실험을 통해 연합학습 환경에서 LLM 파인튜닝을 수행할 때 도메인 이질성과 더불어 round 단위의 학습량을 함께 제어해야 함을 보였다.

키워드: 연합학습, 대규모 언어모델, 파인튜닝, 글로벌 에폭 보존, 토큰 사용량 제어

* Corresponding Author

※이 논문은 서울시 산학연 협력사업 2025년도 인공지능 기술사업화 지원사업(CY250165)의 지원을 받아 수행된 연구임

I. 서론

대규모 언어모델(Large Language Model, LLM)은 단순 질의응답을 넘어 의료, 금융과 같은 전문 도메인에서도 활용 가능성이 확대되고 있다. 전문 도메인에서 LLM을 효과적으로 활용하기 위해서는 사전학습된 모델을 도메인 특화 데이터로 추가 학습하는 파인튜닝(Fine-Tuning)이 필요하다. 그러나 도메인 데이터는 개인정보나 기관 내부 정보 등 민감정보가 포함될 수 있어, 외부 서버로 직접 전송하기 어렵다. 이 문제를 완화하기 위한 방법으로 연합학습(Federated Learning, FL)이 주목받고 있다. 이는 민감 데이터가 서버로 직접 이동하지 않는다는 점에서 데이터 주권이 중요한 도메인에 적합하다.

그러나 기존 연합학습 방법은 LLM 파인튜닝에 그대로 적용하기에 학습량 통제 측면에서 문제가 있다. 각 round마다 client가 고정된 local epoch을 수행한 뒤 서버가 이를 집계하는 구조이다보니 round의 증가가 local dataset 노출횟수의 증가로 이어진다. 이는 제한된 학습량을 전제하는 LLM 파인튜닝에서 의도하지 않은 학습을 발생시킨다.

본 논문에서는 이러한 문제를 해결하기 위해 연합학습 기반 LLM 파인튜닝에서 전체 누적 토큰 학습량에 대응되는 global epoch을 우선 정의하고, 이를 round 수에 따라 round별 sample index 구간으로 분할하는 누적 토큰 학습량 정합화 기반 학습 방법을 제안한다. 제안 방법은 round 수 증가로 인한 비의도적 학습량 증가를 방지하여 중앙집중식 학습과 동일한 누적 토큰 학습량 조건을 보장한다. 본 연구에서는 의료와 금융 도메인 데이터셋을 활용하여 단일 도메인과 복합 도메인 환경에서 제안 방법과 중앙집중식 학습을 비교 분석한다.

II. 관련 연구

2.1 LLM 파인튜닝과 학습량 통제

LLM의 Supervised Fine-Tuning(SFT)은 사전학습된 언어모델을 특정 작업이나 도메인 지식에 맞게 적응시키는 작업이다. 기존의 일반 언어 능력을 유지하면서 제한된 데이터로 세부적인 동작을 조정하는 단계이기 때문에 SFT는 학습량을 보수적으로 통제한다. 대표적인 공개 LLM 파인튜닝 모델인 Llama 2-Chat에서는 SFT를 2 epoch으로 수행하였고 [1], Hu et al.은 Llama-3-8B instruction tuning을 3 epochs로 수행하였다 [2]. 이러한 사례들을 바탕으로 본 연구에서는 global epoch을 2로 설정하고, 연합학습 환경에서 동일한 누적 토큰 학습량 조건을 보장할 수 있는 학습 과정을 제안한다.

2.2 PEFT와 LoRA

본 연구에서는 전체 파라미터 파인튜닝의 계산과 메모리 부담을 줄이기 위해 LoRA 기반 PEFT를 사용하였다 [3]. Qwen/Qwen3-4B-Instruct-2507 모델에 LoRA를 적용하여 medical, finance, mixed 도메인 파인튜닝을 수행하였다 [4].

2.3 연합학습 기반 LLM 파인튜닝

연합학습은 client의 원본 데이터를 중앙 서버로 직접 보내지 않고, 각 client에서 파인튜닝하여 업데이트된 파라미터만 서버에 전달하여 global model을 학습하는 분산 학습 방식이다. 대표적인 집계 방식인 FedAvg는 round마다 client가 local training을 수행한 뒤 서버에서 client의 업데이트된 파라미터의 평균을 계산하여 global model을 만든다 [5]. 그러나 client 간 데이터 분포가 다른 non-IID 환경에서는 update가 편향될 수 있으며, 이는 로컬 학습과 집계 주기가 성능에 영향을 준다는 것을 시사한다 [6][7]. 최근에는 이러한 연합학습 구조를 LLM 파인튜닝에 적용하려는 연구도 진행되고 있다 [8][9][10].

III. 제안방법

3.1 누적 토큰 학습량 불일치 문제

기존 방식에서는 round가 증가할수록 client가 local dataset에 노출 횟수가 증가하며 학습량도 비례하여 증가한다. 이는 중앙집중식 SFT에서 설정한 global epoch과 불일치할 수 있으며, 본 연구에서는 이를 누적 토큰 학습량 불일치 문제로 정의한다.

3.2 누적 토큰 학습량 정합화 방법

본 논문에서는 연합학습 환경에서도 중앙집중식 LLM 파인튜닝과 동일한 누적 토큰 학습량을 보장하기 위한 방법을 제시한다. client k의 local dataset의 크기($|D_k|$)와 전체 누적 토큰 학습량에 대응되는 global epoch(E_g)을 정의하고 이를 round 수(R)로 분할하는 누적 토큰 학습량 정합화 방법을 제안한다. client k가 각 round에서 학습해야 할 sample 수를 S_k 라고 두면 수식은 아래와 같다.

$$S_k = \frac{E_g \times |D_k|}{R}$$

client k가 r번 라운드에서 학습해야할 sample index 구간을 I_k^r 라고 하면 다음과 같이 정의된다.

$$I_k^T = [(r-1)S_k, rS_k)$$

이렇게 구한 구간에 맞게 연합학습을 수행하면 전체 누적 토큰 학습량을 E_g 로 고정할 수 있다.

3.3 실험 구성 및 결과

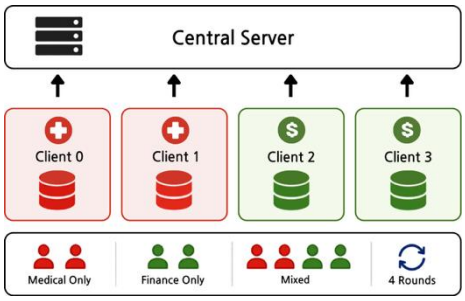


그림 1

그림 1은 본 연구의 연합학습 기반 LLM 파인튜닝 실험 구성을 나타낸다. 전체 client는 4개로 구성되며 client 0,1은 medical 도메인, client 2, 3은 finance 도메인을 담당한다. Medical Only는 0,1만 참여하고, Finance Only는 2,3만 참여하여 단일 도메인 파인튜닝 환경을 시험하며 Mixed는 4개의 client가 모두 참여하며 복합 도메인 환경에서 파인튜닝을 수행하는 설정이다. 데이터셋은 medical 도메인에 openlifescienceai/medmcqa를, finance 도메인에 sujet-ai/Sujet-Finance-Instruct-177k를 사용하였다. client마다 도메인별 데이터셋에서 20,000 samples를 할당하여 실험을 수행하였다.

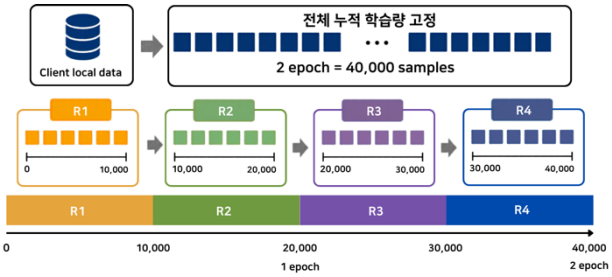


그림 2

연합학습 실험은 global epoch이 2가 되도록 round별 학습 구간을 분리하였다. 그림 2는 이를 시각적으로 나타내었다. 각 client는 라운드마다 10,000 samples를 학습한다.

모든 실험은 Qwen/Qwen3-4B-Instruct-2507 모델에 LoRA를 적용하여 수행하였다. optimizer는 AdamW를 사용하였고, learning rate는 3e-4로 설정하였다. learning rate scheduler는 cosine scheduler를 사용하였으며, warmup ratio는 0.06으로 설정하였다. LoRA 설정은 r=32, alpha=32, dropout=0.05로 구성하였다. 동일한 학습량을 수행한 중앙집중식 학습과 제안하는 방법이 적용된 연합학습을 medical, finance, general 벤치마크를 사용하여 비교하였다. general은 STEM, Social, Humanities, finance는 FPB, FIQA, TFNS, medical은 PubMedQA, MedQA, MedMCQA, CareQA로 평가후 평균을 내었다.

표1. 연합학습 환경 도메인별 점수

Experiment	General	Finance	Medical
vanilla	0.7014	0.5961	0.5977
Medical Only	0.7130	0.7264	0.6638
Finance Only	0.6958	0.8344	0.6536
Mixed	0.7108	0.8065	0.6621

표2. 중앙집중식 학습 환경 도메인별 점수

Experiment	General	Finance	Medical
vanilla	0.7014	0.5961	0.5977
Medical Only	0.6867	0.7160	0.52
Finance Only	0.6630	0.8333	0.6116
Mixed	0.6966	0.8223	0.6514

실험 결과 제안 방법을 적용한 연합학습은 중앙집중식 학습보다 전반적으로 뛰어난 성능을 보였다. 중앙집중식 학습의 Medical 도메인 모델이 Medical 벤치마크에서 낮게 나오는 특이사항이 있는데 답변을 확인해본 결과 PubMedQA에서 yes/no/maybe 형식 대신 학습 데이터셋의 답변 형식에 가까운 응답을 생성한 사례가 확인되었다. 전반적인 학습의 결과에서 파인튜닝의 목적과 유사한 성능 상승 효과를 확인할 수 있어 방법론의 적용이 유의미한 향상을 보였다고 볼 수 있다.

IV. 결론

본 논문은 연합학습 기반 LLM 파인튜닝에서 round 수의 증가가 의도하지 않은 학습량의 증가로 이어진다는 문제를 분석하였다. 이를 해결하기 위해 global epoch를 우선 정의하고, round별 sample index 구간으로 분할하는 누적 토큰 학습량 정합화 기반 학습 방법을 제안하였다. 이는 round 수의 증가가 전체 학습량의 증가로 이어지지 않아, 연합학습과 중앙집중식 학습이 동일한 누적 토큰 학습량 조건에서 비교될 수 있도록 한다. 의료와 금융 도메인의 데이터셋을 활용한 실험에서 제안 방법을 적용한 연합학습 파인튜닝 모델들이 전반적으로 중앙집중식 학습보다도 안정적인 성능 향상을 보였다.

본 연구는 연합학습에서 round 수 증가에 따른 누적 학습량 증가 효과를 통제함으로써, 집계 빈도와 학습량 증가 효과가 모델 성능에 미치는 영향을 분리하여 비교할 수 있도록 하였다는 점에서 의의를 가진다. 다만 적은 수의 client로 수행했다는 점과 dataset의 특성 차이가 고려되지 않았다는 점에서 한계가 존재한다. 향후 본 방법을 기반으로 데이터 이질성, 클라이언트 규모, 로컬 학습 조건 등 연합학습의 다양한 특성들에 대한 후속 연구로 확장할 수 있을 것이다.

참고문헌

[1] H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv preprint arXiv:2307.09288, 2023.

[2] H. Hu et al., "Fine-Tuning Large Language Models with Sequential Instructions," in Proc. NAACL, 2025.

[3] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," ICLR, 2022.

[4] Qwen Team, "Qwen/Qwen3-4B-Instruct-2507," Hugging Face Model Card, https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

[5] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," AISTATS, 2017.

[6] J. Zheng et al., "Safely Learning with Private Data: A Federated Learning Framework for Large Language Model," EMNLP, 2024.

[7] R. Ye et al., "OpenFedLLM: Training Large Language Models on Decentralized Private Data via Federated Learning," KDD, 2024.

[8] R. Singhal, K. Ponkshe, and P. Vepakomma, "FedEx-LoRA: Exact Aggregation for Federated and Efficient Fine-Tuning of Large Language Models," ACL, 2025.

[9] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated Optimization in Heterogeneous Networks," MLSys, 2020.

[10] S. P. Karimireddy et al., "SCAFFOLD: Stochastic Controlled Averaging for Federated Learning," ICML, 2020.

Evidential Deep Learning을 이용한 ABR 이미지 기반 난청 분류 및 불확실성 추정

권진오[O], 민지홍[*], 마준[**], 홍민[***]
[O***]순천향대학교 컴퓨터소프트웨어공학과
[*]순천향대학교 소프트웨어융합학과
[**]국립군산대학교 소프트웨어학과
e-mail: wlsdh338@sch.ac.kr[O], wlghd5524@sch.ac.kr[*], junma@kunsan.ac.kr[**]
mhong@sch.ac.kr[***]

Hearing Loss Classification and Uncertainty Estimation from ABR Images Using Evidential Deep Learning

Jino Kwon[O], Jihong Min[*], Jun Ma[**], Min Hong[***]
[O***]Dept. of Computer Software Engineering, Soonchunhyang University
[*]Dept. of Software Convergence, Soonchunhyang University
[**]Dept. of Software, Kunsan National University

요 약

청각 뇌간 반응(ABR) 이미지 기반 난청 자동 분류에서 기존 Softmax 기반 모델은 오분류 시에도 높은 확신도를 출력하는 과신뢰 문제로 인해 의료 환경에서의 신뢰성 확보에 한계가 있다. 본 연구는 이를 극복하기 위해 Evidential Deep Learning(EDL)을 적용하고, AlexNet, EfficientNetV2-S, DenseNet121, ConvNeXt-Tiny의 4종 백본에 대해 Softmax 기반 모델과 비교 실험을 수행하였다. 총 25,096장의 ABR 이미지를 대상으로 실험한 결과, EDL 기반 모델은 4종 백본 전체에서 ECE 및 Uncertainty AUROC가 일관되게 개선되었으며, Brier Score 또한 대부분의 백본에서 동등하거나 향상되었다. 특히 EfficientNetV2-S에서 ECE가 0.273에서 0.227로 가장 큰 폭으로 개선되었다. 분류 성능(Accuracy, AUC)은 전반적으로 동등 수준을 유지하였으며, Uncertainty-Accuracy Trade-off 분석을 통해 고확신 케이스는 자동 판정하고 고불확실 케이스는 전문의에게 연계하는 단계적 판정 체계의 구현 가능성을 확인하였다.

키워드: 청각뇌간반응, 난청분류, Evidential Deep Learning, 불확실성추정, 모델갤리브레이션

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

I. 서론

청각 뇌간 반응(Auditory Brainstem Response, ABR) 검사는 청각 자극에 대한 뇌간의 전기적 반응을 측정하는 객관적 청력 평가 방법으로, 신생아 청력 선별 검사 및 신경학적 이상 진단에 널리 활용된다[1]. ABR 검사 결과는 파형의 잠복기와 진폭을 분석하여 판독되며, 특히 Wave V 성분의 유무와 역치를 기반으로 난청 여부를 판정한다[1]. 그러나 파형 판독은 전문가의 주관적 판단에 의존하는 경우가 많아 판독자 간 일관성 확보가 어렵고, 대량의 검사 데이터를 처리할 때 판독 오류 가능성이 증가한다[2].

이러한 문제를 해결하기 위해 Ma et al.[3]은 AlexNet, VGG16 등 6종의 CNN 모델을 ABR 이미지에 적용하여 최고 95.93%의 분류 정확도를 달성하였다. 그러나 기존 연구들은 Softmax 기반의 결정론적 출력만을 제공하여, 모델이 틀릴 가능성이 높은 경우에도 높은 확신도를 출력하는 과신뢰 문제를 내포하고 있다[4]. 이는 오탐지가 임상적으로 중대한 결과를 초래할 수 있는 의료 환경에서 모델 신뢰성을 정량적으로 제시하지 못한다는 한계로 이어진다[5].

본 연구에서는 이러한 한계를 극복하기 위해 Evidential Deep Learning(EDL)을 ABR 이미지 난청 분류에 적용한다. EDL은 Dirichlet 분포 기반의 불확실성 추정을 통해 분류 결과와 함께 예측 신뢰도를 정량적으로 산출하여 Softmax 모델의 과신뢰 문제를 구조적으로 해결한다[6]. AlexNet, EfficientNetV2-S[7], DenseNet121[8], ConvNeXt-Tiny[9]의 4종 백본에 대해 Softmax 기반 모델과 EDL 기반 모델을 동일한 조건 하에 비교 실험하고, ECE(Expected Calibration Error)[4], Brier Score[12], Uncertainty AUROC[13]를 통해 EDL의 신뢰도 개선 효과를 정량적으로 검증한다.

II. 관련 연구

2.1 ABR 이미지 기반 딥러닝 분류

McKearney et al.[10]은 심층 CNN을 활용하여 ABR 파형을 3단계로 자동 분류하고 92.9%의 정확도를 달성하였다. Ma et al.[3]은 노이즈 제거 및 파형 정규화 등의 전처리 과정을 거친 ABR 이미지에 AlexNet, DenseNet121, VGG16 등 6종의 CNN 모델을 적용하여 비교 실험을 수행하였으며, AlexNet이 95.93%의 최고 정확도를 기록하였다. 그림 1은 해당 전처리 파이프라인을 통해 구축된 난청 및 정상 ABR 이미지 예시를 나타낸다. 그러나 이들 연구는 모두 Softmax 기반의 결정론적 분류 출력만을 제공하여 예측의 신뢰도를 정량적으로 제시하지 못한다는 한계를 지닌다.

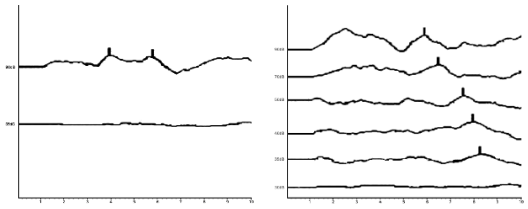


그림 1. ABR 전처리 이미지 데이터
(좌: 난청, 우: 정상)

2.2 Evidential Deep Learning

Evidential Deep Learning(EDL)은 Sensoy et al.[6]이 제안한 방법론으로, 주관적 논리와 Dirichlet 분포를 기반으로 분류 결과와 함께 예측

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

불확실성을 정량적으로 추정한다. EDL 기반 모델은 Softplus로 evidence를 출력하며, Dirichlet 분포의 집중 파라미터 αi와 예측 확률 $\hat{p}i$, 불확실성 u는 식 (1)-(3)과 같이 정의된다[6].

$$\alpha_i = e_i + 1 \tag{1}$$

$$\hat{p}_i = \alpha_i / S, \quad S = \Sigma\alpha_i \tag{2}$$

$$u = K / S \tag{3}$$

여기서 K는 클래스 수이다. 본 연구에서 K는 2이며, u가 1에 가까울수록 예측 신뢰도가 낮음을 나타내며, evidence 부족 시 u가 1에 수렴하는 구조로 Softmax의 과신뢰 문제를 구조적으로 해결한다. Dawood et al.[11]은 EDL 기반 불확실성 추정이 의료 영상 분류에서 모델 캘리브레이션 개선에 효과적임을 확인하였으며, 이는 오판단이 임상적으로 중대한 결과를 초래할 수 있는 의료 환경에서 특히 유용하다.

III. 제안방법

3.1 데이터셋

본 연구는 순천향대학교 부속병원에서 수집된 ABR 검사 이미지를 사용하였다. 데이터셋은 난청 환자 이미지 16,857장과 정상 청력 이미지 8,239장으로 구성된 총 25,096장이며, 2.1절에서 소개한 Ma et al.의 전처리 파이프라인을 통해 구축된 데이터를 활용하였다. 전체 데이터는 학습 70%, 검증 15%, 테스트 15%의 비율로 분할하였다. ABR 검사 특성상 좌우 귀는 독립적인 신경 경로를 측정하므로 동일 환자의 좌우 귀 이미지는 임상적으로 독립된 데이터로 취급하였다. 난청 클래스가 정상 클래스 대비 약 2:1의 불균형을 보이는데, 이는 ABR 검사가 난청이 의심되는 환자를 대상으로 주로 시행되는 임상적 특성에 기인한다.

3.2 모델 구조

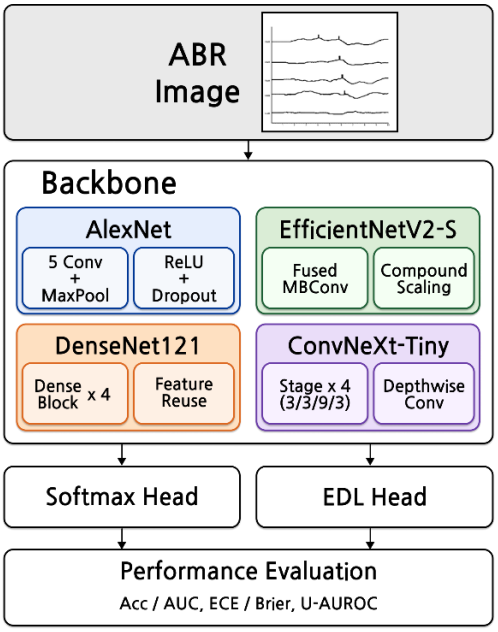


그림 2. 모델 파이프라인 구조

그림 2는 본 연구의 전체 모델 파이프라인 구조를 나타낸다. 본 연구는 ImageNet 사전학습 가중치를 적용한 4종의 백본 네트워크를 동일한 구조 하에 비교 실험하였다. 각 백본의 출력 특징 벡터는 완전연결층을 거쳐 Softmax 헤드와 EDL 헤드에 전달된다. Softmax 기반 모델은 Softmax로 확률을 출력하는 반면, EDL 기반 모델은 Softplus로 evidence를 출력하여 불확실성을 함께 산출한다.

3.3 학습 설정

모델 학습에는 EDL 손실 함수를 적용하였으며, KL 발산 항의 가중치를 점진적으로 증가시키는 annealing 기법을 사용하여 안정적인 불확실성 학습을 유도하였다. Softmax 기반 모델에는 NLL(Negative Log-Likelihood) 손실 함수를 적용하였다. 두 모델 모두 Adam 옵티마이저($lr = 1e-4$)와 CosineAnnealing 스케줄러를 사용하였다.

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

ImageNet 사전학습 모델의 표준 입력 크기에 맞추기 위해 224×224로 리사이즈하였으며, ImageNet 평균과 표준편차로 정규화하였다.

IV. 실험 및 결과

4.1 Softmax 및 EDL 모델 분류 성능 비교

표 1은 4종 백본에 대해 두 모델의 분류 성능을 비교한 결과이다. EDL 기반 모델은 백본에 따라 Accuracy 및 AUC가 소폭 향상되거나 동등 수준을 유지하였다. 이는 EDL이 불확실성 학습에 따른 분류 성능 저하를 최소화하면서 Softmax와 동등한 분류 성능을 달성할 수 있음을 시사한다.

표1. 백본별 분류 성능 비교

Backbone	Softmax		EDL	
	Acc	AUC	Acc	AUC
AlexNet	0.890	0.929	0.885	0.933
DenseNet121	0.895	0.946	0.890	0.943
EfficientNet V2-S	0.878	0.939	0.891	0.942
ConvNeXt-Tiny	0.892	0.940	0.896	0.946

4.2 불확실성 품질 지표 비교

표 2는 두 모델의 불확실성 품질 지표를 비교한 결과이다. EDL 기반 모델은 4종 백본 전체에서 ECE 및 Uncertainty AUROC가 Softmax 대비 일관되게 개선되었으며, Brier Score 또한 대부분의 백본에서 동등하거나 향상되는 경향을 보였다. ECE는 특히 EfficientNetV2-S에서 0.273에서 0.227로 가장 큰 폭으로 감소하였으며, Uncertainty AUROC는 AlexNet에서 0.788에서 0.818로 가장 크게 향상되었다. 이를 통해 EDL이 모델 캘리브레이션과 오분류 탐지 성능을 동시에 개선함을 확인하였다.

표2. Softmax vs EDL 불확실성 품질지표 비교

Backbone	Softmax		EDL	
	ECE	Brier	ECE	Brier
AlexNet	0.236	0.089	0.218	0.085
DenseNet121	0.259	0.081	0.232	0.081
EfficientNet V2-S	**0.273**	0.091	**0.227**	0.083
ConvNeXt-Tiny	0.243	0.083	0.241	0.079

Backbone	Softmax	EDL
	Uncertainty AUROC	
AlexNet	**0.788**	**0.818**
DenseNet121	0.830	0.834
EfficientNet V2-S	0.813	0.820
ConvNeXt-Tiny	0.821	0.828

그림 3은 Softmax와 EDL 기반 모델의 캘리브레이션 곡선을 비교한 결과이다. Softmax 기반 모델은 예측 확률 0.1~0.5 구간에서 곡선이 대각선을 상회하는 경향을 보여 캘리브레이션 편향이 나타난다. 반면 EDL 기반 모델은 전 구간에 걸쳐 대각선에 근접한 분포를 보여, 예측 확률과 실제 정확도 간의 일치도가 전반적으로 향상되었음을 시각적으로 확인할 수 있다. 이는 ECE 수치 개선과 일치하는 경향이다.

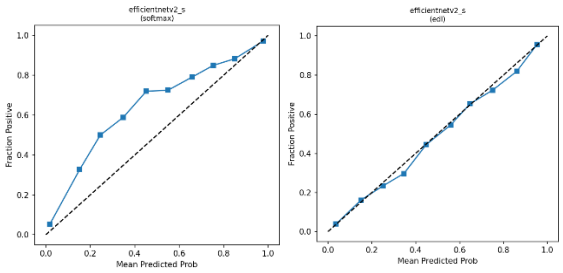


그림 3. Softmax vs EDL Calibration 곡선 비교 (Backbone: EfficientNetV2-S)

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

4.3 불확실성 분포 분석

그림 4는 EDL 기반 모델의 예측 불확실성 분포를 나타낸다. 대표적으로 AlexNet과 ConvNeXt-Tiny를 제시하였으며, 나머지 2종 백본에서도 동일한 경향이 관찰되었다. 공통적으로 정답 케이스가 오답 케이스에 비해 낮은 불확실성 구간에 집중되는 경향을 보였다. AlexNet의 경우 u=0.3 이후 구간에서 오답 케이스의 비율이 급격히 증가하며, 고불확실성 구간에서 오분류 케이스가 집중되는 패턴이 뚜렷하게 나타난다. ConvNeXt-Tiny는 정답 케이스의 대부분이 u=0.1 이하에 밀집되어 있어 정답과 오답 간 불확실성 분리가 더욱 명확하며, 이는 4.1절에서 확인된 ConvNeXt-Tiny의 상대적으로 높은 분류 성능과도 일치하는 경향이다. 이는 EDL이 오분류 가능성이 높은 케이스에 대해 높은 불확실성을 부여하는 방식으로 정상적으로 동작함을 나타내며, 단순 분류 결과 외에 예측의 신뢰도를 정량적으로 구분할 수 있음을 확인하였다.

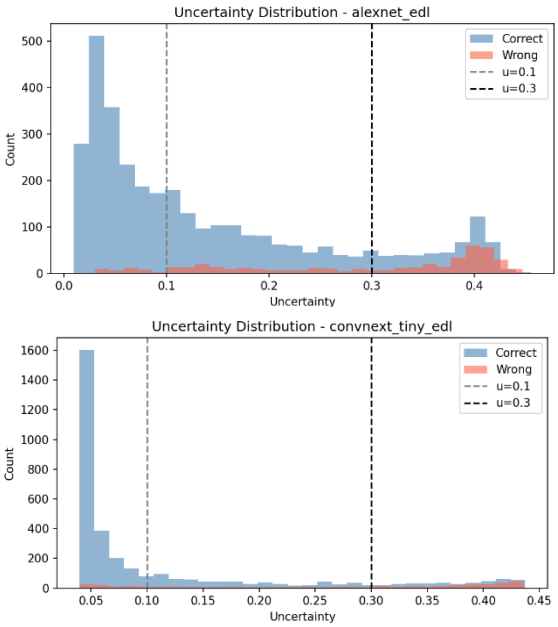


그림 4. 백본별 예측 불확실성 분포

(Backbone: AlexNet, ConvNeXt-Tiny)

4.4 Uncertainty-Accuracy Trade-off 분석

그림 5는 ConvNeXt-Tiny 기반 EDL 모델을 대표 예시로 제시하며, 불확실성 임계값에 따른 분류 정확도 및 전문의 의뢰율(Referral Rate)의 변화를 나타낸다. 불확실성이 높은 케이스를 제외할수록 잔류 케이스의 분류 정확도가 단조 증가하는 Uncertainty-Accuracy Trade-off 특성이 관찰되었다. 임계값 u=0.1 적용 시 의뢰율 약 40%를 감수하는 대신 정확도가 약 0.98까지 향상되었으며, u=0.3 적용 시에는 의뢰율 약 5% 수준에서 정확도 약 0.96을 달성하여 실용적인 운용 범위 내에서도 유의미한 성능 향상이 가능함을 확인하였다. 이는 EDL 기반 불확실성 추정을 활용하여 모델이 확신하는 케이스는 자동 판정하고, 불확실성이 높은 케이스는 전문의에게 연계하는 단계적 판정 체계의 구현 가능성을 시사한다.

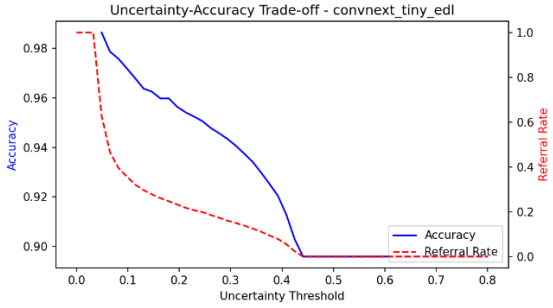


그림 5. 불확실성 임계값에 따른 분류 정확도 변화

(Backbone: ConvNeXt-Tiny)

V. 결론

본 연구에서는 ABR 이미지 기반 난청 자동 분류를 위해 Evidential Deep Learning(EDL)을 적용하고, Softmax 기반 모델과의 비교 실험을 통해 EDL의 신뢰도 개선 효과를 정량적으로 검증하였다. AlexNet, EfficientNetV2-S, DenseNet121,

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

ConvNeXt-Tiny의 4종 백본에 대해 동일한 조건 하에 실험한 결과, EDL 기반 모델은 4종 백본 전체에서 ECE 및 Uncertainty AUROC가 Softmax 대비 일관되게 개선되었으며, Brier Score 또한 대부분의 백본에서 동등하거나 향상되었다. 분류 성능(Accuracy, AUC)은 백본에 따라 동등하거나 소폭 향상되는 양상을 보였다. 이는 EDL이 Softmax의 과신뢰 문제를 구조적으로 완화하고, 오분류 가능성이 높은 케이스에 대해 높은 불확실성을 부여함으로써 신뢰도 기반 판정이 가능함을 나타낸다. 또한 불확실성 임계값에 따른 Uncertainty-Accuracy Trade-off 분석을 통해, 모델이 확신하는 케이스는 자동 판정하고 불확실성이 높은 케이스는 전문의에게 연계하는 단계적 판정 체계의 구현 가능성을 확인하였다.

본 연구의 한계로는 ECE 절대값이 0.2대로 높아 캘리브레이션 품질의 추가적인 개선이 필요하며, Softmax 대비 EDL의 개선 폭이 크지 않은 점을 들 수 있다. 이는 EDL annealing 파라미터 최적화 및 학습 전략 개선을 통해 보완 가능할 것으로 판단된다. 향후 연구에서는 Vision Transformer 등 최신 아키텍처로의 확장과 EDL annealing 최적화를 통한 불확실성 추정 품질 개선, 그리고 임상 데이터를 활용한 불확실성 임계값의 실제 검증을 진행할 예정이다. 본 연구의 의의는 단순한 분류 성능 향상을 넘어, EDL을 통해 모델 스스로 예측의 불확실성을 인식하고 이를 임상에 전달함으로써 의료 AI의 신뢰성과 안전성 향상 방향을 정량적으로 제시한 데 있다.

참고문헌

[1] A. Young, J. Cornejo, and A. Spinner, "Auditory Brainstem Response," StatPearls, Jan. 2023.

[2] K. F. P. Naves et al., "Assessment of inter-examiner agreement and variability in the manual classification of auditory brainstem response," BioMedical Engineering OnLine, vol. 11, no. 86, 2012.

[3] J. Ma et al., "Comparative Analysis of CNN Models for Hearing Loss Classification Using ABR Images," Diagnostics, vol. 14, no. 12, p. 1232, 2024.

[4] C. Guo et al., "On Calibration of Modern Neural Networks," ICML, 2017.

[5] A. Kurz et al., "Uncertainty Estimation in Medical Image Classification: Systematic Review," JMIR Medical Informatics, 2022.

[6] M. Sensoy et al., "Evidential Deep Learning to Quantify Classification Uncertainty," NeurIPS, 2018.

[7] M. Tan and Q. Le, "EfficientNetV2: Smaller Models and Faster Training," ICML, 2021.

[8] G. Huang et al., "Densely Connected Convolutional Networks," CVPR, 2017.

[9] Z. Liu et al., "A ConvNet for the 2020s," CVPR, 2022.

[10] R. M. McKearney and R. C. MacKinnon, "Objective Auditory Brainstem Response Classification Using Machine Learning," International Journal of Audiology, vol. 58, no. 4, pp. 224–230, 2019.

[11] T. Dawood et al., "Addressing Deep Learning Model Calibration Using Evidential Neural Networks," ISBI, 2023.

[12] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," Monthly Weather Review, 1950.

[13] D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and OOD Examples," ICLR, 2017.

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

온프레미스 프레젠테이션 기록, 요약 및 분석 시스템

조진만[O], 민지홍[*], 홍민[**]
[O**]순천향대학교 컴퓨터소프트웨어공학과
[*]순천향대학교 소프트웨어융합학과
e-mail: cjm8685@naver.com[O], wlghd5524@gmail.com[*], minhong71@gmail.com[**]

An On-premise Presentation Recording, Summarization and Analysis System

JinMan Cho[O], JiHong Min[*], Min Hong[**]
[O**]Dept. of Computer Software Engineering, Soonchunhyang University
[*]Dept. of Software Convergence, Soonchunhyang University

요 약

최근 대규모 언어 모델(LLM)을 활용한 자동 요약 및 분석 기술이 증가하고 있다. 하지만 실제 서비스 환경에서는 클라우드 API 의존 구조로 인해 보안 문제와 네트워크 지연 등의 한계가 존재한다. 특히, 기업 회의나 학술 프레젠테이션은 미공개 연구 데이터 등 민감한 정보를 다루는 경우가 많아 외부 데이터 유출을 차단하는 온프레미스(On-premise) 환경의 AI 도입이 필수적이다. 이에 본 연구는 RTX 4090(24GB) 단일 GPU 환경에서 동작하는 온프레미스 발표 기록 및 분석 시스템을 제안한다. 본 시스템은 음성(STT) 텍스트화, LLM 기반 내용 요약, 발표 자료 교차 분석을 통한 누락 정보 및 예상 질의 도출 기능을 제공하며 맞춤형 PPT 구성안까지 자동 기획하는 통합 솔루션이다. 이를 구현하기 위해 STT 단계에는 Faster-Whisper를 사용하고, 요약 및 분석에는 4-bit NF4 및 이중 양자화 기법을 적용한 Qwen2.5-14B-Instruct 모델을 적용하였다. 요약 품질은 Gemini 3.1 Pro 출력을 절대적 정답이 아닌 상용 LLM 기반 비교 기준(Baseline)으로 설정하고 ROUGE 점수를 측정한 결과, 평균 ROUGE-1 0.70, ROUGE-L 0.62를 기록하였다. 이는 제한된 컴퓨팅 자원에서도 상용 모델 출력과 일정 수준의 유사도를 달성할 수 있음을 보여준다. 실행 성능 측면에서는 6분 길이 입력 기준 VRAM 피크가 12.27GB, 첫 토큰 응답 시간(TTFT) 0.04초, 처리량 9.67 tokens/sec를 달성하여 단기 입력에 대한 실용 가능성을 보였다. 반면 입력 길이가 16분 수준의 장문 입력에서는 속도 저하가 관찰되어 이러한 결과는 향후 장문 입력 처리를 위한 텍스트 분할(Chunking)과 같은 추가적인 최적화 기법의 필요성을 시사한다.

키워드: 온프레미스, 대규모 언어 모델, 단일 GPU, STT, 프레젠테이션 분석

※ 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음(2021-0-01399)

I. 서론

최근 대규모 언어 모델(LLM)의 발전으로 프레젠테이션 음성 기록 및 요약, 자료 생성 등 업무 자동화 수요가 증가하고 있다[1]. 그러나 대다수의 상용 서비스는 클라우드 API를 기반으로 동작하여, 기업 내부 회의나 미공개 연구 데이터를 다룰 때 심각한 데이터 유출 및 보안 취약점이라는 한계를 지닌다[2]. 따라서 민감한 데이터를 다루는 환경에서는 외부 API 의존도를 낮추고, 로컬 인프라 내부에서 추론을 수행하는 온프레미스(On-premise) 방식이 하나의 실질적인 대안이 될 수 있다[3]. 이에 본 연구는 로컬 환경에서 데이터를 안전하게 처리하는 단일 GPU 기반 온프레미스 AI 기록 및 분석 시스템을 제안한다. 제안 시스템은 기존 STT-LLM 파이프라인이 전사 후 요약에 초점을 두는 것과 달리, 4-bit 양자화된 로컬 LLM을 활용하여 발표 자료(PPT)와의 교차 분석, 누락 정보 식별, 예상 질의 도출, 맞춤형 PPT 구성안 자동 기획 등 프레젠테이션 특화 통합 기능을 제공하며, 정량적 요약 품질 및 실행 성능 평가를 통해 제한된 자원의 로컬 환경에서도 실무적 활용의 가능성을 제시한다.

II. 관련 연구

2.1 STT 기반 음성 전사

음성인식(Speech-to-Text)은 회의록 작성, 자막 생성, 발표 기록 등 음성을 기반으로 하는 업무 자동화에 활용되며 특히 발표 분석에서는 정확도와 더불어 발화 순서 및 문장 단위 흐름의 유지가 중요하며, 이를 위해 트랜스포머(Transformer) 아키텍처를 도입하는 연구가 중요하게 다루어진다[4]. 최근에는 대규모 오디오 데이터를 사전학습한 Whisper계열 모델이 음성인식의 주요 선택지로 자리잡았으며, 이를 CTranslate2 기반으로 최적화한 Faster-Whisper 구현은 제한된 자원 환경에서 추론 속도와 메모리 효율을 높이는데 유리하다[5].

2.2 LLM 기반 요약 및 발표 분석

대규모 언어 모델(LLM)은 텍스트의 요약, 질의응답, 문서 구조화 등 다양한 자연어 처리 작업에 활용된다[6]. 최근에는 단순한 스크립트 압축을 넘어서 회의 및 발표 요약 벤치마크가 등장하는 등 영역이 확장되는 추세이다[7]. 단일 GPU와 같은 제한된 환경에서 대규모 파라미터 모델을 효율적으로 구동하기 위해 모델의 가중치를 압축하는 양자화(Quantization) 기법이 핵심 연구로 떠오르고 있다[8]. 특히 NF4(NormalFloat4) 및 이중 양자화(Double Quantization)를 결합한 4-bit 양자화 방식은 LLM 본연의 추론 성능 손실을 줄이면서도 VRAM 요구량을 낮출 수 있어, 소비자용 단일 GPU 환경의 LLM 구동에 널리 활용된다[9]. 이러한 로컬 LLM을 활용하여 원본 시각 자료와 실제 발화 내용 간의 정합성을 파악하고 이를 연계 분석하는 기초 연구가 진행되고 있다[10]. 나아가 이를 바탕으로 청중 관점의 예상 질의를 자동 도출하는 등 프레젠테이션에 특화된 분석 기능의 필요성도 점차 대두되고 있다[11].

2.3 온프레미스 AI와 보안성

클라우드 API 기반 AI 서비스는 편의성이 높지만 데이터가 외부 서버로 전송되어 민감한 정보 노출 등 보안 취약점이 존재한다[12]. 따라서 기업 내부 회의나 학술 발표 등 민감한 데이터를 다루는 환경에서는 외부망과 분리된 자체 자원으로 독립 구동하는 온프레미스(On-premise) 방식이 실질적인 대안으로 떠오르고 있다[3].

※ 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음(2021-0-01399)

III. 제안방법

3.1 시스템 구조

제안 시스템은 발표 녹화 이후의 후처리 과정을 로컬 환경에서 수행하기 위한 온프레미스 기반 AI 발표 기록 및 분석 시스템이다. 전체 시스템 흐름도는 그림 1과 같다.

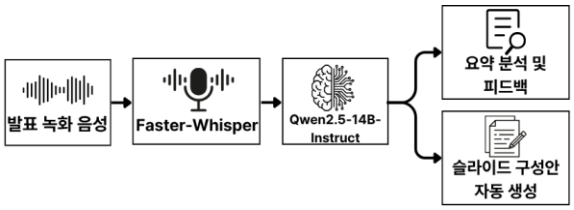


그림 1. 전체 시스템 흐름도

시스템은 발표 음성 전사, 발표 자료 기반 분석, 사용자 스타일 기반 PPT 구성안 생성 기능으로 구성된다. 발표 분석 기능에서는 발표 녹화 파일을 업로드하면 로컬 환경에서 STT 전사와 LLM 기반 요약을 수행한다. 사용자가 필요에 따라 PPT 자료를 추가로 업로드하면, 시스템은 STT 전사문과 PPT 내용을 비교하여 누락된 정보와 예상 질의를 도출한다. PPT 구성안 생성 기능에서는 사용자가 기존에 제작한 PPT와 신규 발표 주제를 입력하면 기존 자료의 구성 방식과 스타일을 반영한 슬라이드 구성안을 생성한다. 발표 기록 및 분석 과정에서는 발표 녹화 음성이 Faster-Whisper에 입력되어 전사문으로 변환된다. 그 후 발표 자료에서 추출한 텍스트와 함께 Qwen2.5-14B-Instruct 모델에 입력되어 요약 및 분석이 수행된다. 모든 과정은 외부 클라우드 API 호출 없이 로컬 환경 내에서 독립적으로 수행되도록 설계하였다.

3.2 발표 기록 및 분석 절차

발표 기록 및 분석 절차는 발표 녹화 음성을 입력으로 받아 STT 전사와 LLM 기반 요약을 수행하는 과정으로 구성되며 데이터 흐름과 파이프라인은 그림 2와 같다.

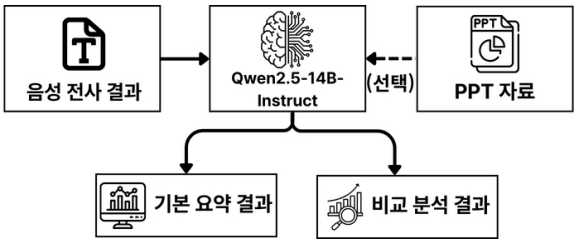


그림 2. 발표 기록 및 분석 절차 파이프라인

먼저 입력된 발표 녹화 음성은 Faster-Whisper를 통해 문맥에 맞는 텍스트 전사문으로 변환된다. 이후 본 연구에서 설계한 로컬 LLM 분석 엔진을 통해 핵심 내용 요약 및 비교 분석이 수행된다. 단일 GPU 환경에서의 자원 최적화와 추론 결과의 일관성 확보를 위해 적용한 LLM 엔진의 세부 설정 및 파라미터는 다음 표 1과 같다.

표 1. LLM 엔진의 세부 설정 및 파라미터

항목	적용 기술 및 설정 값
프레임워크, 라이브러리	HuggingFace Transformers, bitsandbytes
양자화 방식	4-bit NF4(NormalFloat4)
메모리 최적화 기법	이중 양자화 (Double Quantization)
연산 자료형	fp16
탐색 알고리즘	탐욕 탐색 (Greedy Decoding)
반복 페널티	1.1
최대 생성 토큰	1024

사용자는 발표 자료를 별도로 입력하지 않더라도 녹화된 발표 음성만으로 기본적인 핵심 발표 기록과 요약 결과를 확인할 수 있다. 발표 자료와의 비교 분석은 사용자가 필요에 따라 추가로 수행할 수 있는 기능이다. 사용자가 기존 PPT 자료를

※ 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음(2021-0-01399)

업로드하면, 시스템은 STT 전사문 전체와 PPT에서 추출된 내용을 함께 LLM에 입력하여 두 자료 간의 관계를 분석한다. 이를 통해 발표 자료에는 포함되어 있으나 실제 발화에서 설명되지 않은 부분을 식별하고 발표 내용과 자료 사이의 보완이 필요한 부분을 도출한다. 그리고 분석 결과를 바탕으로 청중이나 심사자가 제기할 가능성이 있는 예상 질의를 생성한다. 본 파이프라인을 통해 구현된 실제 발표 기록 및 분석 결과 화면은 그림 3 및 그림 4와 같다.

그림 3. 기록 및 요약 결과 화면

그림 4. 분석 및 예상 질의응답 생성 결과 화면

3.3 사용자 맞춤 PPT 구성안 생성 절차

사용자 맞춤 PPT 구성안 생성 기능은 발표 녹화 분석과 별도로 동작하며 새로운 주제에 적합한 PPT 구성안을 생성하는 것을 목표로 한다. 데이터 흐름과 파이프라인은 그림 5와 같다.

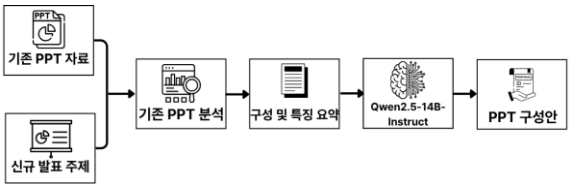


그림 5. PPT 구성안 생성 절차 파이프라인

시스템은 입력된 기존 PPT에서 슬라이드 제목, 본문 구성, 내용 전개 순서 등의 특징을 파악하여 요약한다. 이후 Qwen2.5-14B-Instruct 모델은 분석된 스타일적 특징을 신규 발표 주제에 적용하여 유사한 흐름을 유지하는 슬라이드별 구성안을 생성한다. 생성 결과는 텍스트 기반의 제목, 핵심 내용, 시각 자료 제안 등으로 구성되며, 이를 통해 사용자는 기존 스타일을 유지하면서 새로운 주제의 초기 발표 구성안을 기획할 수 있다. 실제 사용자 스타일을 반영하여 자동 생성된 맞춤형 PPT 구성안은 그림 6과 같다.

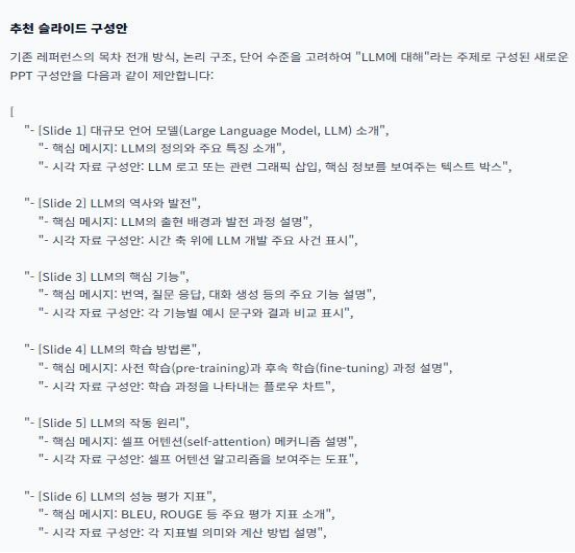


그림 6. PPT 구성안 생성 화면

IV. 실험 및 결과

4.1 실험 환경 및 평가 방법

제안 시스템의 적용 가능성을 확인하기 위해 단일 GPU 기반 로컬 환경에서 실험을 수행하였다. 실험 환경은 AMD Ryzen 9 5900X(12-Core 24-Thread 3.7Ghz) CPU, NVIDIA RTX 4090 (24GB) GPU, DDR4-2666MHz 128GB (32GBx4) RAM이 장착된 Windows 10 Pro 기반의 온프레미스

※ 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음(2021-0-01399)

데스크톱 환경으로 구축하였으며, STT는 Faster-Whisper, LLM은 4-bit 양자화 Qwen2.5-14B-Instruct를 사용하였다. 입력 데이터는 녹화된 데이터를 활용하였다. 본 실험의 평가는 실행 성능과 요약 품질을 중심으로 수행되었다. 실행 성능은 LLM 추론 과정에서의 VRAM 피크 사용량, 첫 토큰 응답 시간(TTFT), 토큰 처리량(tokens/sec)을 기준으로 측정하였다. 요약 품질은 상용 클라우드 모델인 Gemini 3.1 Pro의 요약 출력을 기준으로 제안 시스템의 결과와 비교하여 ROUGE-1 및 ROUGE-L 점수를 산출하였다. 단, 평가에 사용한 기준은 평가자가 작성한 절대적인 정답 요약이 아니므로 ROUGE 결과는 상용 모델 출력과의 상대적인 유사도를 확인하기 위한 지표로 해석한다.

4.2 실행 성능 및 요약 품질 평가

입력 데이터의 길이에 따른 시스템의 추론 성능 변화를 확인하기 위해서 발표 음성의 길이를 1분, 6분, 16분으로 세분화하여 실험을 진행하였다. 제안 시스템의 실행 성능 평가 결과는 다음 표 2와 같다.

표 2. 실행 성능 평가 결과

입력 길이(분)	VRAM 피크 사용량 (GB)	첫 토큰 응답시간 (TTFT, 초)	토큰 처리량 (tokens/sec)
1분	9.39	0.01	10.03
6분	12.27	0.04	9.67
16분	27.49	0.04	0.34

그리고 파이프라인 단계별 처리 시간 평가결과는 다음 표 3과 같다.

표 3. 파이프라인 단계별 처리 시간 평가 결과

입력 길이(분)	STT 전사 시간(초)	LLM 추론 시간(초)	전체 처리 시간(초)
1분	5.14	26.12	31.26
6분	51.49	28.68	80.17
16분	144.91	1369.38	1514.29

실험 결과, 6분 내외의 짧은 발표 입력에서는 단일 GPU 환경에서 충분한 메모리 마진을 유지하며 후처리 기반 발표 요약 시스템으로 활용될 가능성을 확인하였다. 반면, 16분 수준의 장문 입력의 경우 LLM의 어텐션(Attention) 연산 부담으로 VRAM이 물리적 한계인 24GB를 초과하여 드라이버의 공유 메모리 fallback이 동작하였고, 이로 인해 데이터 전송 병목이 추론 속도 저하의 원인으로 작용한 것으로 해석된다.

요약 품질 평가에서는 Gemini 3.1 Pro의 요약 결과를 기준으로 ROUGE를 산출하였으며, 평균 ROUGE-1은 0.70, ROUGE-L은 0.62를 기록하였다. 결과는 다음 표 4와 같다.

표 4. 요약 품질 평가 결과

입력 길이(분)	ROUGE-1	ROUGE-L
1분	0.8030	0.6617
6분	0.6410	0.6026
16분	0.6598	0.5882

1분 입력의 ROUGE가 상대적으로 높은 것은 짧은 입력일수록 참조 요약과의 어휘 중첩이 커지기 때문이며, 전반적으로 로컬 LLM 요약이 상용 모델과 일정 수준의 내용적 유사성을 나타냄을 확인할 수 있다.

※ 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음(2021-0-01399)

V. 결론

본 연구는 RTX 4090 단일 GPU 환경에서 동작하는 온프레미스 AI 발표 기록, 요약 및 분석 시스템 구조를 제안하고 핵심 구성 요소의 동작 가능성을 평가하였다. 실험 결과, 제한된 로컬 GPU 환경에서도 발표 전사 및 기본 요약 기능이 구동 가능하다는 점을 확인하였다. 다만 입력 길이 증가에 따른 속도 저하가 관찰되었으며 정량 평가가 요약 결과 유사도와 LLM 기본 실행 성능을 중심으로만 수행되었다는 한계가 존재한다. 따라서 향후 연구에서는 효율적인 텍스트 분할(Chunking) 및 계층적 요약 전략을 도입하여 연산 효율성을 개선할 예정이다. 또한 본 연구에서 제안한 PPT 발화 비교 분석, 예상 질의 생성, PPT 구성안 생성 등의 기능들은 아직 실질적인 검증이 이루어지지 않았으므로, 향후 실제 사용자 대상의 추가 평가를 진행하여 제안 시스템의 실효성을 검증할 계획이다.

참고문헌

[1] T. B. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[2] N. Carlini et al., "Extracting training data from large language models," in USENIX Security Symposium, 2021.

[3] Y. Yao et al., "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," High-Confidence Computing, vol. 4, no. 2, 2024, Article 100211.

[4] S. Karita et al., "A comparative study on Transformer vs RNN in speech applications," in Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 449–456.

[5] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in Proceedings of the International Conference on Machine Learning (ICML), 2023.

[6] L. Ouyang et al., "Training language models to follow instructions with human feedback," in Advances in Neural Information Processing Systems (NeurIPS), 2022.

[7] Y. Hu et al., "MeetingBank: A benchmark dataset for meeting summarization," in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2023.

[8] E. Frantar et al., "GPTQ: Accurate post-training quantization for generative pre-trained transformers," in Proceedings of the International Conference on Learning Representations (ICLR), 2023.

[9] T. Dettmers et al., "QLoRA: Efficient finetuning of quantized LLMs," in Advances in Neural Information Processing Systems (NeurIPS), 2023.

[10] H. Wang et al., "SlideSpeech: A large scale slide-enriched audio-visual corpus," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024.

[11] S. Chen et al., "PresentCoach: Dual-agent presentation coaching through exemplars and interactive feedback," arXiv preprint arXiv:2511.15253, 2025.

[12] H. Li et al., "PrivLM-Bench: A multi-level privacy evaluation benchmark for language models," in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2024, pp. 54–73.

※ 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음(2021-0-01399)

s

유방촬영술 기반 이진 분류에서 딥러닝 Backbone 아키텍처 성능 비교 분석

배준상 O, 민지홍*, 김태헌** 홍민***
O ***순천향대학교 컴퓨터소프트웨어공학과
* **순천향대학교 소프트웨어융합학과
e-mail : junsangbae12@sch.ac.kr O, wlghd5524@sch.ac.kr*,
taeheon1213@sch.ac.kr**, mhong@sch.ac.kr***

Comparative Analysis of Deep Learning Backbone Architectures for Binary Mammography Classification

Junsang Bae O, Jihong Min*, TaeHeon Kim** Min Hong***
O ***Dept. of Computer Software Engineering, Soonchunhyang University
* **Dept. of Software Convergence, Soonchunhyang University

요 약

유방촬영술(Mammography)은 유방암 조기 진단을 위한 대표적인 영상 검사 방법으로, 최근에는 딥러닝 기반 Computer-Aided Diagnosis(CAD) 시스템 연구가 활발히 진행되고 있다. 그러나 기존 연구들은 서로 다른 데이터셋과 학습 환경에서 성능을 보고하는 경우가 많아 Backbone Architecture 자체의 영향을 객관적으로 비교하기 어렵다는 한계가 있다. 본 연구에서는 VinDr-Mammo 데이터셋을 활용하여 CNN 기반 ResNet-101, Transformer 기반 ViT-B 및 Swin Transformer-B, 그리고 Self-Supervised Learning 기반 Foundation Model 인 DINOv3-B 를 동일한 학습 환경에서 비교 분석하였다. 데이터 전처리 단계에서는 Otsu Thresholding 기반 ROI Crop 을 적용하였으며, 클래스 불균형 완화를 위해 Random Under-Sampling 과 Weighted Binary Cross-Entropy Loss 를 사용하였다. 실험 결과, Transformer 기반 모델이 CNN 기반 모델보다 우수한 성능을 보였으며, DINOv3-B 가 Accuracy 77.63%, AUROC 0.7934 를 기록하여 가장 높은 분류 성능을 나타냈다. 본 연구는 유방촬영술 이진 분류에서 Backbone Architecture 의 영향을 정량적으로 비교하고, 효과적인 모델 선택을 위한 기준을 제시한다.

키워드: Mammography, Classification, Vision Transformer, Medical Image, Deep Learning

※ 본 연구는 2026 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구 결과로 수행되었음(2021-0-01399)

I. 서론

유방암은 전 세계 여성에서 가장 흔하게 진단되는 암 중 하나이며, 조기 진단 여부가 환자의 생존율에 큰 영향을 미친다[1]. 현재 유방암 조기 검진을 위해 유방촬영술(Mammography)이 표준 검사로 널리 사용되고 있으나, 치밀 유방(Dense Breast), 저대비 병변, 촬영 장비 차이 등으로 인해 판독의 어려움이 존재한다[2]. 이러한 한계를 보완하기 위해 딥러닝 기반 Computer-Aided Diagnosis(CAD) 시스템 연구가 활발히 진행되고 있다[3].

최근 유방촬영술 분류 연구에서는 ResNet과 같은 CNN 기반 모델뿐만 아니라 Vision Transformer(ViT), Swin Transformer 등 Transformer 기반 모델이 활용되며 성능 향상이 보고되고 있다.[4] 또한 Self-Supervised Learning 기반 Foundation Model이 컴퓨터 비전 분야에서 우수한 표현 학습 능력을 보이며 의료영상 분석 분야로도 적용 범위가 확대되고 있다.[4] 그러나 기존 연구들은 주로 특정 Backbone Architecture를 활용한 성능 향상에 초점을 두고 있으며, 서로 다른 데이터셋, 전처리 방법 및 학습 환경에서 결과를 보고하는 경우가 많아 Backbone Architecture 자체가 분류 성능에 미치는 영향을 객관적으로 비교하기 어렵다.

따라서 본 연구에서는 유방촬영술 분류 과제에서 Backbone Architecture에 따른 성능 차이를 분석하기 위하여 CNN 기반 모델인 ResNet-101,[5] Transformer 기반 모델인 ViT-Base[6] 및 Swin-Base,[7] 그리고 Self-Supervised Learning 기반 Foundation Model인 DINOv3[8]를 선정하였다. 모든 모델은 동일한 데이터셋과 학습 조건에서 학습 및 평가를 수행하였으며, Accuracy, AUROC, Sensitivity, Specificity 등의 지표를 활용하여 Backbone Architecture에 따른 분류 성능을 비교 분석하였다. 이를 통해 유방촬영술 분류 과제에서 각 Backbone의 특성과 성능을 정량적으로 평가하고, 효과적인 모델 선택을 위한 기준점을 제공하고자 한다.

II. 관련 연구

2.1 딥러닝 기반 유방촬영술 연구

최근 딥러닝 기술의 발전에 따라 유방촬영술 영상 분석 연구가 활발히 수행되고 있다. 유방촬영술 기반 인공지능 연구는 병변 탐지(Detection), 병변 분할(Segmentation), 위험도 예측(Risk Prediction), 유방암 분류(Classification) 등 다양한 분야에서 활용되고 있으며,[9] 특히 Classification 은 추가적인 병변 위치 정보 없이 학습이 가능하여 대규모 검진 데이터에 적용하기 용이하다는 장점이 있다.

초기 연구에서는 ResNet과 같은 CNN 기반 모델이 주로 사용되었으나, 최근에는 ViT 및 Swin Transformer 등 Transformer 기반 모델이 도입되며 우수한 성능을 보이고 있다.[4] 또한 유방촬영술 분류 연구는 입력 영상 구성 방식에 따라 단일 뷰 기반 방법과 다중 뷰 기반 방법으로 구분되며, 다양한 학습 전략이 제안되고 있다.

2.2 다중 뷰 기반 유방촬영술 분류

유방촬영 검사는 표준적으로 양측 유방에 대한 CC(Craniocaudal) 및 MLO(Mediolateral Oblique) 방향의 4 장 영상으로 구성되며, 숙련된 방사선과 전문의는 동측 뷰 및 대측 뷰 간 비교를 통해 병변의 존재 여부와 악성도를 종합적으로 판단한다[10]. 이러한 다중 뷰 판독 과정을 자동화하기 위한 딥러닝 연구가 활발히 수행되어 왔으며, 교차 어텐션 기반 뷰 간 특징 대응[11], Transformer 기반 다중

※ 본 연구는 2026 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음(2021-0-01399)

뷰 융합[12] 등이 단일 뷰 대비 분류 성능 향상을 보고하였다. 그러나 다중 뷰 기반 구조에서는 뷰 간 특징 융합(Feature Fusion) 과정이 추가적으로 개입되어 Backbone Architecture 자체의 영향을 독립적으로 분석하기 어렵다는 한계가 존재한다. 따라서 본 연구에서는 Backbone 간 성능 차이를 보다 명확하게 비교하기 위해 단일 뷰 기반 정상 및 비정상 이진 분류 환경을 구성하였다.

III. 제안방법

3.1 데이터셋 선정 및 준비

본 연구에서는 유방촬영술 영상 분류의 표준 벤치마크 데이터셋인 VinDr-Mammo 를 사용하였다[13]. 데이터 분할은 VinDr-Mammo 에서 제공하는 Train/Test Split 을 그대로 사용하였으며, 학습 데이터의 20%를 환자 단위(Patient-level Split)로 분리하여 검증 데이터로 활용하였다. 또한 기존 VinDr-Mammo 기반 유방촬영술 분류 연구와[14] 동일하게 BI-RADS 1-2 는 정상(Class 0), BI-RADS 3-6 은 비정상(Class 1)으로 정의하여 이진 분류를 수행하였다.

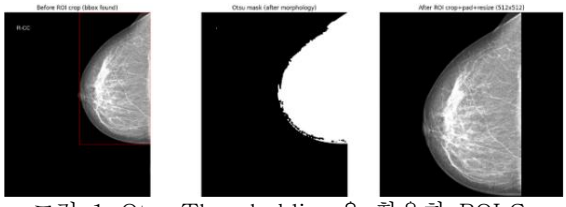


그림 1. Otsu Thresholding 을 활용한 ROI Crop

모델이 유방 영역에 집중하여 특징을 학습할 수 있도록 유방 영역 기반 ROI(Region of Interest) Crop 을 수행하였다. 먼저 pydicom 을 이용하여 DICOM 영상의 16-bit 픽셀 데이터를 추출한 후, Min-Max Normalization 을 적용하여 0-255 범위의 PNG 영상으로 변환하였다. 이후 Otsu Thresholding 을 이용하여 유방 실질 영역의 Binary Mask 를 생성하고, 가장 큰 Contour 를 기준으로 ROI 를 Crop 하였다. 그림 1 은 이러한 Crop 과정을 보여준다.

3.2 학습 모델

그림 2 는 본 연구의 전체 학습 파이프라인 구조를 나타낸다. Backbone Architecture 로는 잔차 연결(Residual Connection)을 기반으로 하는 대표적인 CNN 모델인 ResNet-101 과 이미지를 Patch 단위로 분할하여 Self-Attention 을 수행하는 Vision Transformer 기반의 ViT-Base, 계층적 구조와 Window-based Self-Attention 을 활용하여 고해상도 영상의 특징을 효율적으로 추출하는 Swin-Base, 그리고 대규모 Self-Supervised Learning 기반 사전학습을 통해 강력한 특징 표현 능력을 학습한 Foundation Model 인 DINOv3 ViT-Base (vit_base_patch16_dinov3)를 사용하였다. 모든 모델은 공개된 사전학습 가중치를 이용하여 초기화하였으며, 동일한 전이학습(Transfer Learning) 환경에서 학습 및 평가를 수행하였다.

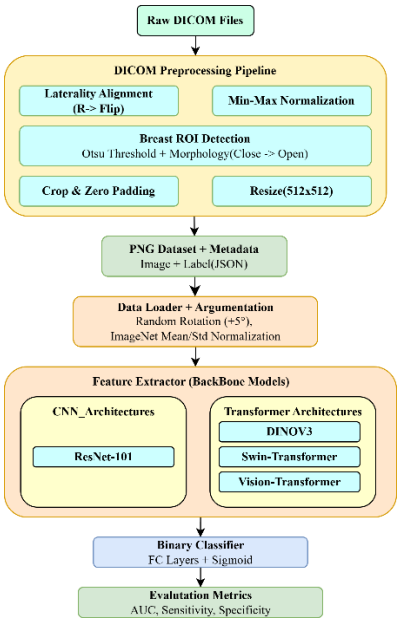


그림 2. 전체 학습 파이프라인

※ 본 연구는 2026 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음(2021-0-01399)

3.3 학습 설정

표 1. 하이퍼파라미터 설정

Parameter	Value
LR	1e-4
Batch_Size	32
Optimizer	AdamW
Scheduler	CosineWR(T_0=10, T_mult=2)
Max_Epochs	100
Patience	10
Loss	Weighted BCE

표 1 은 모델 학습에 사용된 주요 하이퍼파라미터를 나타낸다. Backbone Architecture 에 따른 성능 차이를 객관적으로 비교하기 위해 모든 실험에서 동일한 학습 환경을 적용하였다. 입력 영상은 모든 모델에 대해 384×384 크기로 통일하였으며, 학습 단계에서는 데이터 다양성 확보 및 일반화 성능 향상을 위해 Random Rotation(±5°) 증강을 적용하였다. Optimizer 는 AdamW(learning rate = 1×10^{-4})를 사용하였으며, 학습 초기의 탐색 능력을 확보하고 학습 후반의 안정적인 수렴을 유도하기 위해 Cosine Annealing Warm Restarts Scheduler ($T_0 = 10$, $T_{\text{mult}} = 2$) 을 적용하였다.

또한, 학습 데이터의 클래스 불균형 문제를 완화하기 위해 다수 클래스(정상 클래스)에 대해 무작위 언더샘플링(Random Under-Sampling)을 적용하였다. 언더샘플링은 학습 데이터에만 수행하였으며, 정상 클래스와 비정상 클래스의 비율이 약 2:1 이 되도록 조정하였다. 검증 및 테스트 데이터는 실제 데이터 분포를 유지하여 모델의 일반화 성능을 평가하였다. 그러나 언더샘플링 이후에도 클래스 간 표본 수 차이가 완전히 해소되지 않으므로, 추가적으로 Weighted Binary Cross-Entropy(WBCE) Loss 를 적용하였다. WBCE Loss 는 클래스별 가중치를 부여하여 소수 클래스의 오분류에 더 큰 패널티를 부여하는 손실 함수이다.

마지막으로, 모델의 과적합을 방지하기 위해 Early Stopping 을 적용하였으며, patience 는 10 으로 설정하였다. 또한 모든 실험은 동일한 하드웨어 및 소프트웨어 환경에서 수행하였다. 학습 환경은 Intel Core i7 프로세서와 NVIDIA RTX 4090 GPU(24GB VRAM)로 구성되었으며, 딥러닝 프레임워크로 PyTorch 2.7.0 과 CUDA 12.8 을 사용하였다.

IV. 실험 및 결과

4.1 모델 성능 비교

표 2 는 VinDr-Mammo 데이터셋을 대상으로 4 종 Backbone 의 분류 성능을 비교한 결과를 나타낸다. 모든 실험은 동일한 학습 환경에서 3 회 반복 수행하였으며, 평균과 표준편차를 계산하였다.

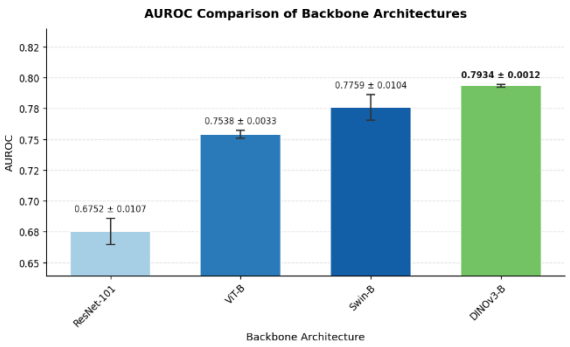


그림 3. 아키텍처 별 AUROC 비교

실험 결과, Transformer 기반 모델이 CNN 기반 모델보다 전반적으로 우수한 분류 성능을 보였다. ResNet-101 은 AUROC 0.6752 를 기록한 반면, ViT-B 는 0.7538, Swin Transformer-B 는 0.7759 를 기록하여 각각 약 7.9%p 및 10.1%p 향상된 성능을 나타냈다. 또한 Sensitivity 역시 ResNet-101

※ 본 연구는 2026 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음(2021-0-01399)

표 2. 모델 학습 결과

Model	ACC	AUROC	Sensitivity	Specificity
ResNet-101	75.35 ± 3.91	0.6782 ± 0.0107	0.4861 ± 0.0536	0.7819 ± 0.0488
ViT-B	74.32 ± 6.96	0.7538 ± 0.0033	0.6198 ± 0.0679	0.7563 ± 0.0842
Swin Transformer-B	75.13 ± 2.15	0.7759 ± 0.0104	0.6502 ± 0.0197	0.7620 ± 0.0259
DINOv3-B	**77.63 ± 2.73**	**0.7934 ± 0.0012**	**0.6684 ± 0.0330**	**0.7878 ± 0.0337**

의 0.4861 에서 ViT-B 0.6198, Swin Transformer-B 0.6502 로 증가하여 비정상 영상을 보다 효과적으로 탐지하는 것으로 확인되었다.

가장 우수한 성능은 DINOv3-B 에서 확인되었다. DINOv3-B 는 Accuracy 77.63%, AUROC 0.7934, Sensitivity 0.6684, Specificity 0.7878 을 기록하며 모든 평가 지표에서 가장 높은 성능을 달성하였다. 특히 AUROC 는 ResNet-101 대비 약 11.8%p 향상되었으며, Sensitivity 역시 약 10.1%p 증가하여 비정상 영상 검출 능력이 크게 향상되었다.

4.2 결과 분석

Backbone Architecture 에 따른 성능 차이를 분석한 결과, Transformer 기반 모델이 CNN 기반 모델보다 유방촬영술 이진 분류에 효과적인 것으로 나타났다. ResNet-101 은 상대적으로 높은 Specificity 를 보였으나 Sensitivity 가 가장 낮게 나타나 비정상 영상 검출에 한계를 보였다. 이는 CNN 기반 모델이 국소적인 특징 추출에 강점을 가지는 반면, Transformer 기반 모델에 비해 전역 문맥 정보를 활용하는 데 상대적인 한계가 있기 때문으로 판단된다.

반면 ViT-B 와 Swin Transformer-B 는 Self-Attention 기반 구조를 통해 영상 전역의 특징을 효과적으로 학습함으로써 ResNet-101 보다 우수한 성능을 보였다. 특히 Swin Transformer-B 는 계층적 구조와 Shifted Window Attention 을 활용하여 지역 및 전역 특징을 동시에 학습함으로써 ViT-B 대비 더 높은 AUROC 와 Sensitivity 를 기록하였다.

DINOv3-B 는 모든 Backbone 중 가장 우수한 성능을 보였으며, AUROC 의 표준편차 또한 ±0.0012 로 가장 낮게 나타났다. 이는 Self-Supervised Learning 기반 Foundation Model 사전학습이 보다 강력한 특징 표현 능력을 제공할 뿐만 아니라, 다양한 초기화 조건에서도 안정적인 성능을 유지할 수 있음을 시사한다. 따라서 유방촬영술 기반 이진 분류에서는 Transformer 기반 Backbone 이 CNN 기반 Backbone 보다 효과적이며, Foundation Model 기반 사전학습이 추가적인 성능 향상에 기여할 수 있음을 확인하였다.

V. 결론

본 연구에서는 유방촬영술 기반 유방암 이진 분류에서 Backbone Architecture 가 분류 성능에 미치는 영향을 분석하기 위해 CNN 기반 ResNet-101, Transformer 기반 ViT-B 및 Swin Transformer-B, 그리고 Self-Supervised Learning 기반

※ 본 연구는 2026 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음(2021-0-01399)

Foundation Model 인 DINOv3-B 를 비교하였다. 동일한 학습 환경에서 수행한 실험 결과, Transformer 기반 모델이 CNN 기반 모델보다 우수한 성능을 보였으며, DINOv3-B 가 Accuracy 77.63%, AUROC 0.7934, Sensitivity 0.6684 를 기록하여 가장 높은 분류 성능을 나타냈다. 또한 DINOv3-B 는 가장 낮은 AUROC 표준편차를 기록하여 우수한 성능과 안정성을 동시에 확인할 수 있었다.

실험 결과는 유방촬영술 이진 분류에서 Backbone Architecture 의 선택이 모델 성능에 중요한 영향을 미치며, 특히 Self-Supervised Learning 기반 Foundation Model 의 사전학습이 효과적인 특징 표현 학습에 기여할 수 있음을 시사한다. 향후 연구에서는 다중 뷰 기반 입력 환경과 다양한 Foundation Model 을 활용하여 Backbone 구조 및 사전학습 방식이 유방촬영술 분류 성능에 미치는 영향을 보다 심층적으로 분석할 예정이다.

참고문헌

[1] H. Sung et al., "Global Cancer Statistics 2020," CA Cancer J. Clin., vol. 71, no. 3, pp. 209–249, 2021.

[2] N. F. Boyd et al., "Mammographic density and the risk and detection of breast cancer," N. Engl. J. Med., vol. 356, no. 3, pp. 227–236, 2007.

[3] A. Rodriguez-Ruiz et al., "Stand-alone artificial intelligence for breast cancer detection in mammography," J. Natl. Cancer Inst., 111(9):916-922, 2019

[4] Li, Jun, et al. "Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives." Medical Image Analysis, vol. 79, 2022, p. 102462.

[5] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[6] Dosovitskiy, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations (ICLR), 2021.

[7] Liu, Ze, et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[8] Siméoni, Oriane, et al. "DINOv3." arXiv preprint arXiv:2508.10104, 2025.

[9] Kumar, S. Arun, et al. "Review on Deep Learning-Based CAD Systems for Breast Cancer Diagnosis." Technology in Cancer Research & Treatment, vol. 22, 2023, pp. 1-13.

[10] L. Tabár et al., "The TNM classification of the radiographic appearance of breast cancer," Radiol. Clin. North Am., 54(3):531-551, 2016.

[11] Y. N. T. Vu et al., "M&M: Tackling false positives in mammography with a multi-view and multi-instance learning sparse detector," in MICCAI 2023, Oct. 2023.

[12] S. Sarker et al., "MV-Swin-T: Mammogram classification with multi-view Swin Transformer," arXiv preprint arXiv:2402.16298, Feb. 2024.

[13] H. H. Pham et al., "VinDr-Mammo: A large-scale benchmark dataset for computer-aided detection and diagnosis in full-field digital mammography," PhysioNet, version 1.0.0, Mar. 2022, doi: 10.13026/br2v-7517.

[14] Petrini, Daniel G. P., et al. "Optimizing Breast Cancer Detection in Mammograms: A Comprehensive Study of Transfer Learning, Resolution Reduction, and Multi-View Classification." arXiv, 2025, arXiv:2503.19945v2.

※ 본 연구는 2026 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음(2021-0-01399)

Unity 엔진 기반 니파 바이러스 민·관 협력 대응 교육 시뮬레이션 시스템 개발

김상호[O], 강병훈[*], 권진오[*], 박민성[*], 전수영[*], 조현우[*], 정영웅[*], 배준상[*], 정현우[**],
민지홍[**], 이하민[**], 김태헌[**], 홍민[*]
[O] [*]순천향대학교 컴퓨터소프트웨어공학과
[**]순천향대학교 소프트웨어융합학과
e-mail: ksh2987625@naver.com[O], kbh0095@gmail.com[*], wlsdh338@sch.ac.kr[*],
tjdrhkwns21@sch.ac.kr[*], tndud5666@sch.ac.kr[*], belgiumb97@sch.ac.kr[*],
wjdduddnd@sch.ac.kr[*], junsangbae12@sch.ac.kr[*], hun398@sch.ac.kr[**],
wlghd5524@gmail.com[**], haminprotoss@sch.ac.kr[**],
taeheon1213@gmail.com[**], mhong@sch.ac.kr[*]

Development of a Unity-based Educational Simulation System for Public-Private Collaborative Response to Nipah Virus

SangHo Kim[O], ByungHoon Kang[*], JinO Kwon[*], MinSung Park[*], SuYeong Jeon[*],
HyeonU Cho[*], YeongWoong Jeong[*], JunSang Bae[*], HyeonWoo Jung[**],
JiHong Min[**], HaMin Lee[**], TaeHeon Kim[**], Min Hong[*]
[O] [*]Dept. of Computer Software Engineering, Soonchunhyang University
[**]Dept. of Software Convergence, Soonchunhyang University

요 약

본 연구는 니파 바이러스 의심 환자 발생 상황에서 민간 의료기관의 대응 절차와 보건소 등 관계기관 연계 과정을 학습할 수 있는 Unity 기반 감염병 교육 시뮬레이션 시스템을 개발하였다. 기존 모의훈련은 단일 기관 중심의 대응 절차에 초점을 두어 관계기관 연계 흐름을 통합적으로 학습하는 데 한계가 있다. 이에 본 연구에서는 개인 보호구 착·탈의, 검체 관리, 밀접접촉자 선별, 환경소독 및 의료폐기물 처리 등 주요 절차를 포함한 시뮬레이션을 구현하였다. 본 시스템은 애니메이션 기반 시나리오와 미니게임 기반 상호작용을 결합하여, 학습자가 전체 대응 흐름을 이해하고 주요 절차를 직접 수행할 수 있도록 설계되었다. 또한 의료 현장 종사자의 자문을 통해 현실성과 교육적 적합성을 검토하였으며, 향후 사용성 평가와 교육 효과를 검증할 예정이다.

키워드: 니파 바이러스, 감염병 교육 시뮬레이션, Linear Interactive Narrative, Unity 엔진, 관계기관 연계

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

I. 서론

신종 감염병의 반복적인 발생은 의료기관 내 감염병 의심 환자 발생 시 신속하고 체계적인 초기 대응의 중요성을 보여주었다[1]. 감염병 위기 상황에서는 환자 선별, 격리, 감염구역 관리, 검체 채취, 개인 보호구 착용과 같은 의료기관 내부 절차가 중요하게 요구된다[2]. 또한 의심 환자 발생 이후에는 보건소 및 관계기관과의 신속한 정보 공유와 협력 대응이 함께 이루어져야 한다. 그러나 기존 감염관리 시뮬레이션 교육은 주로 감염예방 절차와 개별 술기 훈련을 중심으로 활용되어 왔으며[3], 실제 상황에서 요구되는 민간 의료기관과 보건 당국 간 협력 흐름을 통합적으로 학습하는 데에 한계가 있다.

이에 본 연구에서는 니파 바이러스 의심 환자 발생 상황을 기반으로 민간 의료기관의 초기 대응 절차와 보건소 등 관계기관과의 협력 과정을 학습할 수 있는 Unity 엔진 기반 교육 시뮬레이션 시스템을 개발하였다. 제안하는 시스템은 애니메이션과 미니게임이 결합된 Linear Interactive Narrative 형식으로 구성되어 학습자가 전체 대응 흐름을 이해하고, 감염구역 소독, 검체 채취, 개인 보호구 착용 등 주요 절차를 직접 수행할 수 있도록 설계되었다.

II. 관련 연구

2.1 감염병 대응 교육과 니파 바이러스

신종 감염병 대응 교육에서는 의심 환자 발생 시 초기 대응 절차를 정확히 이해하고 수행하는 것이 중요하게 다루어진다. 특히 니파 바이러스와 같은 고위험 감염병은 사람 간 전파 가능성과 높은 치명률로 인해 환자 선별, 격리, 개인 보호구 착용, 검체 채취, 감염구역 관리 등의 절차가 초기 단계에서 요구된다[4].

2.2 감염관리 시뮬레이션 기반 교육

감염관리 교육 분야에서는 실제 감염 위험 없이 반복 학습이 가능한 시뮬레이션 기반 교육이 활용되고 있으며, 감염예방 및 관리 교육에서 실습 능력과 태도 향상에 효과적인 교육 방식으로 다루어지고 있다[5]. 선행 연구에서는 COVID-19 팬데믹 이후 의료 종사자 대상 감염병 예방 교육의 필요성을 제시하고, Unity 엔진 기반 2.5D 시뮬레이션 게임을 통해 병원 내 감염 확산 경로와 관리 전략을 학습할 수 있도록 구현하였다[6]. 이러한 연구는 감염관리 교육에서 시뮬레이션 기반 학습의 활용 가능성을 보여준다.

2.3 Unity 기반 의료 교육 시뮬레이션

Unity 엔진은 2D 및 3D 콘텐츠 개발, 애니메이션, 사용자 상호작용, 이벤트 처리 등을 통합적으로 지원하여 교육용 시뮬레이션 콘텐츠 개발에 활용되고 있다[7]. 또한 가상환경 기반 의료 시뮬레이션은 실제 현장 훈련에서 발생할 수 있는 감염 위험, 공간 및 인력 제약, 반복 훈련의 어려움을 보완하고, 학습자가 필요한 만큼 절차를 반복적으로 경험할 수 있는 교육 환경을 제공할 수 있다[8]. 그러나 감염병 의심 환자 발생 시 의료기관 내부 대응과 보건 당국 간 협력 흐름을 통합적으로 다루는 교육 시뮬레이션 연구는 상대적으로 부족하다. 이에 본 연구에서는 니파 바이러스 의심 환자 발생 상황을 기반으로 의료기관 내부 대응 절차와 관계기관 연계 과정을 함께 학습할 수 있는 Unity 엔진 기반 교육 시뮬레이션 개발에 초점을 둔다.

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

III. 제안방법

3.1 시스템 구조

본 연구에서 개발한 시스템은 니파 바이러스 의심 환자 발생 상황을 중심으로 구성된 Unity 엔진 기반 교육 시뮬레이션이다. 전체 시스템 구조는 그림 1과 같다.

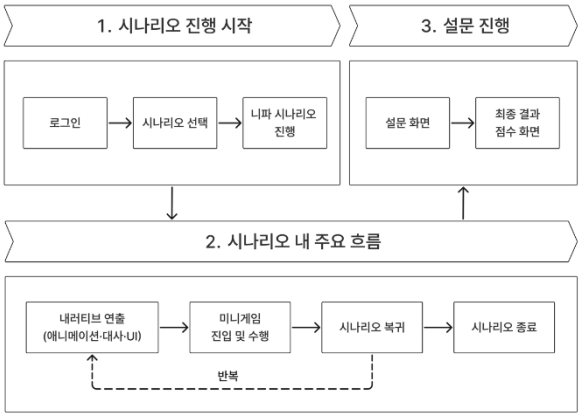


그림 1. 니파 바이러스 시뮬레이션 시스템 구조

시스템은 애니메이션 기반 시나리오와 미니게임 기반 상호작용을 결합한 Linear Interactive Narrative 형식으로 설계하였다. 학습자는 로그인 후 니파 바이러스 시나리오에 진입하며, 애니메이션, 대사, UI를 통해 감염병 대응 상황을 확인한다. 주요 대응 절차가 요구되는 구간에서는 미니게임을 수행하고, 완료 후 다시 시나리오로 복귀한다. 이러한 과정은 시나리오 내 주요 절차마다 반복되며, 전체 시나리오가 종료되면 설문 화면으로 전환된다.

3.2 니파 바이러스 시나리오 및 학습 흐름

니파 바이러스 시나리오는 민간 의료기관에 의심 환자가 내원한 상황에서 시작하여, 감염병 대응 절차와 관계기관 연계 과정을 순차적으로 학습할 수 있도록 구성하였다. 니파 바이러스 시나리오의 전체 학습 흐름은 그림 2와 같다.

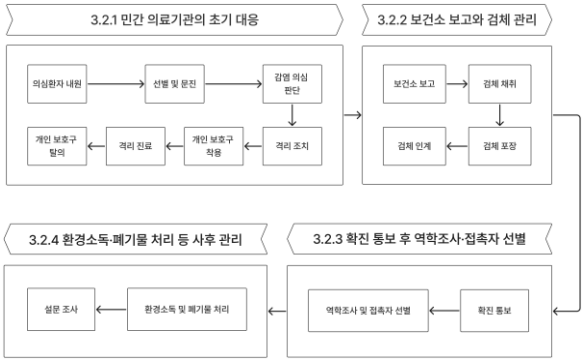


그림 2. 니파 바이러스 시나리오 학습 흐름

그림 2와 같이 시나리오는 의심 환자 내원 이후 초기 대응, 보건소 보고, 검체 관리, 확진 통보, 역학조사, 사후 관리로 이어지는 흐름으로 진행된다. 학습자는 애니메이션과 대사를 통해 상황 전개를 확인하고, 주요 대응 절차에서는 미니게임을 통해 해당 절차를 직접 수행한다. 이를 통해 감염병 의심 환자 발생부터 사후 관리까지의 전체 대응 과정을 단계적으로 이해할 수 있도록 하였다.

3.2.1 민간 의료기관의 초기 대응

환자가 응급실에 내원하면 선별 간호사는 증상과 해외 방문 이력을 확인하여 감염 의심 여부를 판단한다. 이후 환자를 격리구역으로 이동시키고, 의료진은 개인 보호구를 착용한 뒤 격리 진료와 탈의 절차를 수행하도록 구성하였다. 이를 통해 의심 환자 발생 시 의료기관에서 요구되는 선별, 격리, 보호구 착·탈의 등 초기 대응 절차의 순서와 역할을 이해할 수 있다. 그림 3은 민간 의료기관의 초기 대응 과정을 구현한 시뮬레이션 화면의 일부를 나타낸다.

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

그림 3. 민간 의료기관의 초기 대응

3.2.2 보건소 보고와 검체 관리

의료진과 감염관리 담당자는 의심 환자 발생을 보건소에 보고하고, 검체 채취와 포장 절차를 수행한다. 학습자는 시나리오와 미니게임을 통해 검체 채취·포장·인계 과정을 경험하고, 보건소 보고 이후 검체가 관리·전달되는 절차를 확인한다. 이를 통해 의심 환자 발생 시 의료기관과 보건소 간 정보 공유 방식과 검체 관리 흐름을 이해할 수 있다. 그림 4는 보건소 보고와 검체 관리 과정을 구현한 시뮬레이션 화면의 일부를 나타낸다.

그림 4. 보건소 보고와 검체 관리

3.2.3 확진 통보 후 역학조사·접촉자 선별

보건소로부터 확진 통보가 전달되면 의료기관은 확진 상황을 공유하고 추가 전파 가능성을 파악한다. 이후 학습자는 CCTV와 현장 정보를 바탕으로 밀접접촉자 선별 미니게임을 수행하며 접촉자 범위를 확인 및 분류한다. 이를 통해 확진 통보 이후 의료기관과 보건소 간 정보 공유 및 추가 전파 차단을 위한 접촉자 관리 절차를 이해할 수 있다. 그림 5는 확진 통보 이후의 역학조사와 접촉자 선별 과정을 구현한 시뮬레이션 화면의 일부를 나타낸다.

그림 5. 역학조사와 접촉자 선별

3.2.4 환경소독 및 사후 관리

환자 이송 이후에는 사용 공간과 이동 동선에 대한 환경소독, 의료폐기물 처리, 격리구역 정비가 진행된다. 학습자는 미니게임을 통해 오염 가능 구역을 확인하고 의료폐기물을 분류·처리한다. 이를 통해 감염 확산 방지를 위한 환경소독과 의료폐기물 처리 등 사후 관리가 필요함을 이해할 수 있다. 그림 6은 환경소독과 의료폐기물 처리 등 사후 관리 과정을 구현한 시뮬레이션 화면의 일부를 나타낸다.

그림 6. 환경소독과 의료폐기물 처리

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

3.3 미니게임 기반 상호작용 설계

본 시스템은 니파 바이러스 대응 과정에서 수행 순서와 판단이 중요한 절차를 중심으로 미니게임을 배치하였다. 각 미니게임은 안내 화면, 사용자 조작, 정답·오답 피드백, 결과 확인, 시나리오 복귀의 흐름으로 진행되며, 학습자가 주요 대응 절차를 직접 수행하면서 절차 순서, 도구 선택, 오염 관리의 중요성을 반복적으로 확인할 수 있도록 설계하였다.

3.3.1 개인 보호구 착·탈의 미니게임

개인 보호구 착·탈의 미니게임은 격리 진료 전후 단계에 배치하였다. 착용과 탈의 미니게임을 분리하여 학습자가 손위생과 보호구 착·탈의 순서를 단계적으로 수행하도록 설계하였다. 이를 통해 학습자는 보호구 착용과 탈의 과정에서 순서 준수의 중요성을 확인하고, 순서 오류가 오염 확산으로 이어질 수 있음을 이해할 수 있다. 그림 7은 개인 보호구 착·탈의 미니게임의 한 장면이다.

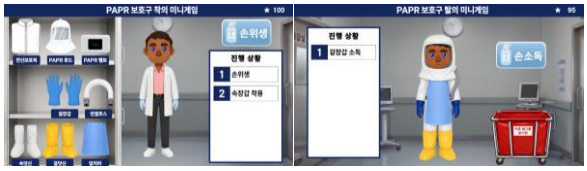


그림 7. 개인 보호구 착·탈의 미니게임

3.3.2 검체 관리 미니게임

검체 관리 미니게임은 보건소 보고 이후 검체 채취와 인계가 이루어지는 단계에 배치하였다. 상기도 검체 채취, 소변 검체 채취, 검체 3중 포장 미니게임을 통해 학습자가 검체 채취·포장·인계 과정을 단계적으로 수행하도록 구성하였다. 이를 통해 학습자는 검체 포장과 인계 과정에서 요구되는 정확성과 안전성의 중요성을 이해할 수 있다. 그림 8은 검체 관리 미니게임의 한 장면이다.

그림 8. 검체 관리 미니게임

3.3.3 역학조사 및 접촉자 선별 미니게임

역학조사 및 접촉자 선별 미니게임은 확진 통보 이후 추가 전파 가능성을 파악하는 단계에 배치하였다. 학습자가 CCTV 및 상황 정보를 바탕으로 밀접접촉자 여부를 판단하도록 설계하였다. 이를 통해 학습자는 밀접접촉자를 신속하게 분류하는 과정이 추가 전파 차단에 중요함을 이해할 수 있다. 그림 9는 밀접접촉자 선별 미니게임의 한 장면이다.

그림 9. 밀접접촉자 선별 미니게임

3.3.4 환경소독·의료폐기물 처리 미니게임

환경소독·의료폐기물 처리 미니게임은 환자 이송 이후 사후 관리 단계에 배치하였다. 학습자는 소독 대상 구역을 확인하고 의료폐기물을 분류하여 적절한 처리 용기를 선택하도록 구성하였다. 이를 통해 학습자는 오염 가능 구역 관리와 의료폐기물 처리가 사후 감염 확산 방지에 필요함을 이해할 수 있다. 그림 10은 환경소독·의료폐기물 처리

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

미니게임의 한 장면이다.

그림 10. 환경소독·의료폐기물 처리 미니게임

또한 의료 현장 종사자의 자문을 반영하여 개인 보호구 착·탈의, 검체 관리, 접촉자 선별, 환경소독 및 의료폐기물 처리 과정이 실제 현장 대응 흐름과 부합하는지 검토하였다. 이를 바탕으로 각 미니게임은 절차 순서, 도구 선택, 오염 확산 방지, 판단 기준을 학습할 수 있도록 구성하였다.

IV. 결론

본 연구에서는 니파 바이러스 의심 환자 발생 상황을 기반으로 민간 의료기관의 초기 대응 절차와 보건소 등 관계기관 연계 과정을 학습할 수 있는 Unity 기반 교육 시뮬레이션 시스템을 개발하였다. 제안 시스템은 애니메이션 기반 시나리오와 미니게임 기반 상호작용을 결합하여, 학습자가 감염병 대응 흐름을 순차적으로 이해하고 주요 절차를 직접 수행할 수 있도록 구성하였다. 이를 통해 학습자는 절차의 순서와 수행 목적을 상호작용으로 확인할 수 있으며, 민간 의료기관 내부 대응부터 보건소 보고, 역학조사, 접촉자 관리로 이어지는 관계기관 연계 흐름을 함께 학습할 수 있다.

다만 본 연구는 시스템 개발과 시나리오 구현을 중심으로 수행되었으며, 실제 의료 현장 종사자를 대상으로 한 사용성 평가와 교육 효과 검증은 아직 이루어지지 않았다는 한계가 있다. 향후 연구에서는 의료 현장 종사자를 대상으로 사용성과 교육 효과를 검증하고, 그 결과를 바탕으로 시나리오와 미니게임 구성을 개선할 예정이다. 또한 다양한 감염병 대응 시나리오로 확장하여 민·관 협력 기반 교육 시스템으로서의 활용 가능성을 높이고자 한다.

참고문헌

[1] June KJ, Choi E. Infection Control of Hospital Nurses: Cases of Middle East Respiratory Syndrome. Korean J Occup Health Nurs. 2016;25(1):1-8.

[2] Alhumaid S, Al Mutair A, Al Alawi Z, et al. Knowledge of infection prevention and control among healthcare workers and factors influencing compliance: A systematic review. Antimicrob Resist Infect Control. 2021;10:86.

[3] Kang M, Nagaraj D, Campbell J, et al. The role of simulation-based training in healthcare-associated infection prevention. Antimicrob Steward Healthc Epidemiol. 2022;2(1):e97.

[4] Aditi, Shariff M. Nipah virus infection: A review. Epidemiol Infect. 2019;147:e95.

[5] Yoshikawa A, Ohtsuka H, Aoki K, et al. Simulation-based infection prevention and control training for medical and healthcare students: A systematic review. Front Med. 2025;12:1529557.

[6] 이하민, 마준, 김태헌, 민지홍, 하정혜, 권진오, 홍민. 모바일 기반 병원 내 감염 통제 시뮬레이션 게임 개발: 의료 종사자 교육을 위한 확장 연구. 2025 한국데이터사이언스학회 춘계종합학술대회 논문집; 2025.

[7] 윤석현. 교육적 관점에서 본 Unity 3D 엔진. 한국컴퓨터정보학회 학술발표논문집. 2015;23(2):246-249.

[8] Gupta S, Wilcocks K, Matava C, et al. Creating a Successful Virtual Reality-Based Medical Simulation Environment: Tutorial. JMIR Med Educ. 2023;9:e41090.

※ "본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

흉부 X-ray 기반 다중 병변 분류 모델의 Grad-CAM 기반 약지도 병변 위치 추정 연구

박현진*, 진수빈*, 전예은*, 류찬*, 마준**, 성낙준*
* 한신대학교 AI·SW대학, ** 국립군산대학교 소프트웨어학과
e-mail:{ guswls, vheh5053, jjh1221jjh, lyuchan, njsung}@hs.ac.kr, junma@kunsan.ac.kr

A Study on Grad-CAM-Based Weakly Supervised Lesion Localization for a Multi-Lesion Classification Model Using Chest X-rays

Yeeun Jun*, Su Been Jin*, Hyeon Jin Park*, Jun Ma**, Nak-Jun Sung*
*School of Computing and Artificial Intelligence, HanShin University
** Department of Software Science and Engineering, Kunsan National University

요 약

본 연구는 XRAYDAR-Multimodal 흉부 X-ray 데이터셋을 이용하여 영상 수준 합의 라벨만으로 병변 위치 단서를 추정하는 약지도 위치 추정 연구를 수행하였다. ImageNet-1K로 사전학습된 ResNet50을 백본으로 사용하여 전역 평균 풀링 기반 멀티라벨 분류 헤드를 결합하고, 학습된 분류기의 Grad-CAM 활성 지도를 이진화하여 예측 병변 박스를 생성하였다. 분류 성능은 macro-F1과 mean Average Precision(mAP)으로, 위치추정 성능은 예측 CAM 박스와 데이터셋에 포함된 방사선 전문의 bounding box 간 mean IoU 및 IoU 0.1 초과 hit rate로 평가하였다. 실험 결과 validation set 기준 최고 mAP 0.5795, test set 51개 기준 macro-F1 0.2901, mAP 0.3704, hit rate@0.1 0.3529, mean IoU 0.1003을 보였다. 실험 결과 제한된 데이터에서도 질환 공존 패턴을 일부 학습하였지만, 단순 Grad-CAM 박스화만으로는 병변의 정밀 위치를 안정적으로 복원하기 어렵다는 점을 확인하였다.

키워드: 흉부 X-Ray, 폐 질환 분류, 약지도학습, eXplainable AI, Grad-CAM

I. 서론

흉부 X-ray 인공지능 모델은 단순 분류 정확도뿐 아니라 모델이 어떤 영상 근거에 기반해 판단했는지를 함께 제시해야 임상적 신뢰성을 확보할 수 있다. 그러나 병변 위치 주석은 영상 수준 라벨보다 수집 비용이 높고 판독자 간 차이도 크다. 이에 따라 이미지 수준 라벨만으로 학습한 분류 모델에서 사후 설명 기법을 이용해 병변 위치를 추정하는 약

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원 사업(경기RISE사업)의 연구결과 입니다.

지도 위치추정은 실용적인 연구 방향이 된다.

본 연구의 목적은 XRAYDAR-Multimodal 데이터셋에서 멀티라벨 분류와 Grad-CAM 기반 위치추정을 결합한 재현 가능한 기준선을 만들고, 분류 성능과 위치 근거 품질 사이의 간극을 정량적으로 분석하는 것이다. 특히 cardiomegaly, pleural effusion, consolidation을 모두 포함하는 복합 소견 샘플을 중심으로 학습하여, 실제 흉부 X-ray에서 흔한 다중 병변 공존 상황에서 CAM 기반 설명이 얼마나 유효한지 평가하였다.

II. 데이터

본 연구에서 활용한 XRAYDAR-Multimodal 데이터셋은 979개의 흉부 X-ray 검사로 구성되며, 각 샘플은 PNG 이미지, 영상 수준의 consensus label, 판독자별 bounding box, 원문 판독문, 보고서를 포함한다. 본 논문에서는 학습을 위해 consensus label 정보와 bounding box 정보를 사용하였다.

전체 979개 샘플에서 가장 빈번한 영상 레이블은 cardiomegaly(632건, 64.56%), interstitial shadowing(631건, 64.45%), object(562건, 57.41%), pleural effusion(548건, 55.98%), consolidation (501건, 51.17%)이었다.

본 연구에서는 가장 빈번하게 등장하며, 실제 병변에 해당하는 3가지 레이블(cardiomegaly, pleural effusion, consolidation)을 모두 포함하는 샘플만 사용하였으며, 이 조건을 만족하는 샘플은 252개였다. 샘플링한 252개의 데이터는 7:1:2 비율로 train(176개) set, validation (25개) set, test (51개) set으로 분할된다. 모든 데이터는 환자 단위로 분할되었으며, 각 서브셋은 데이터 누출이 없는 독립적 데이터셋으로 구성되었다.

III. 연구 방법

3.1 멀티 레이블 분류 모델

본 연구에서는 ImageNet-1K로 사전학습된 ResNet50을 기본 백본으로 사용하였다. 입력 흉부 X-ray 영상은 RGB 형식으로 변환한 뒤 512 x 512 크기로 조정하고, ImageNet 평균과 표준편차를 이용하여 정규화하였다.

제안 모델은 ResNet50의 마지막 convolution feature map에 adaptive average pooling을 적용한 뒤, 선형 분류기를 통해 각 레이블에 대한 logit을 출력하도록 구성하였다. 학습 대상 레이블은 전체 consensus label 중 최소 20회 이상 등장한 레이블로 제한하였다. 멀티라벨 분류 문제이므로 손실 함수는 BCEWithLogitsLoss를 사용하였고, optimizer는 AdamW를 사용하였다. 기본 학습률은 1e-4, batch size는 8, 학습 epoch는 100으로 설정하였다.

3.2 Grad-CAM 기반 위치 추정

병변 위치 추정은 별도의 detection 모델을 학습하지 않고, 학습된 분류 모델의 Grad-CAM을 이용하여 수행하였다. 각 영상에서 sigmoid 확률이 가장 높은 레이블을 예측 레이블로 선택하고, 해당 class logit에 대한 gradient를 마지막 convolution block에 역전파하여 class activation map을 생성하였다. 생성된 CAM은 0과 1 사이의 값으로 정규화하였다. 이후 threshold 0.5 이상인 영역을 활성 영역으로 정의하고, 해당 영역을 포함하는 최소 외접 사각형을 예측 bounding box로 변환하였다. CAM이 threshold를 만족하는 영역을 만들지 못한 경우에는 최소 크기의 기본 박스를 사용하였다.

3.3 정답 박스와의 비교

위치 평가는 예측 레이블과 동일한 레이블을 가진 전문의 bounding box가 존재하는 샘플에 대해서

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원 사업(경기RISE사업)의 연구결과 입니다.

만 수행하였다. 동일 레이블의 정답 박스가 여러 개인 경우에는 예측 박스와의 IoU가 가장 높은 값을 해당 샘플의 위치 추정 성능으로 사용하였다.

분류 성능은 threshold 0.5 기반 macro-F1과 class-wise average precision의 평균인 mAP로 평가하였다. 위치 추정 성능은 mean IoU와 hit rate@0.1로 평가하였다. hit rate@0.1은 예측 박스와 정답 박스의 IoU가 0.1을 초과한 샘플의 비율로 정의하였다. Best checkpoint는 학습 과정에서 validation mAP가 가장 높은 epoch을 기준으로 선택하였다.

IV. 실험

II-데이터 섹션에서 소개한 바와 같이 실험 또한 cardiomegaly, pleural effusion, consolidation을 모두 포함하는 252개 샘플을 먼저 구성한 뒤, 동일 seed로 분할된 test set 51개에서 수행하였다. Test set의 모든 샘플에는 예측 레이블과 비교 가능한 정답 bounding box가 존재하여 localization 평가에도 51개 샘플이 모두 포함되었다.

Test set에서 classification macro-F1은 0.2901, mAP는 0.3704로 측정되었다. Localization hit rate@0.1은 0.3529, mean IoU는 0.1003로 확인됐다. 이는 모델이 영상 수준의 다중 레이블 분류 신호는 일부 학습하였으나, Grad-CAM을 단순 박스 형태로 변환했을 때 정답 병변 위치와의 공간적 일치도는 여전히 낮았음을 의미한다.

표1. 실험 결과

Type	Metrics	Score
Classification	Macro-F1	0.2901
	mAP	0.3704
Localization	Hit @0.1	0.3529
	Mean IoU	0.1003

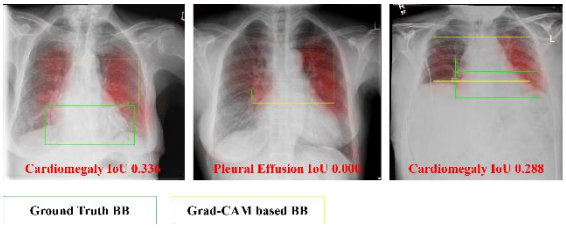


그림1. 실험 결과의 시각화. (좌) Cardiomegaly IoU 0.336, (중) Pleural Effusion IoU 0.000, (우) Cardiomegaly IoU 0.288.

V. 논의

본 연구의 실험 결과는 흉부 X-ray 멀티라벨 분류 성능과 Grad-CAM 기반 위치추정 성능이 서로 다른 평가 목표임을 보여준다. ResNet50 기반 분류 모델은 test set에서 macro-F1 0.2901, mAP 0.3704를 보였으나, 동일 모델의 Grad-CAM 박스를 전문의 bounding box와 비교했을 때 mean IoU는 0.1003에 그쳤다. 이는 영상 수준 라벨만으로 학습한 모델이 병변과 관련된 일부 판별 단서는 포착할 수 있지만, 그 단서가 병변의 전체 공간 범위를 정확히 반영하지는 않음을 의미한다 [1].

이러한 결과는 Grad-CAM 방법의 특성과 흉부 X-ray 병변의 특성을 함께 고려하여 해석해야 한다. Grad-CAM은 분류 판단에 기여한 영역을 사후적으로 시각화하는 방법이며, 병변 경계를 직접 학습하는 검출 모델이 아니다 [2]. 따라서 활성 영역을 threshold 0.5로 이진화한 뒤 최소 외접 사각형으로 변환하는 방식은 병변의 실제 범위를 과소 또는 과대 추정할 수 있다. 특히 pleural effusion과 consolidation처럼 경계가 불명확하거나 넓게 분포하는 소견에서는 CAM이 일부 강한 영상 단서에만 집중할 가능성이 있다. 또한 cardiomegaly는 국소 병변이라기보다 심장 크기와 흉곽 구조의 전역적 관계로 판단되는 소견이므로, 단일 bounding box 기반 IoU 평가가 실제 설명 품질을 충분히 반영하

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원 사업(경기RISE사업)의 연구결과 입니다.

지 못할 수 있다 [3].

본 연구의 데이터 구성 역시 결과 해석에 영향을 준다. 실험은 cardiomegaly, pleural effusion, consolidation을 모두 포함하는 252개 샘플만을 대상으로 수행되었으며, test set은 이 중 51개로 구성되었다. 이 설정은 복합 병변이 공존하는 상황에서 CAM 기반 위치 단서를 분석하기에는 적합하지만, 모든 샘플이 세 핵심 레이블을 포함한다는 점에서 일반적인 흉부 X-ray 질환 분류 환경과는 차이가 있다. 따라서 본 결과는 XRAYDAR-Multimodal 전체 분포에 대한 일반화 성능이라기보다, 복합 병변 조건에서의 약지도 위치추정 기준선으로 해석하는 것이 적절하다.

단순 Grad-CAM thresholding은 추가 위치 주석 없이 병변 위치 단서를 탐색할 수 있다는 장점이 있지만, 전문의 수준의 병변 위치를 안정적으로 대체하기에는 한계가 있다. 향후 연구에서는 분류 성능만을 기준으로 checkpoint를 선택하는 방식에서 벗어나 localization-aware 모델 선택 기준을 적용할 필요가 있다 [4]. 또한 병변별 CAM threshold 조정, Grad-CAM++ 및 Score-CAM과 같은 대체 CAM 방법 비교, annotator 간 bounding box 차이를 반영한 평가를 통해 약지도 위치추정의 신뢰성을 더 체계적으로 검증해야 한다.

VI. 결론

본 연구에서는 XRAYDAR-Multimodal 데이터셋을 이용하여 흉부 X-ray 멀티라벨 분류 모델의 Grad-CAM 기반 약지도 병변 위치 추정 연구를 수행하였다. 실험 결과 영상 수준 라벨만으로 학습한 분류 모델이 병변과 관련된 일부 영상 단서를 학습할 수는 있지만, 단순 Grad-CAM thresholding만으로 전문의 bounding box 수준의 병변 위치를 안정적으로 복원하기는 어렵다는 점을 확인하였다. 따라서 흉부 X-ray 설명가능성 연구에서는 분류 성능뿐 아니라 위치 설명의 품질을 별도로 평가해야 하며, 이러한 초기 연구 결과를 바탕으로 추후에는 localization-aware checkpoint selection, 병변별 CAM threshold 조정, 다양한 CAM 방법 비교, annotator-aware 평가 연구를 수행할 계획이다.

참고문헌

[1] Choe, J., Oh, S. J., Lee, S., Chun, S., Akata, Z., & Shim, H. (2020). Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3130-3139)

[2] Keshavamurthy, K. N., Eickhoff, C., & Juluru, K. (2021). Weakly supervised pneumonia localization in chest X-rays using generative adversarial networks. *Medical Physics, 48*(11), 7154-7171

[2] Fan, W., Yang, Y., Qi, J., Zhang, Q., Liao, C., Wen, L., Wang, S., Wang, G., Xia, Y., Wu, Q., Fan, X., Chen, X., He, M., Xiao, J., Yang, L., Liu, Y., Chen, J., Wang, B., Zhang, L., ... Zhang, D. (2024). A deep-learning-based framework for identifying and localizing multiple abnormalities and assessing cardiomegaly in chest X-ray. *Nature Communications, 15*(1), 1347

[4] Han, Y., Holste, G., Ding, Y., Tewfik, A., Peng, Y., & Wang, Z. (2023). Radiomics-guided global-local transformer for weakly supervised pathology localization in chest X-rays. *IEEE Transactions on Medical Imaging, 42*(3), 750-761

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

개발자를 루프 안에 유지하기: 소프트웨어 개발에서의 과업 의존적 AI 위임

Shi Bowen[1o], 김장현 [1,2]
성균관대학교 인터랙션사이언스학과 [1] / 인간 AI 인터랙션융합전공 [2]
e-mail: bowenshi@skku.edu, alohakim@skku.edu

Keeping Developers in the Loop: Task-Contingent AI Delegation in Software Development

Shi Bowen[1o], Jang Hyun Kim[1,2]
Dept. of Interaction Science[1] / Human-Artificial Intelligence Interaction[2], Sungkyunkwan University

요 약

생성형 AI 가 소프트웨어 개발에 점점 더 깊이 통합되고 있으나, 개발자는 모든 과업에서 AI 에게 동일한 자율성을 부여하지 않는다. 본 연구는 AI 사용을 단일한 채택 결과가 아니라 과업 수준의 인간-AI 통제 배분(task-level human–AI control allocation)으로 본다. 2025 년 Stack Overflow 개발자 설문을 사용해 개발자 30,494 명, 13 개 과업에서 383,761 개의 개발자-과업 관측치를 분석하였으며, 다항 로짓 모형으로 AI 무사용, AI 보조적 사용, AI 주도적 사용을 구분하였다. 가장 두드러진 패턴은 과업 의존성이다. AI 위임의 변량 대부분은 개발자 간이 아니라 개발자 내 과업 간에 존재한다. 맥락의존성과 결과 노출도(consequence exposure)는 위임을, 특히 AI 주도적 위임을 크게 억제한다. 검증용이성은 보다 구체적인 비대칭을 만든다. 즉, 여러 대안적 과업 코딩에 걸쳐 계획된 AI 주도적 위임보다 계획된 AI 보조적 위임과 일관되게 더 강하게 연관되며, 이는 쉬운 검증이 루프 안에 머무르는 것의 옵션가치를 높임을 시사한다. 검증가능성은 인간 감독을 불필요하게 만들기보다 오히려 더 가치 있게 만들어, 개발자를 루프 안에 유지한다.

키워드: 인간-AI 상호작용, 생성형 AI, AI 위임, 기술수용, 소프트웨어 개발

I. 서 론

생성형 AI 코딩 도구가 실험을 넘어 일상적 소프트웨어 업무로 이동하면서, 조직은 2 차적 질문에 직면하고 있다. 단순히 AI 를 사용할 수 있는지가 아니라, 어디에서 더 깊은 AI 개입을 허용할 가치가 있는가이다. 최근 보도는 기업들이 빠른 내부 도입 이후 값비싼 에이전트형 코딩 도구의 접근을 재검토하기 시작했음을 보여주며[1], 이는 관심을 AI 사용량에서 특정 AI 지원 작업의 가치로 이동시킨다. 이러한 조직적 전환은 더 근본적인 인간-AI 상호작용 문제를 부각한다. 개발자는 AI 를 사용할지 여부만이 아니라, 과업마다 얼마만큼의 통제권을 보유할지를 결정한다. 따라서 우리는 AI 사용을 과업 수준의 인간-AI 통제 배분으로 보고, AI 보조적(assisted) 위임과 AI 주도적(dominant) 위임을 구분한다.

생성형 AI 는 단일하게 채택하거나 거부하는 기술이 아니라, 개발자가 서로 다른 수준의 자율성을 부여할 수 있는 유연한 행위자로 소프트웨어 개발에 들어왔다. 이는 단순한 채택 문제가 아니라 위임 문제를 만든다. 개발자는 답 검색·코드 초안·디버깅에는 AI 를 허용하면서도 배포 모니터링·기획·분석에는 AI 의 개입을 거부할 수 있다. 동일한 개발자가 열성적 AI 사용자이면서 동시에 신중한 위임자일 수 있다. 따라서 핵심 질문은 개발자가 AI 를 사용하는지가 아니라, 과업에 걸쳐 통제권을 어떻게 배분하는지가 된다.

자동화 의존(automation reliance)에 관한 기존 이론은 직관적 예측을 제시한다. 시스템 출력을 관찰·평가·검증하기 쉬울수록 사용자는 자동화에 더 기꺼이 의존한다는 것이다[2,3]. 검증은 불확실성을 줄이고 신뢰 보정을 개선하며 오류 탐지를 용이하게 한다[4]. 이 논리를 소프트웨어 개발의 생성형 AI 에 적용하면 검증이 쉬운 과업일수록 AI 에 더 위임 가능하다는 결론에 이른다. 본 연구는 AI 사용을 과업 수준의 자율성 배분으로 다룰 때 이 통상적 예측이 불완전해진다고 주장한다.

핵심 논지는 다음과 같다. AI 출력을 더 안전하게 사용하게 만드는 바로 그 검증 용이성이, 인간 감독을 보존할 가치 또한 높일 수 있다. 검증이 어려울 때는 인간이 루프 안에 남아 있어도 오류 탐지가 어려워 추가적 통제 이득이 작다. 반대로 출력이 검증하기 쉬울 때 AI 보조적 위임은 특히 매력적이 된다. 개발자가 AI 생성 도움을 누리면서도 그 출력이 올바른지에 대한 실질적 판단을 계속 행사할 수 있기 때문이다. 그 결과 검증 용이성은 AI 주도적 위임보다 AI 보조적 위임과 더 강하게 연관되며, 우리는 이를 인간 감독의 옵션가치(option value of human oversight)로 설명한다. 이 비대칭은 채택 이력·도구 가용성·조직 정책에 의해 형성되는 현재 사용보다, 선택이 여전히 열려 있는 계획된 의도(prospective intent)에서 더 뚜렷하게 드러날 것이다. 본 연구는 (1) AI 사용 연구의 대상을 채택에서 자율성 배분으로, (2) 분석 단위를 사용자에서 사용자-과업 쌍으로 이동시키며, (3) 검증용이성의 'AI 보조적 위임 우위 비대칭(assisted-over-dominant asymmetry)'을

감독의 옵션가치로 이론화한다. 분석에는 2025 년 Stack Overflow 개발자 설문(개발자 30,494 명, 13 개 과업, 383,761 개 개발자-과업 관측치)을 사용한다.

II. 이론적 배경 및 가설

2.1 채택에서 과업 수준 자율성 배분으로

기술수용모형(TAM)과 통합기술수용이론(UTAUT)은 채택을 사용자 수준의 결과로 다룬다[5,6]. 과업-기술 적합(TTF)은 과업과 기술의 적합이 성과를 형성함을 인정하지만, 사용자가 과업에 걸쳐 AI 에 서로 다른 수준의 자율성을 어떻게 배분하는지는 규정하지 않는다[7]. 범용 AI 에서는 이 한계가 분명해진다. 동일한 개발자가 코드 검색에는 AI 를 광범위하게, 디버깅에는 부분적으로, 배포에는 전혀 쓰지 않을 수 있다. Russo[8]는 생성형 AI 채택이 일반적 유용성 신념보다 기존 작업 흐름과의 양립성에 의해 형성됨을, Zacharias 등[9]은 개발자가 생성형 AI 를 생산성·창의성의 원천이자 과의존·역량 저하·정체성 긴장의 원천으로 동시에 경험함을 보인다. AI 사용이 안정적 사용자 성향이라면 빈번 사용자는 과업 전반에서 사용하겠지만, 자율성 배분 문제라면 동일 개발자가 과업에 걸쳐 크게 달라질 것이다.

가설 1. AI 위임은 사용자 의존적이기보다 과업 의존적이며, 위임의 변량은 개발자 간보다 개발자 내(과업 간)에서 더 클 것이다.

2.2 비대칭적 위임 상태로서의 보조적·주도적 위임

기존 연구는 더 많은 AI 개입을 단일 연속선상의 이동으로 다루는 경향이 있다. 그러나 AI 보조적 위임과 AI 주도적 위임은 정도뿐 아니라 구조에서 다르다. 보조적 위임은 인간의 개입 권한을 보존한다. 개발자는 AI 출력을 검토·수정·거부·통합할 수 있고 검증과 최종 책임에 능동적으로 관여한다. 주도적 위임은 과업 과정의 더 많은 부분을 AI 에 이양하고 오류 발견·수정·통합에서 인간의 역할을 축소한다. 차이는 'AI 를 덜 쓰는가 더 쓰는가'가 아니라 인간-AI 통제의 구조이다[10]. 소프트웨어 개발에서는 많은 AI 출력이 그럴듯하지만 옳지 않다. 컴파일되지만 경계 사례에서 실패하는 코드, 일관되게 들리지만 원인을 잘못 짚는 디버깅 설명이 그 예다. 따라서 동일한 과업 속성이 두 위임 상태에 서로 다르게 작용할 수 있으며, 이를 가시화하려면 다항 모형이 필요하다.

2.3 과업 속성과 비대칭적 억제

우리는 전문 업무를 AI 에 위임할 때의 세 가지 통제 제약을 종합한다. 출력 평가가능성(검증용이성), 결과 노출도(consequence exposure), 상황적 지식 적합(맥락의존성)이다. 이 세 차원이 모든 과업 특성을 망라한다고 주장하지는 않는다. 알고리즘 의존은 과업 호의성에 단조적이지 않으며, 알고리즘 회피와 알고리즘 선호가 맥락에 따라 갈린다[11,12].

결과 노출도는 오류가 전문적·조직적·법적·평판적 결과를 낳는 정도를 가리킨다. 결과가 클수록 위임은 줄지만, 그 억제는 주도적 위임에서 더 강할 것이다. 보조적 위임은 검토와 개입을 보존해 문제 출력을 수정·거부할 수 있는 반면, 주도적 위임은 그 통제를 축소하기 때문이다.

가설 2a. 결과 노출도는 보조적·주도적 위임 모두와 부적으로 연관되며, 주도적 위임에서 더 강한 부적 연관을 보일 것이다.

맥락의존성은 과업 수행이 코드베이스·시스템·팀·조직 특유의 지식을 요구하는 정도를 가리킨다. 보조적 위임에서는 개발자가 국지적 지식으로 출력을 보완할 수 있으나, 주도적 위임에서는 시스템이 지속적 인간 통합 없이 더 많은 과정을 처리하도록 기대된다. 따라서 그 억제 역시 주도적 위임에서 더 강할 것이다.

가설 2b. 맥락의존성은 보조적·주도적 위임 모두와 부적으로 연관되며, 주도적 위임에서 더 강한 부적 연관을 보일 것이다.

2.4 감독의 옵션가치

검증용이성은 결과 노출도·맥락의존성과 이론적으로 다르다. 후자는 개발자가 왜 더 많은 자율성을 거부하는지를 설명하나, 검증용이성은 개발자가 왜 주도적 위임보다 보조적 위임을 선호하는지를 설명한다. 검증용이성에는 두 효과가 있다. 첫째는 위험 감소 효과로, 검증이 쉬우면 AI 사용이 더 안전해진다. 둘째는 옵션가치 효과로, 검증이 쉬우면 인간 개입을 보존할 가치가 커진다. 실물옵션 이론에서 옵션은 더 나은 정보 아래 나중에 행동할 권리를 보존할 때 가치가 있다[13,14]. AI 보조 작업에서의 인간 감독이 바로 이 구조를 갖는다. 개발자는 출력을 검토·수정·거부·통합하고 결과의 수용 여부를 결정할 권리를 보존한다. 이 옵션은 감독이 유용한 신호를 만들 때에만 가치가 있고, 검증용이성은 감독의 정보성을 높인다[2]. 보조적 위임은 두 효과 모두에서 이득을 보지만, 주도적 위임은 개입의 내장된 옵션 상당 부분을 포기한다. 적정 의존(appropriate reliance) 관점에서도 옳은 조언을 수용하고 틀린 조언을 거부하는 능력은 검증 행위가 실제로 일어날 때에만 작동하는데, 주도적 위임은 그 행위를 봉쇄한다[4]. 따라서 검증용이성은 AI 주도적 위임보다 AI 보조적 위임과 더 강하게 연관될 것이다.

가설 3. 계획된 AI 위임에서 검증용이성은 계획된 AI 주도적 위임보다 계획된 AI 보조적 위임과 더 강하게 연관될 것이다. 쉬운 검증이 인간 감독을 보존할 가치를 높이기 때문이다.

가설 4(시간 프레임 비교). 검증용이성이 계획된 AI 주도적 위임보다 계획된 AI 보조적 위임과 더 강하게 연관되는 'AI 보조적 위임 우위 비대칭'은 현재 사용보다 계획된 의도 응답에서 더 뚜렷할 것이다. 이는 강한 인과 주장이 아니라 시간 프레임 비교이다. 계획된 의도는

아직 AI 에 위임되지 않아 숙고적으로 열려 있는 응답자-과업 쌍에서 보고되며, 현재 사용은 채택 이력·도구 가용성·조직 제약을 추가로 반영한다.

III. 연구방법

3.1 자료와 표본

우리는 2025 년 Stack Overflow 연례 개발자 설문을 분석한다. 전체 설문은 2025 년 5–6 월에 177 개국에서 49,196 명의 응답을 받았다. 필수 응답자 수준 공변량에 유효 응답을 보이고 AI 워크플로 문항에 최소 한 개 이상 응답한 개발자로 표본을 제한하면 30,494 명이 된다. 장형 재구조화는 모형 A(현재 위임)용으로 383,761 개의 완전한 개발자-과업 관측치를 산출한다. 계획된 사용 응답은 현재 AI 사용으로 분류되지 않은 응답자-과업 쌍에서만 관측되므로, 모형 B 는 29,385 명에서 도출된 265,025 개 관측치의 조건부 부분표본에서 추정된다. 개발자는 평균적으로 13 개 중 12.6 개 과업에 응답했고 과업 수준 무응답은 3% 미만이다. Stack Overflow 표본은 공개 기술 커뮤니티에 참여하는 자기선택 표본이므로 전체 전문 개발자 모집단에 대한 대표성을 주장하지 않으며, 모형 B 는 모집단 모형이 아니라 조건부 계획-의도 모형으로 해석한다.

3.2 변수

종속변수. AI 워크플로 모듈은 각 과업에 대해 다섯 가지 상호배타적 위임 상태(현재 대부분 AI, 현재 부분 AI, 대부분 AI 사용 계획, 부분 AI 사용 계획, 사용 계획 없음)를 선택하게 한다. 이를 두 개의 3 수준 결과로 구성한다. 현재 위임은 무사용=0, 부분(보조)=1, 대부분(주도)=2; 계획 위임도 동일하다. 부분 AI 를 보조적, 대부분 AI 를 주도적으로 명명한다(설문 범주와 구성개념의 구분). 이진으로 축약하면 보조-주도의 핵심 변량이 사라지므로 3 수준 범주 구조를 보존한다.

독립변수. 13 개 과업을 세 차원에서 3 점 척도(1=낮음, 3=높음)로 코딩한다. 두 코더(저자 1 인과 가설·구성개념에 눈가림된 독립 평정자)가 사전 루브릭으로 39 개 판단을 독립 평정하였다. 이차가중 Cohen 카파는 결과 노출도 0.83, 맥락의존성 0.93 에서 almost perfect, 검증용이성 0.70(13 개 중 9 개 정확 일치, 69.2%)에서 substantial 이었다. 전체 39 개 판단의 정확 일치는 79.5%(31/39), Cohen κ=0.82 였다. 모든 불일치는 1 점 차이였고 토론으로 조정되었으며, 검증용이성은 완전한 검증이 아니라 일상적 작업 흐름에서의 작업자-지각 평가가능성으로 해석한다. 검증용이성은 세 차원 중 가장 해석 민감도가 높아, 대안적 검증 코딩에서 보조 우위 비대칭이 유지되는지를 보충자료에서 추가 검토한다. 합의 코딩에서 세 속성의 상관은 |r|=0.35(검증-노출), 0.44(검증-맥락), 0.55(노출-맥락)이며 분산팽창계수는 1.20, 1.49, 1.51 로 임계치 이하이다. 개인 수준 통제로 AI 정확성 신뢰(5 점, −2~+2 평균중심화), 연령, 전문 코딩 경력, 개발자 유형(5 범주), Parasuraman 단계 고정효과를 포함한다[17].

3.3 모형

주 분석은 무사용을 기준 범주로 하는 다항 로짓이다[18]. 표준오차는 응답자 수준에서 군집화한다[19]. 두 결과가 개념적으로 구별되므로 시간 프레임을 통합하지 않고 모형 A(현재)와 모형 B(계획)를 따로 추정한다. 이진 모형은 보조-주도 구분을 붕괴시키고, 비례 오즈 순서형 모형은 동일 예측변수가 보조-주도 경계에서 같은 방향으로 작용한다고 가정하므로, 핵심 결과를 보이려면 다항 사양이 필요하다.

IV. 분석결과

4.1 변량분해: 분석단위로서의 개발자-과업 쌍

가설 1 을 검증하기 위해 30,494 명×13 과업 설계에서 현재 AI 사용의 변량을 분해한다. 13 개 중 8 개 이상에 응답한 개발자(n=29,949)에서 개발자 내 성분(약 62%)은 개발자 간 성분(약 38%)보다 실질적으로 크다. 과업 간 변량은 전체의 약 10%이며 그중 세 과업 속성이 약 68%를 설명한다. 계획된 사용에서는 개발자 내 약 56%, 개발자 간 약 44%, 과업 간 약 5%(속성 설명 약 73%)이다. AI 위임을 사용자 수준 결과로 다루면 체계적 변량의 약 60%를 버리게 된다. 개발자-과업 쌍이 적절한 분석 단위이다(표 1).

표 1. 개발자-과업 관측치의 AI 위임 변량분해

변량 출처	현재 사용	계획된 사용
개발자 내(within)	62%	56%
개발자 간(between)	38%	44%
과업 간 변량 비중	10%	5%
ㄴ과업속성 설명 비중	68%	73%

주: 13 개 과업 이진 위임 지표 행렬에 대한 분해(8 개 이상 응답: 현재 n=29,949). 반올림으로 합이 100%가 아닐 수 있음. 마지막 행은 과업 간 변량 중 속성 설명 비중.

4.2 주효과: 비대칭적 억제와 보조 우위 비대칭

표 2 는 모형 A·B 의 다항 로짓 추정치를 보고한다. 결과 노출도와 맥락의존성은 두 위임 상태 모두를 억제하되 주도적 위임을 더 강하게 억제한다. 모형 A 에서 결과 노출도는 보조 −0.296, 주도 −0.511(비율 1.73), 맥락의존성은 보조 −0.395, 주도 −0.535(비율 1.35)이다. 모형 B 에서 결과 노출도는 −0.418/−0.699(비율 1.67), 맥락의존성은 −0.158/−0.228(비율 1.44)이다. 네 비율 모두 1.35–1.73 구간에 모여 주도적 위임이 더 강하게 억제됨을 보인다. 따라서 가설 2a·2b 가 지지된다.

검증용이성의 패턴은 다르다. 모형 A 에서는 보조 +0.217, 주도 +0.195 로 둘 다 정적이며 크기가 거의 같다. 그러나 모형 B 에서는 두 상태가 서로 다르게 반응한다. 검증용이성은 계획된 보조적 위임과 +0.145(p<.001)로 정적으로 연관되는 반면, 합의 코딩에서 계획된 주도적 위임과는 −0.040(p<.001)으로 연관된다. 이론적으로 중요한 것은 두 계수 사이의 차이이다. 검증용이성은 계획된 주도적 위임보다 계획된 보조적 위임과 훨씬 더

강하게 연관된다(차이=+0.185). 검증용이성을 낮음에서 높음으로 올리면 계획된 보조적 위임의 예측확률은 .340 에서 .412 로(상대적 +21%) 상승하는 반면, 계획된 주도적 위임의 예측확률은 거의 변하지 않는다(.143→.120). Wald 검정에서 두 계수의 차이는 Δβ=+0.185(SE 0.013), χ^2=215.53, p<.001 로 유의하다. 합의 코딩에서 두 계수는 부호가 반대이지만, 더 중요한 것은 보조측 계수가 주도측보다 유의하게 크다는 점이다(약 3.7 배). 주도측 계수 자체는 작고(검증용이성 1 단계 상승당 계획된 주도적 위임의 오즈 약 3.9% 감소) 그 부호는 대안적 검증 코딩에 걸쳐 강건하지 않으므로, 우리는 음의 주도측 계수를 핵심 결과로 보지 않는다. 강건한 것은 비대칭이다. 추정한 모든 검증 코딩에서 검증용이성은 계획된 주도적 위임보다 계획된 보조적 위임과 유의하게 더 강하게 연관된다. 이 'AI 보조적 위임 우위 비대칭'이 이론이 예측하고 자료가 일관되게 지지하는 결과이다. 가설 3 이 지지되며 가설 4 와도 일치한다(서술적 비교).

표 2. 과업 의존적 AI 위임의 다항 로짓 추정치

과업 속성 / 변수	현재 보조	현재 주도	계획 보조	계획 주도
검증용이성	+.217***	+.195***	+.145***	−.040***
결과 노출도	−.296***	−.511***	−.418***	−.699***
맥락의존성	−.395***	−.535***	−.158***	−.228***
AI 정확성 신뢰	+.470***	+.915***	+.465***	+.842***

주: 기준 범주는 무사용(A)/사용계획 없음(B). 값은 로그 오즈 계수. 응답자 수준 군집-강건 표준오차. 과업 속성 3 점 척도(1=낮음,3=높음), AI 신뢰 평균중심화. *** p<.001. 관측치 N=383,761(A)/265,025(B), 의사 R^2=.108(A)/.082(B).

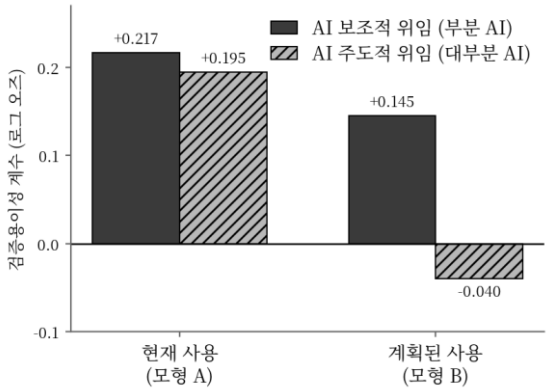


그림 1. 검증용이성의 다항 로짓 계수(로그 오즈). 현재 사용에서는 보조·주도 모두와 정적으로 연관되나, 계획된 사용에서는 AI 주도적 위임보다 AI 보조적 위임과 더 강하게 연관된다(보조 우위 비대칭).

4.3 신뢰 및 강건성

AI 정확성 신뢰는 두 모형 모두에서 가장 강한 개인 수준 예측변수이다. 모형 A 에서 보조 +0.470, 주도 +0.915(비율 1.94), 모형 B 에서 +0.465/+0.842(비율 1.81)이다. 신뢰는 AI 사용을 확대하되 주도적 위임을 약 2 배 더 강하게 확대하면서도 과업 구조를 지우지는 않는다. 학생·초기 경력자는 일관되게 낮은 위임을 보이고(β≈−0.30~−0.47), 전문 코딩 경력은 주도적 위임에 작은 부적 효과를 보인다(−0.023). 다섯 가지 강건성 점검(16 개국 제한 n=275,402, 이진·순서형 대안 척도, 심층 검증 코딩, 2 속성 다중공선성 점검, leave-one-task-out 재추정)에서 보조 우위 비대칭(계획 보조와 계획 주도의 검증용이성 계수 차이)은 모두 정적·유의하게 유지된다. 세 가지 검증 코딩(합의·Coder 1·심층 검증)에서 이 차이는 +0.165~+0.204 로 모두 유의하며, leave-one-task-out 에서는 13 개 중 12 개 사양에서 양(+0.097~+0.283)이다. 유일한 예외는 Predictive analytics 를 제외한 경우로 차이가 영에 가까워지고(−0.012), 가장 영향력이 큰 Deployment and monitoring(최고 결과 노출도·최저 주도 의도)를 제외하면 주도측 계수가 양(+0.301)으로 뒤집힌다. 강건한 것은 두 상태 간 관계의 비대칭이지 검증용이성의 절대적 주도 억제가 아니며, 이것이 우리가 절대적 억제를 주장하지 않는 이유이다.

V. 논의 및 결론

본 연구는 AI 사용이 사용자 수준 채택보다 자율성 배분으로 더 잘 이해됨을 보였다. 이론적 기여는 세 가지다. 첫째, 연구 대상을 채택에서 자율성 배분으로 이동시킨다. 대부분의 변량이 개발자 내 과업 간에 있으므로, 사용자 수준 모형은 서로 다른 과업 수준 메커니즘을 평균 처리해 버릴 위험이 있다. 둘째, 보조적·주도적 위임을 단일 연속선의 점이 아니라 비대칭적 상태로 이론화한다. 셋째, 검증용이성의 보조 우위 비대칭을 인간 감독의 옵션가치로 설명한다. 쉬운 검증은 인간을 루프에서 제거하지 않고 루프를 더 가치 있게 만든다. 이 설명은 자기위협[10]이나 비용 하의 통제가능성[20] 논의와 구별된다. 단조적 위험 감소 메커니즘이라면 검증이 쉬운 과업에서 주도적 위임도 그만큼 늘어야 하나, 우리는 보조 쪽이 훨씬 더 강하게 연관됨을 관찰했다. 또한 시간불변 과업 교란이라면 두 프레임 모두에서 작동해야 하나, 비대칭은 계획된 의도에서 더 뚜렷하다.

실무적 함의는 네 가지다. 첫째, 자동화만이 아니라 보조적 위임을 위해 설계하라. 검증이 쉬운 과업에서도 개발자는 대화형 제안, 수락/거부 인터페이스처럼 감독을 보존하는 도구를 선호할 수 있다. 둘째, 감독 인프라를 작업 흐름에 내장하라. 투명한 diff, 자동 테스트 생성, 가정·불확실성 표시, 롤백이 보조적 위임의 가치를 만든다. 셋째, 일반 AI 정책 대신 과업별 거버넌스를 사용하라. 맥락의존성 결과는 장벽이 일반적 신뢰가 아니라 상황적 지식임을 보인다. 넷째, 주도적 위임은 결과 노출·맥락이 낮고 사후 감사가 가능한 과업에 한정하라.

한계는 다음과 같다. Stack Overflow 표본은 자기선택적이고, 설문은 횡단면이며, 의도는 실제 행동이 아니다(의도-행동 간극). 검증용이성은 작업자-지각 평가가능성으로 조작화되므로 이 구성개념 관련 결과는 해당 조작적 정의에 관한 증거로 해석해야 하며, 세 과업 속성은 13 개 과업에 걸쳐서만 변하므로 효과는 13 개 워크플로 과업에 대한 증거로 본다. 과업 속성을

식별하는 유효 변량은 과업 수준(13 개)인 반면 표준오차는 응답자 수준에서 군집화되었으므로, 과업 속성 계수의 정밀도는 과업 수준 불확실성을 과소평가할 수 있다. 따라서 핵심 근거는 개별 p 값이 아니라 leave-one-task-out 과 대안 코딩에 걸친 비대칭의 안정성에 둔다. 결론적으로 생성형 AI 는 보조에서 자동화로 가는 단순한 행진을 만들지 않는다. 개발자는 과업마다 자율성을 배분하며 감독이 가치 있는 곳에서 인간 통제를 보존한다. 검증가능성은 인간 감독을 불필요하게 만들기보다 더 가치 있게 만들어, 개발자를 루프 안에 유지한다.

참고문헌

[1] Fortune, "Uber burned through its entire 2026 AI budget in four months: Now its COO is questioning whether it's worth it," Fortune, May 2026.

[2] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," Human Factors, Vol. 46, No. 1, pp. 50-80, 2004.

[3] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," Human Factors, Vol. 39, No. 2, pp. 230-253, 1997.

[4] M. Schemmer, N. Kuehl, C. Benz, A. Bartos, and G. Satzger, "Appropriate reliance on AI advice: Conceptualization and the effect of explanations," Proc. 28th Int. Conf. on Intelligent User Interfaces, pp. 410-422, 2023.

[5] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Quarterly, Vol. 13, No. 3, pp. 319-340, 1989.

[6] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," MIS Quarterly, Vol. 27, No. 3, pp. 425-478, 2003.

[7] D. L. Goodhue and R. L. Thompson, "Task-technology fit and individual performance," MIS Quarterly, Vol. 19, No. 2, pp. 213-236, 1995.

[8] D. Russo, "Navigating the complexity of generative AI adoption in software engineering," ACM Transactions on Software Engineering and Methodology, Vol. 33, No. 5, Article 135, 2024.

[9] J. Zacharias, A. Popova, M. von Zahn, J. Chen, and O. Hinz, "Developers' dilemma: Opportunities and pitfalls of generative AI for software development," Business & Information Systems Engineering, pp. 1-27, 2026.

[10] M. Adam, C. Diebel, M. Goutier, and A. Benlian, "Navigating autonomy and control in human-AI delegation," Decision Support Systems, Vol. 180, Article 114193, 2024.

[11] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," Journal of Experimental Psychology: General, Vol. 144, No. 1, pp. 114-126, 2015.

[12] J. M. Logg, J. A. Minson, and D. A. Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," Organizational Behavior and Human Decision Processes, Vol. 151, pp. 90-103, 2019.

[13] A. K. Dixit and R. S. Pindyck, Investment under Uncertainty, Princeton University Press, 1994.

[14] L. Trigeorgis, Real Options: Managerial Flexibility and Strategy in Resource Allocation, MIT Press, 1996.

[15] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, Vol. 20, No. 1, pp. 37-46, 1960.

[16] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, Vol. 33, No. 1, pp. 159-174, 1977.

[17] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," IEEE Trans. Systems, Man, and Cybernetics—Part A, Vol. 30, No. 3, pp. 286-297, 2000.

[18] D. McFadden, "Conditional logit analysis of qualitative choice behavior," in Frontiers in Econometrics, P. Zarembka (Ed.), Academic Press, pp. 105-142, 1974.

[19] A. C. Cameron and D. L. Miller, "A practitioner's guide to cluster-robust inference," Journal of Human Resources, Vol. 50, No. 2, pp. 317-372, 2015.

[20] L. Wang and H. C. B. Lee, "How costs influence preferences for control in generative AI: Human-guided vs. GenAI-based delegated search," Information Systems Research, Advance online publication, 2026.

사 사

이 논문은 2026 년도 교육부 및 서울특별시의 재원으로 서울 RISE 센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다 (2026-RISE-01-018-04).

This research was supported by the “Regional Innovation System & Education (RISE)” through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government (2026-RISE-01-018-04).

생성형 AI 사용 고지

본 논문 작성 과정에서 저자는 언어 교정, 가독성 개선, 그림·표 생성을 위한 코드 작성 보조 목적으로 생성형 AI 도구를 사용하였다. 연구 내용·자료 분석·결과 해석·이론적 주장은 AI 로 생성하지 않았으며, 저자는 모든 내용을 검토·수정하고 최종 책임을 진다.

LawSuit: RAG 및 쿼리 확장 기반 민법 상담 시스템의 환각 완화 연구

전예은*, 진수빈*, 박현진*, 성낙준*
* 한신대학교 AI·SW대학
e-mail:{jjh1221jjh, vheh5053, guswls, njsung}@hs.ac.kr

LawSuit: A RAG and Query Expansion-Based Civil Law Consultation System for Hallucination Reduction

Yeeun Jun*, Su Been Jin*, Hyeon Jin Park*, Nak-Jun Sung*
*School of Computing and Artificial Intelligence, HanShin University

요 약

시민들은 일상에서 임대차, 상속, 소비자 분쟁 등 생활 법률 문제에 빈번히 직면하지만, 법조문의 복잡성과 높은 비용으로 인해 법률 상담 접근성에 큰 제약을 겪고 있다. 본 논문은 이러한 문제를 완화하기 위해 일반 시민이 생활 법률 질문을 입력하면 실제 조문 근거와 함께 행동 지침을 제공하는 AI 법률 상담 시스템 LawSuit를 제안한다. Lawsuit은 조문 근거를 통한 상담을 제공하기 위해 민법 PDF 파일을 전처리해 조문 단위로 청킹한 뒤, BGE-M3 임베딩과 FAISS 의미 검색을 결합한 검색 구조를 적용하였다. 또한, 사용자 질문을 일상 표현과 법률 용어로 다각화하는 멀티 쿼리 확장을 도입하여 관련 조문 탐색 범위를 넓혔다. 10개의 테스트 질문을 활용하여 GPT-4o 단독(C0), RAG(C1), RAG 및 멀티 쿼리 확장(C2) 조건의 실험을 수행한 결과, C0은 10건 중 4건(40%), C1은 1건(10%), C2는 2건(20%)의 환각 또는 조문 오적용이 발생하였다. 이는 RAG가 조문 검색을 통해 학습 데이터 의존 오류와 법령 혼용을 줄이는 데 효과적이나, 쿼리 확장은 검색 범위를 넓히는 동시에 노이즈 조문 유입 가능성을 높일 수 있음을 시사한다.

키워드: LawSuit, 법률 상담, RAG, 쿼리 확장, 환각 완화

I. 서론

생활 법률 문제는 임대차 보증금 반환, 계약 해지, 상속, 손해배상 등 일상에서 빈번하게 발생하지만, 일반 시민이 법조문을 직접 검색하고 자신의 상황에 정확히 적용하기는 매우 어렵다. 기존의 법령 검색 포털은 법조문 원문이나 단편적인 정보를 제공하는 데 그치는 경우가 많으며, 전문적인 변호사 상담은 높은 비용과 대기 시간이라는 제약이 존재한다. 최근 대규모 언어 모델(LLM)을 활용한 상담형

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

서비스가 확산되면서 이러한 장벽을 낮출 수 있는 가능성이 열렸으나, 법률 영역에서는 모델이 내부 지식에만 의존하여 존재하지 않는 조문 번호를 생성하거나 부정확한 법적 효과를 단정하는 환각(Hallucination)이 치명적인 위험으로 이어질 수 있다.

본 연구는 이러한 문제를 완화하기 위해 외부 민법 조문 데이터베이스를 기반으로 검색된 근거만으로 답변을 생성하는 RAG(Retrieval-Augmented Generation) 시스템인 LawSuit을 제안한다. 본 연구의 주요 기여는 다음과 같다. 첫째, 민법 PDF 파일을 전처리하여 조문 단위로 청킹된 지식베이스를 구축하였다. 둘째, BGE-M3 임베딩, FAISS 검색 및 쿼리 확장을 결합한 검색 파이프라인을 설계하였다. 셋째, 10개의 생활 법률 질문을 대상으로 GPT-4o 단독, RAG 단독, RAG와 쿼리 확장 적용 조건을 비교하는 실험을 수행하여 환각 발생 양상과 조문 오적용 형태를 분석하였다.

II. 관련 연구

2.1 RAG와 법률 질의 응답

RAG는 언어 모델이 내부 파라미터에만 의존하지 않고, 신뢰할 수 있는 외부 문서를 검색하여 생성 과정의 컨텍스트로 활용하는 기법이다 [1]. 법률 질의응답 도메인에서는 답변의 유창성이나 설득력보다 근거의 정확성이 절대적으로 중요하므로, 관련 조문을 먼저 검색하고 이를 프롬프트에 포함하는 RAG 구조가 필수적이다. 특히 민법과 같이 조문 번호, 항, 호의 연결과 적용 요건이 엄격한 도메인에서는 검색된 지식의 정밀도가 전체 답변의 신뢰도를 좌우한다.

2.2 쿼리 확장 실험

일반 사용자의 법률 질문은 전문 용어보다 일상 표현으로 입력되는 경우가 많아 관련 조문 검색이 누락될 수 있다. 이에 본 연구는 멀티 쿼리 확장을 통해 검색 범위를 보완하고, 비교 실험을 통해 그 효과를 분석하였다.

III. 제안방법

3.1 데이터 구축 및 전처리

LawSuit의 데이터베이스는 국가법령정보센터의 민법 PDF를 기반으로 구축하였다 [2]. PDF 텍스트에서 페이지 번호, 헤더, 푸터, URL, 공백 등 노이즈를 제거한 뒤, “제○조”, “제○조의○” 패턴을 기준으로 조문 단위 청크를 생성하였다 [3]. 최종적으로 chunk size 512, overlap 50을 적용하여 총 1,247개의 청크를 구축하였다.

3.2 LawSuit 파이프라인

전체 파이프라인은 질문 입력, 쿼리 확장, 임베딩 및 검색, 답변 생성 단계로 구성된다. 사용자의 원본 질문이 입력되면 Prompt Injection 위험 패턴(예: “이전 지시 무시” 등)을 정제한 후, QueryExpander가 OpenAI의 GPT-4o API [4]를 이용해 질문을 3개의 쿼리(원본 1개, 변형 2개)로 확장한다. 확장된 쿼리들은 BGE-M3 모델을 통해 1024차원의 임베딩 벡터로 변환되며, FAISS(IndexFlatIP)를 통해 병렬로 조문이 검색된다. 검색 결과 중 유사도 점수가 가장 높은 상위 5개(top-k 5)의 조문이 선별된다 [5-7].

답변 생성 단계에서는 선별된 조문을 프롬프트 컨텍스트로 구성하고, “컨텍스트에 없는 법조문을 임의로 창작하지 말 것”이라는 엄격한 제약을 부여한다. 일관성 확보를 위해 모델의 temperature는

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

0.1로 고정하였으며, 생성되는 응답은 '상황 요약', '지금 할 일', '법적 근거', '주의사항', '전문가 상담 필요 여부'의 구조화된 항목으로 출력된다.

IV. 실험 결과 및 평가

평가 문항은 소멸시효, 취득시효, 유류분, 법정이자, 임대차 묵시갱신, 도급·수급, 상계, 계약해제, 공동불법행위 등 조문 적용 난도가 높은 민법 쟁점 10개로 구성하였다. 평가를 위한 '환각'의 정의는 다음과 같이 설정하였다. 첫째, 존재하지 않는 조문 번호 인용, 둘째, 실제 조문 내용 또는 적용 요건 오류, 셋째, 근거 없는 단정, 넷째, 민법 범위 밖 타 법령의 혼용으로 규정하였다. 본 연구는 GPT-4o 단독 조건(C0), 원본 질문만을 검색에 활용한 RAG 조건(C1), 원 질문과 확장 쿼리 2개를 함께 사용하는 RAG+멀티 쿼리 조건(C2)을 비교하였다.

4.1 종합 결과

비교 실험 결과, C0에서는 4건(40%), C1에서는 1건(10%), C2에서는 2건(20%)의 환각 및 오적용이 발생하였다. 그림 1은 조건별 오답 발생 위치를 문항 단위로 나타낸 것이다.

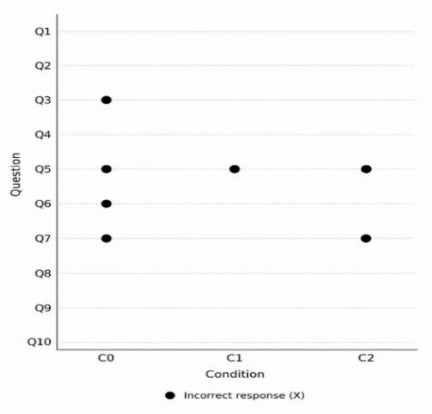


그림1. 실험에 따른 오답 분포

4.2 환각 발생 사례 분석

C0의 오류는 주로 모델의 사전 지식에 의존한 조문 요건 추가와 법령 혼용에서 발생하였다. Q3에서는 민법 제245조에 없는 취득시효 요건을 추가하였고, Q6에서는 민법 범위를 벗어난 주택임대차보호법을 혼용하였다. 반면 C1과 C2에서는 이러한 법령 혼용 오류가 줄어 RAG의 환각 억제 효과가 확인되었다. 다만 Q5는 세 조건 모두에서 오류가 발생하였으며, Q7은 C2에서 유사 조문이 혼재되어 오적용이 발생하였다.

표1. 환각 발생 케이스 요약

조건	문항	핵심 오류
C0	Q3	제245조에 없는 취득시효 요건 추가
	Q5	제379조 적용 요건 오판
	Q6·Q7	특별법 혼용 및 담보책임 기간 오인용
C1	Q5	제379조 적용 요건 연결 오류
C2	Q5·Q7	유사 조문 혼동 및 확장 쿼리 노이즈

4.3 결과 해석 및 한계

본 실험 결과는 RAG 기반 조문 검색(C1)이 기본 모델(C0) 대비 환각을 크게 감소시키고, 조문에 근거한 정확한 설명 제공에 기여함을 입증한다. 반면, 쿼리 확장(C2)은 사용자의 일상 표현과 법률 용어 간의 간극을 줄여 검색 커버리지를 높이는 장점이 있으나, 본 실험의 Q7 사례처럼 노이즈 조문이나 유사 조문을 유입시켜 최종 답변의 혼란을 가중시키는 부작용도 존재함을 보여준다. 본 연구의 한계점으로는 10문항이라는 소표본 평가로 통계적 일반

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원 사업(경기RISE사업)의 연구결과 입니다.

화가 제한된다는 점과 단일 민법 데이터베이스만 활용하여 실제 복합 법률 상담 환경을 완전히 반영하지 못했다는 점이 있다.

V. 논의

실험 결과는 법률 도메인 RAG 시스템 설계 시 단순히 조문의 '존재 여부'를 확인하여 주입하는 것을 넘어, 질문 상황과 조문의 '적용 적합성'을 검증하는 메커니즘이 필수적임을 시사한다. 이를 위해 향후 연구에서는 검색된 여러 조문의 문맥적 정합성을 평가하는 재순위화(Re-ranking) 과정의 도입을 통해 검색 결과의 유효성을 향상시킬 계획이다 [8]. 또한 주택임대차보호법, 근로기준법 등 특별법을 포함한 다중 법령 검색과 대법원 판례 데이터의 결합을 통해 상담의 실효성을 높여야 한다. 본 시스템은 법률 자문을 직접 대체하는 것이 아니라, 사용자가 자신의 법적 권리를 확인하고 다음 행동을 모색할 수 있도록 돕는 보조 도구로써 기능해야 한다.

VI. 결론

본 연구는 민법 조문 단위 지식베이스를 기반으로 한 RAG 법률 상담 시스템 LawSuit를 제안하고, GPT-4o 단독, RAG, RAG+ 멀티 쿼리 조건을 비교하였다. 실험 결과, RAG는 환각률을 40%에서 10%로 낮추며 법률 상담 답변의 근거성을 개선하였다. 반면 멀티 쿼리 확장은 검색 범위를 넓히는 장점이 있으나, 유사 조문 유입으로 인한 오적용 가능성도 확인되었다. 향후에는 검색 조문의 적용 적합성을 검증하는 재순위화와 다중 법령·판례 기반 검색 확장이 필요하다.

참고문헌

[1] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems, 2020.

[2] 국가법령정보센터, "민법 (법률 제21454호)", https://www.law.go.kr, 2026.

[3] Finardi, P., Avila, L., Castaldoni, R., Gengo, P., Larcher, C., Piau, M., ... & Caridá, V. (2024). The chronicles of rag: The retriever, the chunk and the generator. *arXiv preprint arXiv:2401.07883*.

[4] OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2024.

[5] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, Vol. 7, No. 3, pp. 535-547, 2019.

[6] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, Vol. 3, No. 4, pp. 333-389, 2009.

[7] J. Chen, et al., "BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation," arXiv preprint arXiv:2402.03216, 2024.

[8] Zhang, L., Wang, B., Qiu, X., Reddy, S., & Agrawal, A. (2025, November). Rearank: Reasoning re-ranking agent via reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 2458-2471).

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

ISSN 3092-3301
Vol. 0. No. 0, pp. -, 2025
https://data.doi.or.kr/10.23392/JKADS.2025.0.0.0113

KNHANES 기반 근로 청년층의 의료 미충족 예측 모델 개발 및 고위험군 특성 분석

김명하* · 박한결 · 강정엽 · 김서윤 · 김은수 · 이상민†
차의과학대학교 헬스케어융합학부 †차의과학대학교 AI의료데이터학 전공
* 제1저자 · † 교신저자 (Professor) e-mail: qktkrehsrktm150333@naver.com

Development of a Machine Learning Prediction Model for Unmet Medical Care among Working Young Adults Using KNHANES and Identification of High-Risk Subgroups

Myoungha Kim · Hangyeol Park · Jeongyeop Kang · Seoyun Kim · Eunsu Kim · Sangmin Lee†*
Dept. of Healthcare Convergence, CHA University †Dept. of AI Medical Data Science, CHA University

요 약 연구목적: 청년 근로자의 의료 미충족은 건강 악화 및 만성질환 진행과 관련된 공중보건 문제이나, 머신러닝 기반 위험 예측 및 그 해석에 관한 국내 연구는 제한적이다. 본 연구는 국민건강영양조사(KNHANES) 2020-2024년 자료로 19-39세 근로 청년의 미충족 의료를 예측하는 머신러닝 모델을 개발하고 SHAP 기법으로 고위험군 특성을 규명하고자 하였다. 연구방법: 5개년 자료를 통합하여 경제활동 청년 4,122명을 분석하였다. 23개 독립변수로 Logistic Regression(LR), XGBoost, LightGBM을 구축하고 Optuna로 50회 튜닝, 5-fold 교차검증하였으며, Platt scaling 사후 캘리브레이션과 Bootstrap 95% 신뢰구간을 적용하였다. 결과: 미충족률은 8.61%였다. 보정된 LR의 AUC-ROC는 0.730, Brier score는 0.075로 보정 전 대비 약 67% 개선되었다. 평소 스트레스 인지가 가장 강력한 예측 변수였으며(OR=0.634), BP1=1 청년 미충족률(22.3%)은 BP1=4(4.4%)의 약 5배였다. 정신상담 경험(OR=1.495)·자살생각(OR=2.162)도 동일한 양의 연관을 보였다. 최고위험군의 실제 미충족률은 50%로 기저율의 약 6배였고(검증표본 N=4로 소수), 상위 30% 스크리닝으로 위험자의 약 59%를 식별하였다. 결론: 청년 미충족 의료는 다수 위험요인의 누적 효과로 발생하며, 스트레스·정신건강 관리가 의료 접근성 향상의 핵심 축임을 시사한다. 가장 강력한 예측 요인인 정신건강 지표가 격년으로만 조사된 점은 국가 건강조사에서 청년 정신건강 문항의 상시 조사 필요성을 제기한다.

주제어 : 의료 미충족, 청년 근로자, 머신러닝, 설명가능 인공지능(SHAP), 캘리브레이션, 위험 계층화

Abstract This study develops machine learning models predicting unmet medical care among 4,122 economically active young adults aged 19-39 using KNHANES 2020-2024, and identifies high-risk subgroups via SHAP. Logistic Regression (LR), XGBoost, and LightGBM were tuned with Optuna (50 trials) and calibrated using Platt scaling, with bootstrapped 95% confidence intervals. The unmet care rate was 8.61%. Calibrated LR achieved AUC-ROC 0.730 and Brier score 0.075, a 67% improvement after calibration. Perceived stress was the strongest predictor (adjusted OR=0.634); the unmet rate of the most-stressed youth (22.3%) was about five times that of the least-stressed (4.4%). Mental health counseling (OR=1.495) and suicidal ideation (OR=2.162) showed consistent positive associations. The highest-risk stratum (predicted probability >=20%) showed a 50% unmet rate (test n=4), about six times the baseline, and screening the top 30% identified about 59% of unmet cases. Managing stress and mental health is a key axis for improving healthcare access among working young adults.

Key Words : Unmet Medical Care, Young Workers, Machine Learning, Explainable AI (SHAP), Calibration, Risk Stratification

1. 서론

의료 미충족(unmet medical care)이란 의료적 필요(need)가 발생하였음에도 적절한 의료서비스를 이용하지 못하는 상태를 의미하며[1], 미진단·미치료 만성질환의 진행, 응급의료비 증가, 사망률 상승 등 부정적 결과를 초래하여 의료 형평성 평가의 핵심 지표로 활용된다.

우리나라 청년층(19-39세)은 학업·취업·경제활동의 전환기에 위치하여 사회적 안전망의 사각지대에 놓이기 쉽다. 특히 임시·일용직, 자영업 등 비전형 노동에 종사하는 청년 근로자는 직장 의료보장과 시간적 여유의 부족으로 미충족 의료 위험이 높은 것으로 보고된다[2]. 또한 COVID-19 팬데믹 초기 의료접근성이 청년·여성 등 특정 집단에서 저하되었으나[3], 팬데믹 정점기와 회복기를 포함한 5개년 통합 자료로 시기별 효과를 정량 분석한 연구는 제한적이다.

본 연구를 구상하는 과정에서 연구자는 근로 청년을 직접 대상으로 하는 의료 지원 정책이 어떻게 마련되어 있는지 확인하고자 청년 정책 현장에 자문을 구하였다. 그러나 돌아온 답변은 근로 청년만을 표적으로 한 의료 정책은 많지 않으며, 흩어진 정보를 청년이 스스로 찾아내기도 쉽지 않다는 것이었다. 이러한 정책 공백은 청년 미충족 의료 문제가 학술적 분석에 앞서 정책 설계의 우선순위에서부터 충분히 다루어지지 못하고 있음을 시사하며, 본 연구가 '어떤 청년에게 어떤 의료 지원을 먼저 제공할 것인가'라는 실천적 질문에 근거를 제공해야 한다는 문제의식의 출발점이 되었다.

기존의 미충족 의료 연구는 주로 단일 시점의 로지스틱 회귀에 머물러[4] 변수 간 비선형 상호작용을 포착하기 어려웠다. 최근 머신러닝 기법의 활용이 확대되어 잠재 위험군을 식별하려는 연구가 늘고 있으나[5], 모델의 블랙박스 특성을 보완하기 위해 SHAP 등 설명 가능 인공지능 기법의 병행이 권고된다[6]. 또한 머신러닝 예측 확률은 실제 발생률과 체계적 차이를 보여 사후 캘리브레이션이 필요하나[7], 국내 미충족 의료 연구에서 이를 적용한 사례는 드물다.

이에 본 연구는 (1) 5개년 통합 자료로 근로 청년의 미충족 의료 예측 모델을 개발하고, (2) 세 모델 성능을 Optuna 튜닝과 Bootstrap 신뢰구간으로 비교하며, (3) Platt scaling으로 신뢰성 있는 위험 확률을 산출하고, (4) SHAP으로 주요 예측 요인을 규명하며, (5) 위험 계층화로 정책 우선 타겟을 식별하는 것을 목적으로 한다.

2. 연구방법

2.1 연구설계 및 분석자료

본 연구는 KNHANES 2020-2024년 5개년 원시자료를 활용한 횡단면 연구이다. KNHANES는 질병관리청이 매년 시행하는 전국 단위 복합표본조사로 건강설문·검진·영양조사로 구성된다[8]. 본 연구에서는 건강설문조사의 의료이용·정신건강·경제활동 영역과 검진조사의 신체계측 영역을 사용하였다.

2.2 연구대상

총 34,640명의 5개년 통합 자료에서 (1) 만 19-39세이며 (2) 경제활동 상태가 '일을 하였음'(EC1_1=1)인 자를 포함하고, 미충족 의료 응답이 유효하지 않거나 핵심 변수에 결측이 있는 자를 제외하였다. 단계별 표본 흐름은 Table 1과 같으며, 최종 분석 대상은 4,122명, 미충족 의료 경험자는 355명(8.61%)이었다.

Table 1. Sample selection flow

Step	N	Excluded
Integrated (2020-2024)	34,640	-
Aged 19-39	6,986	-27,654
Economically active	4,391	-2,595
Valid unmet response	4,235	-156
Complete cases (final)	4,122	-113

2.3 변수의 정의

종속변수는 '최근 1년간 병의원 진료·검사가 필요하였으나 받지 못한 경험'(M_2_yr)을 이진화하여 사용하였다(예=1, 아니오=0). 독립변수는 Andersen의 의료이용 행동모형[9]에 따라 소인성(성별·연령·거주지역·교육수준), 가능성(개인·가구소득·건강보험·직업·종사상지위·근로시간·민간보험), 필요(주관적 건강·의사진단 질병 개수·활동제한·평소 스트레스 인지·정신상담 경험·흡연·음주·운동·비만도) 요인의 23개 변수를 선정하였다.

이 중 의사진단 질병 개수(n_disease)는 5개년에 일관되게 조사된 7개 질환, 즉 고혈압·이상지질혈증·당뇨병·천식·알레르기비염·아토피피부염·신부전의 의사진단 여부(코드 1만 1로, 0·8·9는 0으로) 7개 변수를 합산한 파생변수이다. 대사·순환계 질환뿐 아니라 청년에서 유병이 높은 알레르기·호흡기 질환을 함께 포함하여, 청년이 실제로 경험하는 질병 부담을 현실적으로 반영하였다. 시기 효과 통제를 위해 연도 변수를 포함하였다.

2.4 분석방법

2.4.1 자료 전처리 및 클래스 불균형 처리

범주형 명목변수(성별·거주지역·보험종류·직업·종사상지위·근로시간 범주·비만도·연도)는 원-핫 인코딩을 적용하여 42개 입력변수로 확장하였다. 모든 결측값(특수코드 9, 888, 999 포함)은 완전 사례 분석으로 제외하였다. 자살생각(BP6_10)은 2020·2024년 미조사되어 본분석에서 제외하고 보조분석에만 활용하였다. 전체 자료를 8:2로 학습용(N=3,297)·검증용(N=825)으로 층화 분할하고, 사후 캘리브레이션 학습을 위해 학습용을 다시 8:2로 분리하여 모델 학습용 2,637명, 캘리브레이션용 660명으로 구분하였다. 미충족 양성 비율이 약 8.6%로 불균형하므로 LR·LightGBM

은 class_weight='balanced'를, XGBoost는 scale_pos_weight를 적용하였다.

2.4.2 모델 학습 및 하이퍼파라미터 튜닝

Logistic Regression(L1·L2 정규화), XGBoost[10], LightGBM[11] 세 모델을 비교 학습하였다. 하이퍼파라미터는 Optuna[12]의 Tree-structured Parzen Estimator(TPE) 알고리즘으로 모델당 50회 탐색하였으며, 5-fold 층화 교차검증의 평균 F1-score를 목적함수로 설정하였다. 탐색 공간은 LR의 정규화 강도(C), 정규화 유형(L1/L2)과, 트리 모델의 트리 개수·최대 깊이·학습률·L1/L2 정규화 항(reg_alpha, reg_lambda)·잎 노드 수 등을 포함하였다.

2.4.3 사후 캘리브레이션

머신러닝 모델의 예측 확률이 실제 발생률과 차이를 보이는 문제를 해결하기 위해, 캘리브레이션 자료(N=660)에서 Platt scaling(시그모이드 적합)과 Isotonic regression(단조 비모수 회귀)을 학습 후 비교하였다[7]. 보정 품질은 Brier score와 reliability diagram으로 평가하였으며, Isotonic regression은 유연하나 표본이 작을 때 과적합 위험이 있어, 표본 크기를 고려하여 더 안정적인 Platt scaling을 최종 채택하였다.

2.4.4 평가지표 및 신뢰구간

Test 자료에서 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-score, AUC-ROC, AUC-PR, Brier score를 산출하였다. 재현율은 실제 미충족자 중 모델이 포착한 비율로 위험군을 놓치지 않는 능력을, AUC-PR은 양성 비율이 낮은 불균형 자료에서의 식별력을, Brier score는 예측 확률과 실제 결과의 평균제곱오차로 보정 품질을 나타낸다. 위험군을 놓치지 않는 능력이 핵심이므로 재현율과 AUC-PR을 주된 지표로 강조하였다. 모든 지표의 95% 신뢰구간은 1,000회 부트스트랩 재추출로 산출하였고, 의사결정 임계값은 학습용에서 F1-score를 최대화하는 값으로 설정하였다.

2.4.5 모델 해석·위험 계층화·민감도 분석

트리 기반 LightGBM에 SHAP TreeExplainer[6]를 적용하여 변수별 기여도와 작용 방향을 시각화하였다. 보정된 예측 확률을 5단계(Very Low~Very High)로 계층화하여 각 계층의 실제 미충족률을 산출하고, lift chart로 상위 X% 스크리닝 시 미충족 위험자 식별률을 추정하는 정책 시뮬레이션을 제시하였다. 결과의 강건성은 (1) 연령 19-34세, (2) 성별, (3) 종사상지위, (4) 시기(코로나 초기·정점·회복기) 등 9개 시나리오의 민감도 분석과, 가중치 적용·미적용 분석의 주요 변수 OR 비교로 검증하였다. 모든 분석은 Python 3.11(scikit-learn 1.4, XGBoost 2.0, LightGBM 4.3, Optuna 3.6, SHAP 0.45, statsmodels 0.14)으로 수행하였으며 양측 P<0.05를 유의수준으로 하였다.

3. 결과

3.1 연구대상자의 일반적 특성

최종 분석 대상 4,122명 중 여성이 51.4%, 평균 연령 30.4세였다. 미충족군(N=355)은 비미충족군(N=3,767)에 비해 여성 비율(63.1% vs 50.3%), 비임금근로자 비율(70.1% vs 62.7%), 나쁜 주관적 건강(26.8% vs 12.9%), 활동제한(15.2% vs 3.5%), 정신상담 경험(13.8% vs 6.1%), 의사진단 질병 개수가 유의하게 높았다(Table 2).

Table 2. Characteristics by unmet medical care

Variable	No-unmet	Unmet	P
Age, mean±SD	30.36±5.80	30.45±5.82	.794
Female, %	50.3	63.1	<.001
Urban residence, %	88.2	87.9	.95
College+ , %	66.9	66.2	.59
Personal income (low), %	19.4	25.6	.041
Household income (low), %	4.6	8.7	<.001
Non-waged worker, %	62.7	70.1	.004
Working >=53h/wk, %	9.6	13.0	.09
Poor self-rated health, %	12.9	26.8	<.001
Activity limitation, %	3.5	15.2	<.001
High stress (BP1=1), %	6.1	18.6	<.001
Mental counseling, %	6.1	13.8	<.001
Current smoking, %	19.4	21.4	.39
Monthly drinking, %	68.9	65.6	.23
No. of diseases, mean	0.48	0.59	.003
Weekly work hrs, mean	37.7	39.3	.050

* Chi-square test or independent t-test. N=3,767 (no-unmet) vs 355 (unmet).

3.2 머신러닝 모델의 예측 성능

Optuna 50회 탐색 결과 최적 하이퍼파라미터는 LR이 C=0.021·L1 정규화, XGBoost가 n_estimators=237·max_depth=5·learning_rate=0.029, LightGBM이 n_estimators=300·max_depth=4·num_leaves=24·learning_rate=0.027이었다. 모든 트리 모델이 강한 정규화와 작은 max_depth로 수렴하였는데, 이는 표본 크기와 특성 수의 균형에 따른 과적합 방지 효과로 해석된다.

Platt scaling 보정 후 Test 자료에서 LR의 AUC-ROC는 0.730(95% CI 0.669-0.787), AUC-PR 0.248(0.176-0.352), F1 0.287, Recall 0.408로 세 모델 중 가장 우수하였으며, LightGBM(0.686)·XGBoost(0.668)는 다소 낮았으나 신뢰구간이 겹쳐 유의차를 단정하기 어려웠다(Table 3, Fig. 1).

Table 3. Model performance after calibration (Test, N=825)

Model	ACC	Prec	Rec	F1	AUROC	AUPR
LR	0.825	0.221	0.408	0.287	0.730	0.248
XGB	0.879	0.277	0.254	0.265	0.668	0.184
LGB	0.868	0.203	0.183	0.193	0.686	0.179

* 95% CI by 1,000 bootstrap resampling; threshold maximizing F1 on training set.

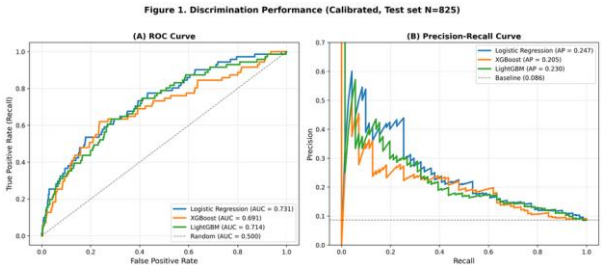


Fig. 1. ROC and Precision-Recall curves (Test, N=825)

3.3 사후 캘리브레이션의 효과

Platt scaling 적용 전 모든 모델은 미충족 확률을 체계적으로 과대 추정하였으며, 예측 확률 0.5 구간의 실제 미충족률은 약 0.10에 불과하였다. 보정 후 Brier score가 LR 0.229→0.075, XGBoost 0.197→0.076, LightGBM 0.214→0.076으로 약 60-67% 감소하여 reliability diagram에서 예측-실제 곡선이 이상적 보정선(perfect calibration line)에 근접하였다(Fig. 2). AUC-ROC는 보정 전후 동일하였는데, 이는 Platt scaling이 단조변환(monotonic transformation)이라는 성질에 기인한다. 이로써 모델의 예측 확률이 실제 발생률에 부합하는 신뢰성 있는 값으로 변환되어 위험 계층화와 정책 시뮬레이션의 토대가 마련되었다.

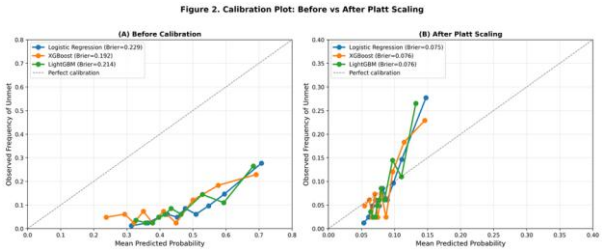


Fig. 2. Reliability diagram before/after Platt scaling

3.4 SHAP 기반 변수 중요도

LightGBM에 SHAP TreeExplainer를 적용한 결과 영향력 상위 변수는 평소 스트레스 인지(BP1), 주당 근로시간, 성별, 연령, 종사상지위, 주관적 건강상태, 시기(2022·2023) 순이었다(Fig. 3). 스트레스를 많이 느끼는 청년일수록 SHAP 값이 미충족 위험 증가 영역에 분포하여 다변량 회귀 결과(OR=0.634)와 일관되었다. 반면 의사진단 질병 개수는 상위 변수에 들지 못해, 청년 집단에서 진단 질병 수의 기여가 주관적 건강·스트레스에 비해 제한적임을 보였다.

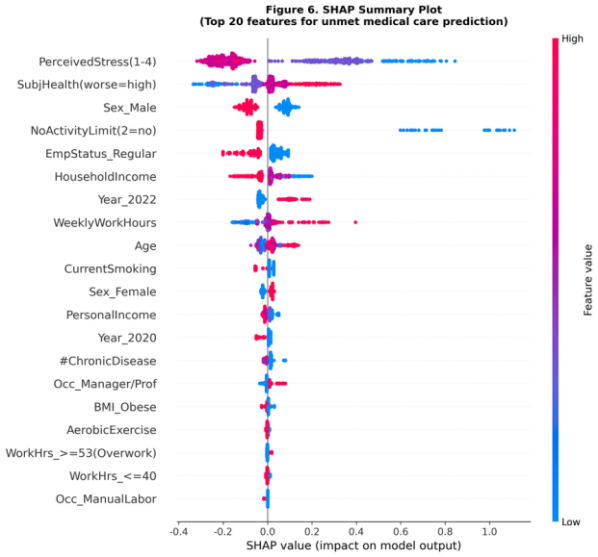


Fig. 3. SHAP summary plot (top features)

3.5 평소 스트레스 인지(BP1)의 심층 검증

BP1=1(대단히 많이) 청년의 미충족률은 22.3%로 가장 높았고 BP1=4(거의 안 느낌)에서 4.4%로 가장 낮아 약 5배 차이를 보였다. 단변량 분석에서 BP1 1점 증가(스트레스 한 단계 감소) 시 미충족 위험이 약 50% 감소하였으며(OR=0.504), 인구사회·경제·임상 변수를 보정한 다변량 분석에서도 관계가 유지되었다(OR=0.634, 95% CI 0.544-0.740, P<0.001). BP1=1군의 정신상담 경험률(25.0%)·자살생각률(25.7%)은 BP1=4군(2.5%, 1.2%)보다 10-20배 높았다. 정신상담 경험(OR=1.495)·자살생각(OR=2.162)도 동일한 양의 연관을 보여(Table 4), 스트레스 인지·정신상담·자살생각의 세 지표가 일관된 정신건강 부담 지표군임을 확인하였다.

Table 4. Multivariable ORs of mental health indicators

Variable	OR (95% CI)	P
Perceived stress (1-4)	0.634 (0.544-0.740)	<.001
Mental counseling (0/1)	1.495 (1.042-2.144)	.029
Suicidal ideation (0/1)	2.162 (1.336-3.501)	.002

* Adjusted for sex, age, education, income, self-health, no. of diseases, activity limitation, smoking, drinking, private insurance, working hours.

3.6 SHAP Dependence 상호작용 분석

SHAP dependence plot 분석 결과, 스트레스가 높은 청년군(BP1=1,2)에서 정신상담 경험자가 비경험자보다 더 높은 미충족 위험 기여를 보여 두 정신건강 지표가 상승작용함을 시사하였다. 또한 주당 근로시간이 약 30시간을 초과하는 시점부터 미충족 위험 기여가 양의 방향으로 전환되었으며, 이 경향은 비임금근로자에서 더 강하게 나타났다(Fig. 4).

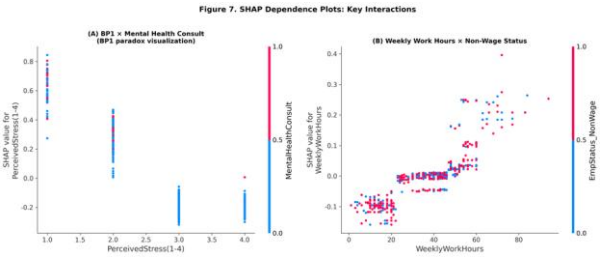


Fig. 4. SHAP dependence plot (key interactions)

3.7 위험 계층화 및 정책 시뮬레이션

보정된 LR 예측 확률을 5단계로 계층화한 결과, 위험 단계에 따라 실제 미충족률이 명확히 증가하였다(Table 5). 최고위험군(≥20%)의 실제 미충족률은 50.0%로 표본 평균(8.6%)의 약 6배였다(검증표본 N=4로 소수). Lift chart에서 예측 확률 상위 30% 청년 스크리닝 시 전체 미충족 위험자의 약 59%를 식별할 수 있어, 무작위 추출(30%)의 약 2배 효율이었다(Fig. 5).

Table 5. Risk stratification by predicted probability

Risk group	Pred. prob.	N	Actual unmet
Very Low	<5%	12	0.0%
Low	5-10%	621	5.3%
Medium	10-15%	169	17.2%
High	15-20%	19	36.8%
Very High	>=20%	4	50.0%

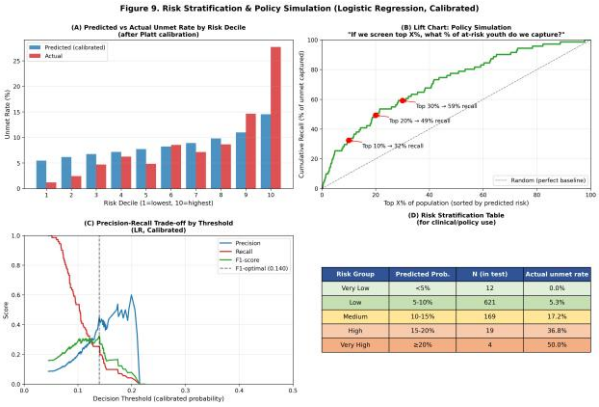


Fig. 5. Risk stratification and policy simulation

3.8 민감도 분석

9개 시나리오 민감도 분석 결과(Table 6), 비임금근로자 단독 분석에서 AUC-ROC 0.720, Recall 0.520으로 전체 표본과 유사한 강건성을 보였다. 코로나 정점 시기(2022) 미충족률은 13.3%로 다른 시기(7-8%)의 약 2배였다(해당 시기 AUC 0.611). 가중치 적용 여부와 무관하게 주요 변수의 연관 방향·유의성은 일관되었다.

Table 6. Sensitivity analysis (LR)

Scenario	N	Unmet%	AUROC	F1	Rec
Base (19-39)	4,122	8.6	0.732	0.302	0.479
Age 19-34	2,839	8.6	0.687	0.286	0.408
Male	2,002	6.5	0.687	0.151	0.154
Female	2,120	10.6	0.645	0.211	0.267
Waged	1,510	7.0	0.649	0.222	0.238
Non-waged	2,612	9.5	0.720	0.317	0.520
COVID early '20-21	1,685	7.2	0.636	0.171	0.250
COVID peak '22	797	13.3	0.611	0.297	0.524
COVID recov.'23-24	1,640	7.7	0.753	0.267	0.400

* Uncalibrated LR performance; Scenario 0 (Base) corresponds to the main analysis.

4. 고찰

본 연구는 KNHANES 5개년 통합 자료로 근로청년의 미충족 의료 예측 모델을 개발하고 SHAP과 사후 캘리브레이션으로 해석 가능성과 임상 적용성을 동시에 확보하였다. 주요 발견과 그 함의는 다음과 같다.

첫째, 세 머신러닝 모델이 거의 동등한 식별력을 보였으며 단순 모델인 LR이 트리 모델에 뒤지지 않았다. 이는 미충족 의료의 결정요인이 다수 위험요인이 동시에 작용할 때 폭발적으로 증가하는 비선형 상호작용보다, 각 요인이 독립적으로 누적되는 효과로 작용함을 시사한다. 다만 AUC-ROC 0.730은 중등도 식별력이므로, LR 위험 점수만으로 정책 개입을 단독 결정하기보다 보정된 확률 기반 위험 계층화와 함께 보조적 의사결정 도구로 활용할 때 가치가 있다. 이는 복잡한 머신러닝 인프라 없이도 일선 보건소·청년 의료 사업에서 구현 가능함을 의미한다.

둘째, 가장 강력한 예측 변수인 평소 스트레스 인지와 미충족 의료의 강한 양의 연관에 주목할 필요가 있다. BP1=1과 BP1=4 사이 미충족률이 약 5배 차이를 보였고 다변량 보정 후에도 유지되었다(OR=0.634). 이는 기존 연구[2]의 보고와 일치한다. BP1=1군의 정신상담 경험률·자살생각률이 BP1=4군보다 10-20배 높아, BP1이 객관적 정신건강 위험을 일관되게 반영하는 지표임을 보였다. 메커니즘적으로는 높은 스트레스가 시간적·심리적 여유 부족으로 이어져 의료를 미루게 되는 경로로 해석된다.

여기서 연구자가 특히 주목한 지점은, 정신건강 부담이 의료 '진입'뿐 아니라 '지속' 단계에서도 미충족을 낳는다는 점이다. 우울증 외래 적정성 평가에 따르면 첫 방문 후 3주 이내 재방문율은 42.3%에 그쳐 신규 환자의 절반 이상이 초기 단계에서 치료를 이어가지 못하였고, 8주 이내 3회 이상 방문율은 23.3%에 불과하였다[13]. 또한 우리나라 정신건강서비스 이용률은 7.2%로 미국(43.1%)·캐나다(46.5%)에 비해 현저히 낮으며, 코로나19 시기 오히려 16.5%(2016)에서 11.5%(2021)로 감소하였다[14]. 본 연구에서 정신건강 지표가 가장 강력한 예측 요인으로 나타난 것은, 청년이 정신건강 문제로 한 번 의료에 접근하더라도 치료를 지속하기 어려운 구조와 맞물려 있을 가능성을 시사한다. 즉 미충족 의료는 일회적 사건이 아니라 정신건강 부담이 높은 청년에게 반복적으로 누적되는 과정으로 보아야 하며, 정책 또한 단발성 진료 지원을 넘어 치료 지속성을 담보하는 방향으

로 설계될 필요가 있다고 판단된다.

셋째, 코로나 정점 시기(2022) 미충족률이 13.3%로 다른 시기의 약 1.7-1.8배였고 SHAP에서 year_2022의 영향력이 7번째로 높았다. 청년 근로자는 의료적 필요가 상대적으로 적음에도 정점 시기 미충족이 급증하였다는 점은, 위기 상황에서 청년 의료가 우선순위에서 밀려날 수 있음을 시사한다.

넷째, 비임금근로자(자영업) 청년의 미충족률(9.5%)은 상용직(6.1%)보다 약 1.5배 높았으며, 단독 분석에서도 모델 성능이 유지되었다(AUC 0.720). 자영업자는 보험료 부담과 영업시간 손실 기회비용 때문에 의료 이용을 미루는 경향이 있어, 자영업 청년의 의료보장 사각지대를 정량적으로 확인하였다[2].

다섯째, 사후 캘리브레이션으로 Brier score가 67% 개선되어 위험 계층화와 lift chart 기반 정책 시뮬레이션이 가능해졌다. 상위 30% 스크리닝으로 미충족 위험자의 약 59%를 식별할 수 있다는 결과는 한정된 보건 자원의 효율적 배분 근거가 된다. 국내 미충족 의료 연구에서 사후 캘리브레이션으로 임상 활용 가능한 확률을 산출한 사례는 드물어, 본 연구가 방법론적 진전을 제공하였다.

여섯째, 의사진단 질병 개수는 미충족군이 유의하게 많았으나(0.59 vs 0.48, P=0.006) SHAP 영향력은 상위에 들지 못하였다. 이는 청년의 미충족 의료가 진단된 질병의 객관적 개수보다 주관적으로 인지하는 건강 악화와 정신적 스트레스에 더 강하게 좌우됨을 시사한다. 대사·순환계 질환만으로 측정했다면 청년 유병이 드물어(약 95%가 0개) 연관 포착이 어려웠을 것이나, 알레르기·호흡기 질환을 포함함으로써 유의한 연관을 확인할 수 있었다. 그럼에도 그 예측력이 정신건강 지표에 미치지 못한다는 점은, 개입 대상 식별 시 진단 질병 수 못지않게 주관적 건강 인식과 정신건강 부담에 주목해야 한다는 본 연구의 핵심 함의를 뒷받침한다.

본 연구의 한계는 다음과 같다. 첫째, 횡단면 자료로 인과관계 추론이 어렵다. 둘째, 미충족과 강하게 연관된 자살생각 문항이 2020·2024년 미조사되어 3개년 보조분석으로만 활용되었는데, 이는 단순한 자료 제약을 넘어 청년 정신건강과 의료 접근성의 관계를 일관된 시계열로 추적하려는 연구가 국가 건강조사 설계 단계에서부터 제약받고 있음을 드러내는 발견이기도 하다. 셋째, 경제활동 청년에 한정되어 비경활·실업 청년은 별도 분석이 필요하다. 넷째, 이환 조사 항목의 연도별 차이로 질병 개수를 7개 질환으로 한정하였다.

5. 결론

본 연구는 KNHANES 5개년 자료로 근로 청년의 미충족 의료 예측 모델을 개발하고 SHAP·사후 캘리브레이션으로 해석 가능성과 임상 적용성을 확보하였다. 보정된 LR은 AUC-ROC 0.730을 달성하고 Brier score가 67% 개선되었으며, 최고위험군의 실제 미충족률은 50%로 기저율의 약 6배(검증표본 N=4 소수), 상위 30% 스크리닝으로 위험자의 약 59%를 식별하였다.

주요 위험 요인은 평소 스트레스 인지, 성별, 주관적 건강상태, 종사상지위, 활동제한, 주당 근로시간이었으며, 스트레스 인지의 강한 양의 연관은 정신상담·자살생각 등 객관적 정신건강 지표와 일관되었다. 비임금근로자·여성 청년과 코로나 정점 시기에 미충족 위험이 두드러졌다.

본 연구가 가장 강조하고자 하는 정책적 제언은 국가 건강조사에서 청년 정신건강 지표를 상시 조사해야 한다는 점이다. 정신건강 부담이 미충족 의료의 가장 강력한 예측 요인으로 확인되었음에도 자살생각 문항이 격년으로만 조사되어 5개년 일관 분석이 불가능하였던 본 연구의 경험은, 역설적으로 이 제언의 필요성을 뒷받침한다. 스트레스 인지·정신상담 경험·자살생각이 매년 일관되게 조사되어야만 청년 정신건강과 의료 접근성의 관계를 종단적으로 추적하고 정책 효과를 평가하는 후속 연구가 가능하다. 본 연구가 청년 근로자의 의료 미충족이라는 사회적 사각지대를 정량적으로 가시화하고 머신러닝 기반 위험 예측의 정책 활용 가능성을 제시하였다는 점에서 학술적·실무적 기여를 기대한다.

REFERENCES

[1] R. Allin, M. Grignon, and J. Le Grand, "Subjective unmet need and utilization of health care services in Canada," Social Science & Medicine, Vol. 70, No. 3, pp. 465-472, 2010.

[2] S. Woo, M. Son, G. Kim, and M. Choi, "Differences in unmet healthcare experience by employment type and sex," Health and Social Welfare Review, Vol. 40, No. 1, pp. 592-616, 2020.

[3] J. Kim, M. You, and C. Shon, "Impact of the COVID-19 pandemic on unmet healthcare needs in Seoul, South Korea," BMJ Open, Vol. 11, No. 8, e045845, 2021.

[4] G. Kim and J. Seo, "Differences in unmet medical service experience before and after COVID-19," Journal of Humanities and Social Sciences 21, Vol. 13, No. 6, pp. 929-942, 2022.

[5] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, Vol. 380, No. 14, pp. 1347-1358, 2019.

[6] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, Vol. 30, pp. 4766-4777, 2017.

[7] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," Proceedings of ICML, pp. 625-632, 2005.

[8] Korea Disease Control and Prevention Agency, Guidelines for the Use of KNHANES (2022-2024) Raw Data, Cheongju: KDCA, 2024.

[9] R. M. Andersen, "Revisiting the behavioral model

and access to medical care," Journal of Health and Social Behavior, Vol. 36, No. 1, pp. 1-10, 1995.
[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of ACM SIGKDD, pp. 785-794, 2016.
[11] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," Advances in Neural Information Processing Systems, Vol. 30, pp. 3146-3154, 2017.
[12] T. Akiba et al., "Optuna: A next-generation hyperparameter optimization framework," Proceedings of ACM SIGKDD, pp. 2623-2631, 2019.
[13] Health Insurance Review & Assessment Service, 2023 (2nd) Depression Outpatient Quality Assessment Results, Wonju: HIRA, 2024.
[14] Ministry of Health and Welfare, 2021 Survey of Mental Disorders in Korea, Sejong: MOHW, 2021.

성인학습이론에 기반한 비전공자 시니어 대상 파이썬 교육의 문제점 분석 및 생성형 AI 활용 개선 방안 연구

이준석, 김장현 성균관대학교 사범대학 성균관대학교 글로벌융합학부
인공지능융합전공 e-mail:s33367183@skku.edu, alohakim@skku.edu

Improving Python Learning Experiences for Senior Non-Majors : An Adult Learning Framework with Generative AI Support

Jun Suk Lee, Jang Hyun Kim Convergence and Open Sharing System , Sungkyunkwan University
Artificial Intelligence Convergence Major, School of Convergence, Sungkyunkwan University

요 약

초고령사회 진입과 디지털 전환이 가속화되는 현 시점에서, 고졸 학력의 60~70세 비전공자 시니어를 대상으로 한 파이썬 교육 설계 연구는 여전히 사각지대에 놓여 있다. 본 연구는 Knowles의 성인학습이론 6원칙을 이론적 기반으로, 서울·경기 소재 30개 대학의 파이썬 강의 유형 분석과 선행 연구 검토를 통해 시니어 대상 파이썬 교육의 핵심 문제점으로 디지털 리터러시 격차, 학습자 간 수준 비균질성, 교육 콘텐츠의 맥락 부재, 프로그램 지속성 미흡, 낮은 디지털 자기효능감의 다섯 가지를 도출하였다. 이를 해소하기 위해 기초 PC 환경 적응과 생성형 AI 친숙화를 포함한 단계적 온보딩 체계 구축, 사전 진단 기반 Level 1~3 수준별 트랙 분리 운영, 약 복용 기록·혈압 판단·가계부 자동화 등 생활밀착형 예제로의 전면 전환, 스파이럴 복습 구조가 결합된 15주 연속 정규 과정 운영, 완성 코드 산출과 언어 재프레이밍을 통한 성취 경험 중심 교수법의 5대 개선 방안을 제안하였다. 특히 ChatGPT·Gemini 등 생성형 AI를 코드 생성 보조, 오류 해설, 개념 질의응답의 보조 도구로 통합함으로써 시니어의 학습 진입 장벽을 효과적으로 낮출 수 있음을 제시하였다. 본 연구의 설계 방안은 시니어 파이썬 교육 커리큘럼의 이론적 근거이자 실천적 기준으로 활용될 수 있으며, 향후 실제 교육 현장 적용을 통한 실증적 검증 연구가 필요하다.

키워드 : 시니어, 성인학습이론, 파이썬 교육, 생성형 AI 활용, 비전공자 교육

I. 서론

4차 산업혁명과 코로나19 팬데믹 이후 디지털 역량은 전 연령층의 필수 생존 기술로 부상하였다. 국내외 주요 정책 기관들은 디지털 소양을 미래 사회의 핵심 기초 역량으로 명시하고 있으며, 정부의 디지털 인재양성 종합방안은 초등학교 5·6학년과 중학교의 코딩 교육 의무화를 2025년부터 시행하는 등 교육 전반의 디지털화를 강력히 추진하고 있다[1]. 그러나 이러한 정책적 흐름은 주로 청소년과 대학생에 집중되어 있으며, 고졸 학력의 60세 이상 비전공자 시니어 학습자를 위한 체계적 교육 설계는 여전히 사각지대에 놓여 있다. 우리나라 65세 이상 노인 인구는 2025년 전체 인구의 20.3%를 넘어 초고령사회에 공식 진입하였으며, 디지털 전환의 가속화에 따라 시니어의 프로그래밍 기초 교육 필요성이 그 어느 때보다 높아지고 있다[2]. 그러나 한국의 55~65세 성인 중 디지털 역량 레벨 2~3단계에 해당하는 비율은 4%에 불과하며 이는 OECD 20개국 중 최하위권에 머무르고 있다. 특히 고졸 학력의 60~70세 시니어는 추상적 알고리즘 사고 경험이 제한적이고 디지털 기기 노출 빈도도 낮아, 프로그래밍 학습 진입에 가장 높

은 장벽을 가진 계층에 해당한다. 파이썬은 현재 전공·비전공을 불문하고 가장 널리 교육되는 프로그래밍 언어로, 미국 상위 대학 중 69%가 입문용 교육 언어로 채택하고 있으며, 국내 대부분의 대학도 전교생 대상 SW 교육의 핵심 도구로 활용하고 있다[3]. 파이썬은 문법이 간결하고 진입 장벽이 낮으며 데이터 분석·자동화·인공지능 등 다양한 분야로 확장 가능하여 비전공자에게 특히 적합한 언어로 평가된다. 이와 동시에 ChatGPT·Gemini 등 생성형 AI의 급속한 확산은 프로그래밍 교육의 패러다임 자체를 바꾸고 있다. 생성형 AI를 보조 도구로 활용하면 비전공자도 자연어로 코드를 생성하고 오류 원인을 즉각 설명받을 수 있어, 시니어의 학습 진입 장벽을 효과적으로 낮출 수 있다. 그럼에도 불구하고 기존의 파이썬 교육 모델은 대체로 대학교 1~2학년을 주요 대상으로 설계되어 있으며, 관련 선행 연구 역시 대학생·비전공자·중고등학생 중심으로 편중되어 있는 반면 60~70세 학습자를 대상으로 한 파이썬 교육 설계 연구는 매우 부족한 실정이다. 이에 본 연구는 성인학습이론을 이론적 기반으로 시니어 대상 파이썬 교육의 구조적 문제점을 진단하고 생성형 AI를 결합한 실질적인 교육 커리큘럼 설계 방안을 제안하고자 한다.

II. 관련 연구

2.1 파이썬 교육 선행연구

서울·경기에 있는 30개의 대학에서 진행하는 파이썬 관련 강의(전공, 교양)를 분석한 결과 교육 방식은 크게 3가지로 분류하였다.

표 1. 파이썬 강의 유형 분류

유형	설명	주 대상	강좌 수
A	파이썬 기초	1학년	37
B	라이브러리 중심	1학년 2학기 ~ 2학년 1학기	18
C	심화 교육	2학년	16

대학에서 운영되고 있는 파이썬 교과목을 내용과 난이도에 따라 크게 세 가지 유형으로 구분하였다. A 유형은 1학년을 주요 대상으로 하는 파이썬 기초 교과로, 변수, 함수 등 기본 문법을 중심으로 한 입문 수준의 커리큘럼으로 구성된다. B 유형은 라이브러리 활용을 중심으로 하는 고급 기초 교과로, 파이썬 기초를 이미 학습한 학습자를 대상으로 다양한 표준 및 외부 라이브러리를 활용하여 문제 해결 능력을 확장하는 데 초점을 둔다. C 유형은 데이터 분석, 영상 처리 등 심화 내용을 다루는 전공 심화 교과로, 주로 전공자를 대상으로 설계되어 있으며, 실제 산업·연구 현장에서 요구되는 응용 역량을 강화하는 것을 목표로 한다. 한편, 일부 대학에서는 이와 별도로 컴퓨팅사고를 주제로 한 온라인 수업을 개설하여, 입학 예정자를 대상으로 사전 학습 형태의 교양·기초 교육을 제공하는 사례도 확인된다.

2.2 성인학습이론과 시니어 학습자 특성

미국의 사회교육학자 멀콤 놀즈(Malcolm Knowles)는 성인 학습자가 아동과 근본적으로 다른 학습 패턴을 지닌다는 전제에서 출발하여 안드라고지(Andragogy) 이론을 체계화하였다. 그는 성인 학습자의 특성을 “자기주도적 자아개념”, “경험의 학습자원화”, “학습 준비도”, “문제중심 학습 지향”, “내재적 동기”, “학습 필요성 인식” 여섯 가지 원칙으로 정의하였으며, 이는 단순한 이론적 명제가 아니라 교육과정 설계의 실천적 기준으로 작동하고 있다[4] . 60~70세 시니어는 수십 년간의 직업 및 생활에 대한 전반적인 경험을 보유하고 있으며, 외부 보상보다 내적 성취감에 의해 동기가 유발되는 경향이 강하다는 점에서 이 여섯 가지 원칙이 특히 강하게 발현되는 집단이다. 따라서 각 원칙을 시니어 대상 파이썬 교육의 커리큘럼 구조와 내용 배열을 결정짓는 핵심 설계 기준으로 아래 표 2과 같이 직접 연결할 수 있다.

표 2. 안드라고지 원칙에 따른 시니어 학습자 특성

안드라고지 원칙	시니어 학습자 특성
자기주도적 자아개념	스스로 학습 속도와 방향을 결정하려는 욕구가 강함
경험의 학습자원화	수십 년의 직업·생활 경험을 보유하나 추상적 이론과 연결 어려움
학습 준비도	동일 연령대 내 디지털 경험·인지 기능 개인차 극심
문제중심 학습 지향	추상적 원리보다 "내 삶의 문제 해결"에서 학습 동기 극대화
내재적 동기	외부 보상보다 성취감·사회 참여·자기 성장에 의해 동기 유발
학습 필요성 인식	"왜 배워야 하는가"가 명확할 때 학습 참여 결정

2.3 생성형 AI의 시니어 파이썬 교육 활용 가능성

교육에서 생성형 AI의 주요 장점으로는 코드 생성 보조, 오류 해설 튜터링, 개념 질의응답 기능을 들 수 있다. ChatGPT, Gemini 등 생성형 AI의 빠른 발전으로, 교육 현장에서 가장 활발히 활용되는 영역은 프로그래밍 및 디지털 리터러시 교육이며, 학습자는 코드 오류의 원인과 수정 방법을 자연어로 질문하고 즉각적인 설명을 제공받는 개인 맞춤형 24시간 튜터를 활용할 수 있게 되었다. 특히 비전공자 시니어에게 생성형 AI는 예제 코드 및 연습문제 작성을 돕는 코드 생성 보조, 디버깅 과정에서 오류의 원인과 해결 전략을 설명하는 오류 해설 튜터, 프로그래밍 및 컴퓨팅 개념에 대한 자연어 기반 질문에 응답하는 개념 질의응답 도구라는 세 가지 교육적 기능을 제공함으로써, 자기주도적 학습과 학습 장벽 완화에 기여하는 핵심 매개체로 기능한다.

표 3. 파이썬 교육에서의 생성형 AI활용 유형

활용 유형	구체적 방법	기대 효과
코드 생성 보조	약 복용 기록 파이썬 코드 만들어 줘	코드 작성 부담 감소, 진입 장벽 완화
오류 해설 튜터	오류 메시지를 AI에 붙여 넣고 원인 질문	즉각 피드백, 심리적 안전망 형성
개념 질의응답	for 반복문 쉽게 설명해 줘	강사 부재 시에도 자기주도 학습 가능

2.4 시니어 학습의 문제점 및 한계

시니어 학습자는 높은 학습 의지와 실습 중심 학습에 대한 강한 선호를 보유하고 있음에도 불구하고, 실제 교육 현장에서는 구조적·인지적·환경적 측면에서 복합적인 한계에 직면하고 있다. 특히 디지털 리터러시 격차, 학습자 간 수준 편차, 콘텐츠의 실용성 부족, 프로그램 지속성 미흡, 낮은 자기효능감은 상호 연쇄적으로 작용하며 시니어 대상 파이썬 교육의 효과성을 저해하는 핵심 요인로 이야기 할 수 있다 또한 시니어 학습자들은 AI 학습의 필요성을 강하게 인식하면서도 "어디서부터 시작해야 할지 모른다"는 생성형 AI 사용의 초기 진입 장벽을 공통적으로 경험하는 것으로 나타났다.

2.5 생성형 AI의 교육적 활용 시 유의사항

생성형 AI는 시니어 파이썬 교육에서 유용한 보조 도구이지만, 동시에 다음과 같은 한계를 지닌다. 첫째 환각 현상으로 인해 오류가 포함된 코드나 잘못된 설명을 정확한 정보로 받아들이는 경우 학습자는 스스로 검증하기 어려워 오히려 학습 혼란을 가져올 수 있다. 둘째, AI 응답에 대한 과도한 의존은 학습자의 자기주도적 문제 해결 역량의 발달을 저해할 수 있다. 이에 따라 본 커리큘럼에서는 생성형 AI를 보조 도구로 제한적으로 활용하며 학습자의 역량이 일정 수준에 도달할 때까지 담당 강사의 지속적인 검증과 학습자의 직접적인 프로그램 입력을 통한 검증 과정을 병행하여야 한다.

2.6 시니어 학습자를 위한 효율성 향상 원칙

앞선 선행연구 분석을 바탕으로, 시니어 대상 프로그래밍 교육의 효율성을 높이기 위한 다섯 가지 원칙을 다음과 같이 제시하고자 한다.

표 4. 시니어 학습자를 위한 효율성 향상 원칙

원칙	내용	설명
1	즉각적 시각화	실행 즉시 시각화 결과 출력
2	생활 연계형 콘텐츠	학습자 생활과 밀착한 실습
3	단계적 범위 확장	충분한 시간 및 반복 학습
4	동질 집단 확보	수강 전 기초/심화의 구분
5	학습 연속성 보장	체계적인 단계별 학습

첫번째 즉각적 시각화로 Turtle 그래픽스와 같이 실행 즉시 결과가 화면에 나타나는 도구를 활용하여 추상적 개념을 구체적 결과물로 연결하여 시니어 학습자의성취감과 내재적 동기를 높이며, 두번째 생활 연계형 콘텐츠 실습으로 가계부 계산, 약 복용 일정 관리, 간단한 텍스트파일 처리 등 시니어의 실제 생활과 연결된 예제를 활용하여 학습자의 생활과 연계가능성을 고려한 수업 내용 구성으로 학습자의 참여도를 높이며, 세 번째 단계적 범위 확장을 통해 기초 문법에서 점진적으로 함수, 라이브러리 활용까지 진행을 하는부분에 있어 한 주제에 1~3주 충분히 시간과 반복 숙달 기회를 제공과 함께 네 번째인 동질 집단 진단으로 사전 진단을 통한 집단동질성 확보가 시니어 교육에서도 매우 중요한 요소로 자리를 잡고 있기에 수강 전 디지털 역량 자가진단을 실시하여 기초반/심화반을 구분 운영하는 것이 효율성을 높인다. 마지막 다섯 번째로 학습 연속성 보장을통해 체계적 학습 경로를 명확히 제시하여 학습자가 다음 단계를 예측할 수 있게 해야 한다.

III. 시니어 대상 파이썬 교육의 문제점 및 개선 방안

앞서 도출한 다섯 가지 문제점을 해소하기 위해 안드라고지 6원칙과 국내 실증 연구를 결합한 구체적 개선 방안을 제시하고자 한다. 각 방안은 문제점과 1:1로 대응하여 커리큘럼 설계 전 단계에서 실행 단계까지 순차적으로 적용 가능하도록 구성하였다.

3.1 단계적 온보딩 체계 구축

파이썬 본 교육 시작 전, '디지털 친해지기' 과정을 독립 모듈로 2회 선행 운영한다. 1회차에는 키보드 타이핑·마우스 조작·파일 저장·인터넷 검색 등 기초 PC 환경 적응에 집중하고, 2회차에는 파이썬 실행 환경 구성과 함께 ChatGPT 또는 Gemini 계정 생성을 병행하여 진행하며 이때 학습자에게 생성형 AI를 언제나 옆에 있는 친절한 선생님으로 소개하고, 파이썬이 무엇인지 쉽게 설명해줘와 같은 단순 질문 실습을 통해 AI에 대한 심리적 친숙감을 먼저 형성하여 AI에 대한 신뢰 형성이 선행될 때 도구 활용도가 유의미하게 높아지기 때문이다. 아울러 학습자가 일상적으로 사용하는 이때 학습자가 일상적으로 사용하는 친숙한 앱들이 모두 파이썬과 같은 코드로 만들어졌음을 시각 자료와 함께 설명함으로써 파이썬 학습의 필요성과 친근감을 자연스럽게 형성한다. 화면과 키보드에 익숙하지 않은 학습자를 위해서는 알고리즘 순서를 카드로 직접 배열하는 언플러

그드 활동을 병행한다 이는 노인 복지관 코딩 교육 연구(이승미 외, 2024)에서 물리적 교구를 활용한 언플러그드 방식이 초기 진입 장벽을 낮추는 데 효과적임이 확인된 바 있으며[5], 온보딩 단계에서도 이와 동일한 원리를 적용한다. 또한 가족·지인과의 동반 수강 제도를 운영하여 낯선 학습 환경에서 오는 심리적 부담을 완화하고, 초기 학습 지속 동기를 함께 확보한다.

3.2 사전 진단 기반 수준별 트랙 분리 운영

수강 신청 시 간이 진단 평가를 실시하여 학습자를 세 트랙으로 배정한다. 트랙 간 이동은 학습자가 원할 경우 또는 4주, 8주 시점 재진단을 통해 상향 조정이 가능하며 강사가 학습 속도를 강요하지 않는 자기주도적 진행 환경을 보장한다. 이는 안드라고지의 '학습 준비도' 원칙에 따라 개별 학습자의 준비 상태를 교육의 출발점으로 삼는 설계 원칙을 실현하는 것이다.

표 5. 학습자 별 트랙 구분

분반	대상 기준	주요 학습 내용
Level 1 (기초입문)	타이핑 미숙, 코딩 경험 전무	변수·출력·조건문·반복문 기초, Turtle 그래픽스
BLevel 2 (기초활용)	기초 디지털 가능, 코딩 경험 미미	함수·리스트·파일 입출력, 실생활 자동화
Level 3 (심화응용)	스마트폰 능숙, 기초 코딩 경험 있음	데이터 분석·시각화·API 활용·간단한 앱

3.3 시니어 생활밀착형 예제로의 전면 전환

모든 파이썬 문법 개념을 시니어의 실제 일상과 직결된 내용으로 진행한다. 단순히 예제를 교체하는 수준이 아니라, 매 수업 도입부에 '이 코드로 오늘 내 삶의 무엇이 달라지는가'를 설명하여 선언함으로써 안드라고지의 '학습 필요성 인식' 원칙을 구조적으로 실현한다.

표 6 문법에 따른 학습 필요성 예제

문법 개념	기존 추상적 예제	시니어 맥락 전환 예제
print() 출력	Hello World 출력	오늘의 혈압·혈당 수치 기록 출력
for 반복문	숫자 순차 출력	30일 약 복용 체크리스트 자동 생성
if-else 조건문	조건에 맞는 출력	혈압 수치 정상·주의·위험 자동 판단 알림
리스트 (list)	임의 숫자 모음	이번 달 식비·교통비·의료비 지출 목록
파일 입출력	텍스트 파일 읽기·쓰기	가계부 자동 저장 및 월별 지출 합계 분석
Turtle 그래픽스	도형 그리기	손자에게 보내는 생일 카드 그리기

나아가 최종 산출물을 가족에게 실제로 전달하는 미니 프로젝트로 설계하여, 학습 결과가 현실 관계 속에서 의미를 갖도록 연결한다.

3.4 스파이럴 구조 기반 연속 정규 과정 운영

단기 일회성 교육의 한계를 극복하기 위해 15주 정규 연속 과정으로 설계하되, 인지 노화를 고려한 스파이럴(나선형) 학습 구조를 적용한다. 매 수업 도입 5~10분은 이전 수업의 핵심 코드를 함께 다시 실행하는 복습으로 시작하며, 이전 개념이 새로운 개념의 토대가 되도록 내용을 중첩 배치한다. 결석 발생 시에는 5분 이내 핵심 요약 영상을 카카오톡으로 자동 발송하여 학습 공백을 최소화한다. 중간 이탈을 방지하기 위해 3주·8주·15주 시점에 미니 발표회를 운영하여 학습 책임감과 성취 동기를 주기적으로 환기한다. 수료 후에도 월 1회 '파이썬 커피클럽' 형태의 자율 학습 모임을 유지하여 학습 공동체가 지속되도록 지원한다.

3.5 성취 경험 설계를 통한 디지털 자기효능감 강화

자기효능감 이론에 따르면 효능감 강화의 가장 강력한 경로는 직접적 성취 경험이다. [6] 이를 위해 수업마다 당일 안에 반드시 실행할 수 있는 완성 코드를 산출하도록 설계하여 학습자가 수업을 마칠 때 항상 성취감을 가질 수 있는 구조를 만들어야 한다. 또한 오류 발생 시에는 틀렸습니다 와 같은 부정적인 표현보다는 오류는 컴퓨터가 보내는 힌트와 같아요 같은 언어 재프레이밍을 사용하여 오류 발생도 학습의 일부로 받아들일 수 있도록 자연스럽게 인식하는 것이 중요하다. 4주, 8주, 15주 기간별 자기효능감 자가 체크리스트를 작성하게 함으로써 학습자 스스로 성장을 체감하는 기회를 구조적으로 제공한다. 나아가 학습자가 생성형 AI와 협력하여 코드를 완성하는 경험은 단순 실패-수정 반복보다 빠른 성취감을 제공하며 대리 경험과 언어적 설득을 동시에 활성화하는 효과가 있으나, AI가 대신 해줬다 는 의존감이 아닌 AI와 함께 내가 만들었다 와 같은 주도감을 유지하도록 AI 도움을 받은 후 반드시 코드를 스스로 한 줄씩 소리 내어 읽고 설명하는 코드 낭독 확인 절차를 도입함으로써, 시니어의 파이썬 자기효능감이 AI 의존이 아닌 실질적 역량 강화로 이어지도록 보장할 필요가 있다.

IV. 결론

IV. 결론

본 연구는 초고령사회 진입과 디지털 전환이 동시에

가속화되는 사회적 맥락 속에서 시니어 비전공자 대상 파이썬 교육이 직면한 구조적 문제점을 분석하고 성인 학습이론에 기반한 실천적 개선 방안을 제안하였다. 선행연구 분석을 통해 디지털 격차, 학습자 간 수준 비균질성, 교육 콘텐츠의 맥락 부재, 프로그램 지속성 미흡, 낮은 디지털 자기효능감의 다섯 가지 핵심 문제점이 도출되었으며, 이들은 각각 독립적으로 작용하는 것이 아니라 상호 연쇄적으로 시니어 학습자의 학습 지속 의지와 성취 수준을 저해하는 복합적 구조를 형성하는 것을 보여주고 있다. 이를 해소하기 위해 안드라고지 6원칙과 국내 실증 연구를 결합하여 다음의 5대 개선 방안을 제안하였다. 첫째 파이썬 본 교육 전 기초 PC 환경 적응과 언플러그드 활동을 포함한 단계적 온보딩 체계를 구축한다. 둘째 사전 진단 평가에 기반한 Level 1~3 수준별 트랙 분리 운영으로 개별 학습 준비도를 반영한다. 셋째 약 복용 기록, 혈압 판단, 가계부 자동화 등 시니어 생활밀착형 예제로 모든 문법 개념을 맥락화 한다. 넷째 스파이럴 복습 구조가 결합된 15주 연속 정규 과정으로 학습 지속성을 구조적으로 보장한다. 다섯째 매 수업 완성 가능한 산출물 제시와 언어 재프레이밍을 통한 성취 경험 중심 교수법으로 디지털 자기효능감을 단계적으로 강화한다. 특히 본 연구는 생성형 AI를 교학습 전반에 걸친 보조 도구로 통합하는 설계를 제안하였다. 코드 생성 보조, 오류 해설, 개념 질의응답, 복습 자료 생성, 맞춤형 예제 생성의 다섯 가지 활용 유형을 통해 생성형 AI는 강사 부재 시에도 학습자의 자기주도적 학습을 지원하는 24시간 개인 튜터로 기능할 수 있다. 다만 환각, 오류와 과도한 의존으로 인한 역량 개발 저해 가능성을 고려하여 생성형 AI는 어디까지나 학습을 보조하는 도구임을 인식하고 강사의 확인과 학습자의 직접 입력 및 검증 과정을 반드시 병행하도록 설계해야 한다. 본 연구가 제안하는 5대 개선 방안은 시니어 대상 파이썬 교육 커리큘럼 설계의 이론적 근거이자 실천적 기준으로 활용될 수 있으며 단순한 기술 전달을 넘어 학습자의 자기효능감과 내재적 동기를 동시에 구축하는 설계 철학은 파이썬 교육의 효과가 디지털 역량 강화에서 삶의 질 향상으로 확장될 수 있음을 시사할 수 있다. 다만 본 연구는 선행 문헌 분석에 기반한 이론적 제안에 머무르는 한계로 향후 연구에서는 제안된 커리큘럼을 실제 시니어 교육 현장에 적용하여 수료율, 자기효능감, 삶의 만족도 변화를 실증적으로 검증하는 후속 연구를 진행할 예정이다.

참고문헌

[1] 이철현, “소프트웨어정책연구소” 월간SW중심사회 포커스 3 4차 산업혁명 시대 디지털 역량 함양을 위한 소고. Vol 11, pp. 52-61, November 2020.

[2] 이성숙 “국가데이터처 사회통계기획과” 2025 고령자 통계. September 2025.

[3] 이원주, “소프트웨어정책연구소” 월간SW중심사회 포커스 4 효과적인 대학의 SW·AI 교육 개선 방향 제안. Vol 11, pp. 62-69, November 2020.

[4] 권대봉 외 4, “박영스토리” 평생교육의 세가지 지평. ISBN : 9791187010623.

[5] 이승미 외 3, “한국정보기술학회” 노인의 디지털 리터러시 향상을 위한 코딩교육 프로그램이 노인 코딩 흥미도 및 효능감에 미치는 영향. Vol.22, pp 223-233, Jan 2024.

[6] 알베르트 반두라 저 / 윤운성 역, “지학사” 변화하는 사회 속에서의 자기효능감. ISBN : 9788975487255.

and access to medical care," Journal of Health and Social Behavior, Vol. 36, No. 1, pp. 1-10, 1995.
[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of ACM SIGKDD, pp. 785-794, 2016.
[11] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," Advances in Neural Information Processing Systems, Vol. 30, pp. 3146-3154, 2017.
[12] T. Akiba et al., "Optuna: A next-generation hyperparameter optimization framework," Proceedings of ACM SIGKDD, pp. 2623-2631, 2019.
[13] Health Insurance Review & Assessment Service, 2023 (2nd) Depression Outpatient Quality Assessment Results, Wonju: HIRA, 2024.
[14] Ministry of Health and Welfare, 2021 Survey of Mental Disorders in Korea, Sejong: MOHW, 2021.

전통시장 활성화를 위한 가상 방문객 기반 유동인구 최적화 시뮬레이션 연구

전이현*, 성낙준*
*한신대학교 AI·SW대학
e-mail: {komosjs44, njsung}@hs.ac.kr

A Study on Digital Visitor Twin-based Pedestrian Flow Optimization Simulation for Revitalizing Traditional Markets

Yi-hyun Jeon*, Nak-Jun Sung*
* School of Computing and Artificial Intelligence, Hanshin University

요 약

연구 배경 및 목적: 본 연구는 전통시장 활성화를 위해 실측 데이터에 기반한 가상 방문객(Digital Visitor Twin) 모델을 제안한다. 2022년 오산시 오색시장의 유동인구, 공영주차장 이용량, 기상 조건을 탐색적으로 검토하여 외부 환경 변화에 따른 방문객의 반응 특성을 규명하고자 하였다. 연구 방법: 방문객을 단일 집단이 아닌 차량 접근성 의존도와 환경 조건(기온, 장날 여부)에 따른 가상 Twin으로 모델링하였다. 특히 장날 주기(3·8일)와 기상 조건의 상호작용이 유동인구에 미치는 비선형적 영향을 시뮬레이션에 반영하였다. 주요 결과: ① 주차 접근성: 공영주차장 이용량과 방문객 수 사이에서 약한 양의 상관관계(r=0.27)가 관찰되었으며, 주차 접근성이 방문객 유입과 일정 수준의 관련 가능성을 시사하였다. ② 기상 역설(Weather Paradox): 영하권 한파 속에서도 장날에는 평균 37,001명이 유입되어(2022년 1~6월 기준), 전체 평일 평균(31,427명) 대비 약 17.7% 높은 유동인구가 발생하는 환경 저항성 현상을 도출하였다. 결론 및 시사점: 구축된 Digital Visitor Twin 모델은 주차 확충 및 이벤트 개최 등 정책 시나리오별 유동인구 변화를 탐색적으로 시뮬레이션하였다. 이는 전통시장 활성화를 위한 정책 수립 시, 기상 조건과 장날 주기를 결합한 데이터 기반 사전 검증 도구로서 활용 가능성을 지닌다.

키워드: Digital Visitor Twin, 전통시장, 유동인구, 주차 접근성, 환경 저항성

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

I. 서 론

전통시장은 지역 상권의 핵심 축으로 기능해 왔으나, 최근 대형 유통시설의 확산과 온라인·모바일 기반 소비의 급증으로 인해 구조적인 방문객 감소 문제에 직면하고 있다. 특히 지역 기반 전통시장은 상권의 물리적 범위가 제한적이며, 방문객의 접근 방식과 방문 빈도가 매출에 직접적인 영향을 미치는 만큼, 외부 환경 변화에 더욱 취약한 구조를 지닌다.

이러한 상황에도 불구하고, 기존의 전통시장 활성화 정책은 축제 개최, 환경 개선, 홍보 강화와 같은 사후적·경험적 접근에 집중되어 왔다. 이러한 정책들은 시행 이후의 성과를 정성적으로 평가하는 데에는 의미가 있으나, 정책 시행 이전에 예상 효과를 정량적으로 검증하거나 복수의 정책 대안을 비교·평가하는 데에는 구조적 한계를 가진다.

한편, 최근 도시 교통·에너지·스마트시티 분야에서는 디지털 트윈(Digital Twin) 솔루션이 정책 시뮬레이션 및 의사결정 지원 도구로 활발히 활용되고 있다. 디지털 트윈은 실측 데이터를 기반으로 현실 시스템의 구조와 행위 규칙을 가상 공간에 복제하고, 정책 변화나 외부 환경 변화가 시스템에 미치는 영향을 사전에 반복적으로 실험할 수 있다는 강점을 지닌다.

기존의 유동인구 예측이나 회귀 분석이 특정 시점의 관계성 설명에 초점을 둔다면, 디지털 트윈은 '정책 개입 → 행태 변화 → 결과'의 과정을 하나의 연속적 시스템으로 다룬다는 점에서 구조적 차별성을 갖는다. 즉, 단순 예측 도구를 넘어 정책의 인과적 메커니즘을 시뮬레이션할 수 있는 의사결정 지원 체계로 기능한다 [1].

본 연구에서는 이러한 디지털 트윈 개념을 전통시장 방문객 분석에 적용하여, 오산시의 전통시장인 오색시장을 대상으로 가상 방문객(Digital Visitor Twin) 모델을 구축한다. 오색시장의 방문자 데이터를 기반으로 기상 조건과 장날 주기가 결합될 때 나타나는 방문객 반응의 비선형적 특성인 환경 저항성을 분석하고, Digital Visitor Twin이 정책 사전 검증 도구로 활용될 수 있음을 실증하는 것을 목표로 한다.

II. 선행 연구

2.1 전통시장 유동인구 분석 연구

전통시장 유동인구에 관한 기존 연구들은 주로 설문조사나 통계청 자료를 활용하여 방문 빈도, 소비 성향, 만족도 등을 분석해 왔다. 이러한 연구는 방문객의 인식과 태도를 파악하는 데에는 유용하지만, 실제 방문객 수의 일별 변동성이나 정책 개입 효과를 정량적으로 검증하는 데에는 한계가 있다 [2].

최근에는 통신 데이터나 카드 사용 데이터를 활용한 유동인구 분석 연구가 증가하고 있으나, 전통시장 단위에서 장기간의 실측 데이터를 활용하여 정책 시뮬레이션으로까지 확장한 연구는 여전히 제한적이다. 본 연구는 오산시 제공 일별 유동인구 실측 데이터를 활용하여 방문객 수 변화를 정량적으로 분석하고, 이를 정책 시나리오 평가로 확장한다는 점에서 기존 연구와 차별성을 갖는다 [3-4].

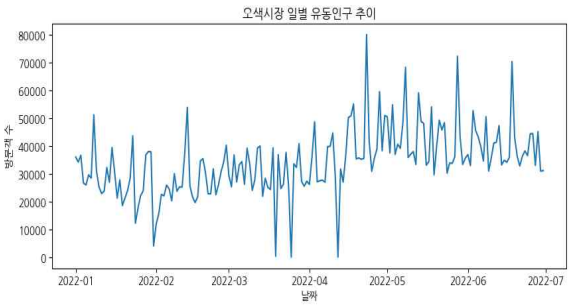


그림1. 오색시장 유동인구 추이

2.2 접근성, 주차, 그리고 방문객 유입

상업시설의 접근성과 관련된 다수의 연구에서는, 주차 공간의 확보 여부와 수용 능력이 방문객 유입에 중요한 영향을 미친다는 결과를 제시하고 있다. 특히 차량 이용 비율이 높은 지역 상권의 경우, 주차 접근성은 방문 결정에 영향을 미치는 중요한 요인 중 하나로 작용한다.

이에 본 연구에서는 공영주차장 입차 건수를 차량 접근성의 대리 변수(proxy variable)로 설정하였다. 실증 분석 결과, 공영주차장 이용량과 오색시장 방문객 수 사이에서 약한 양의 상관관계($r = 0.272$)가 관찰되었으며, 주차 접근성이 방문객 유입과 일정 수준의 관련 가능성을 시사하였다. 다만 이는 약한 상관에 해당하므로 단독으로 인과 관계를 주장하기는 어려우며, 장날 여부·요일·기상 조건 등 추가 변수에 대한 고려가 필요하다.

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

2.3 기상 요인과 상권 활동

기존 상권 분석 연구에서는 일반적으로 기온이 쾌적할수록 유동인구가 증가하는 경향이 있는 것으로 알려져 있다. 그러나 이러한 연구는 주로 대형 상업시설이나 도심 상권을 대상으로 하며, 정기적 이벤트(장날)가 존재하는 전통시장을 대상으로 한 분석은 상대적으로 부족하다.

이러한 연구의 공백을 보완하기 위해, 본 연구는 기상 변수와 장날 주기를 결합하여 분석함으로써, 전통시장의 기능적·문화적 방문 목적성이 기상 조건의 영향을 어떻게 상쇄하는지를 실증적으로 분석한다. 이는 기상 조건과 장날 변수를 독립적으로 다룬 기존 연구와 구별되는 본 연구의 주요 기여점이다.

III. 데이터 및 연구 방법

3.1 데이터 구성

본 연구에서는 다음의 세 가지 실측 데이터를 통합하여 활용하였다. 단, 분석 기간이 2022년 상반기(1~6월)로 한정되어 있어, 이후 절에서 도출되는 결과는 해당 기간의 특성을 반영한 것임을 유의해야 한다.

1) 유동인구 데이터: 오산시 오색시장 일대의 일별 방문객 수 (2022.01.01~2022.06.30, 2022년 상반기)
2) 공영주차장 이용 데이터: 오색시장 인근 공영주차장의 입차 시각 데이터 및 일별 입차 건수
3) 기상 데이터: 기상청 제공 2022년 일별 평균기온 데이터(수원 관측소)

모든 데이터는 날짜 기준으로 병합되었으며, 3일 혹은 8일이 포함되는 경우를 장날로 정의하였다.

3.2 가상 방문객(Digital Visitor Twin) 정의

본 연구에서는 방문객을 단일한 동질 집단으로 가정하지 않고, 접근 방식에 따른 이질적 반응 구조를 반영하기 위해 가상 방문객 트윈 모형을 다음과 같이 정의하였다.

1) 주차 의존형 Twin: 공영주차장 이용량이 상대적으로 높은 날에 주로 유입되는 방문객 집단
2) 비주차형 Twin: 도보 또는 대중교통 접근 비중이 높은 방문객 집단

Twin 분류 기준은 공영주차장 이용량의 중앙값(median)을 기준으로 설정하였다. 사회 데이터의 특성상 명확한 군집 경계가 존재하지 않는 경우가 많으므로, 데이터 불균형을 최소화하고 해석 가능성을 높이기 위해 중앙값 기반 이분 분류를 적용하였다. 다만 이 역시 탐색적 기준에 해당하며, 향후 연구에서는 보다 엄밀한 군집 분석 기법을 적용하여 기준의 타당성을 검증할 필요가 있다.

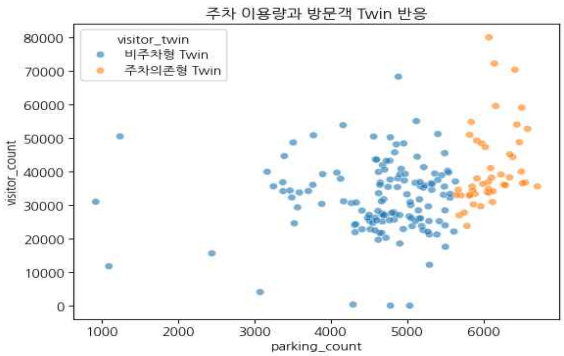


그림 2. 주차 이용량과 시장 방문객 Twin 반응

IV. 실험 결과

4.1 주차 접근성과 방문객 반응

분석 결과, 주차 이용량이 높은 날일수록 전체 방문객 수가 증가하는 경향이 확인되었다. 특히 주차 의존형 Twin은 주차 수용 능력 변화에 민감하게 반응하는 것으로 나타났으며, 이는 주차 정책이 유동인구 확대에 일정한 영향을 미칠 가능성을 시사한다. 다만 공영주차장 이용량과 방문객 수의 상관계수($r = 0.272$)는 약한 양의 상관 수준으로, 주차 접근성이 방문객 유입과 관련된 변수 중 하나일 수 있음을 시사하는 것으로 해석하는 것이 적절하다.

4.2 기상 변수와 장날의 상호작용: '환경 저항성' 분석

기온과 방문객 수의 관계를 전체 데이터 기준으로 분석한 결과, 기온 상승에 따라 방문객 수가 완만하게 증가하는 일반적인 경향이 확인되었으며, 이는 기존 상권 분석 연구의 결과와 일치한다.

그러나 장날 여부를 고려하여 분석한 결과, 기

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

온 변화에 대한 방문객 반응 구조가 근본적으로 달라지는 특이점이 관찰되었다. 장날에는 기온 하락에도 불구하고 방문객 수가 감소하지 않거나 오히려 증가하는 경향이 나타났으며, 이는 기상 조건이 방문 결정에 미치는 영향이 장날에는 현저히 약화됨을 의미한다.

특히 평균기온 0℃ 이하의 한파 구간에서, 장날의 평균 방문객 수는 37,001명으로 전체 평일 평균(31,427명) 대비 약 17.7% 높은 수준을 기록하였다(이상 2022년 1~6월 상반기 데이터 기준). 본 결과는 해당 기간에 한정된 것으로, 연도별·지역별 일반화에는 추가 검증이 필요하다. 그럼에도 장날이라는 조건이 저온 환경에서의 방문 억제 효과를 상쇄하고 있음을 보여준다는 점에서 시사점을 갖는다.

본 연구는 이와 같이 불리한 기상 조건에도 방문객 유입이 유지·강화되는 현상을, 전통시장 방문객이 외부 환경 변화에 대해 일정 수준의 내성을 가지는 행태적 특성으로 해석하고, 이를 환경 저항성(Environmental Resilience)으로 정의한다.

그림 3. 장날 변수 생성

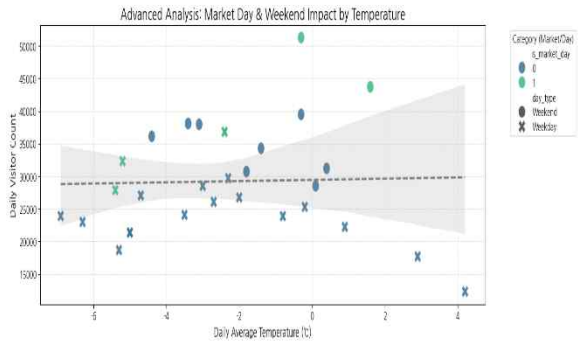


그림 4. 장날/평시와 주말/평일을 결합한 분석

4.2.1 요일·장날·방문객 Twin 조건 결합에 따른 반응 구조 분석

분석 결과, 방문객 반응은 단일 변수에 의해 설명되기보다는 장날 여부를 중심으로 요일과 접근성 조건이 결합된 계층적 구조를 가지는 것으로 나타났다. 요일 효과만을 고려할 경우 주말에 방문객 수가 증가하는 일반적인 경향이 관찰되었으나, 장날이 포함될 경우 해당 효과는 크게 증폭되었다. 특히 평일이라도 장날에는 비장날 주말 수준을 상회하는 방문객 유입이 발생하였다.

방문객 Twin 유형별로는, 주차 의존형 Twin이 장날 조건에서 요일에 따른 변동 폭이 크게 확대된 반면, 비주차형 Twin은 상대적으로 완만한 증가 양상을 보였다. 종합하면 오색시장 방문객 반응은 '장날 여부 → 접근성(Twin) → 요일' 순으로 조건부 분화되는 계층적 구조를 가진다.

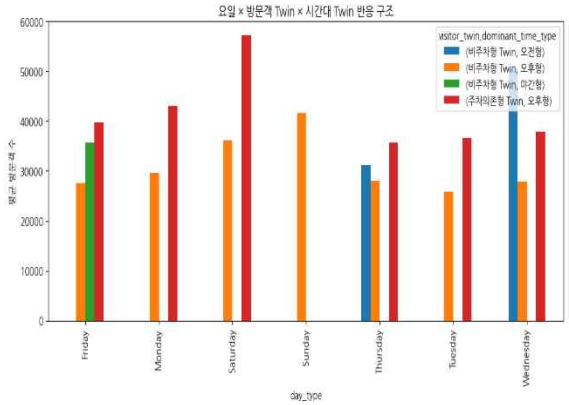


그림 5. 요일·방문객Twin·시간대 Twin 반응 구조

4.3 정책 시나리오 기반 시뮬레이션

Digital Visitor Twin 모델을 활용하여 다음 세 가지 정책 시나리오를 시뮬레이션하였다: ① 주차 수용 능력 20% 확충, ② 이벤트 개최(방문 유인 효과 30% 증가 가정), ③ 주차 확충 + 이벤트 병행.

이벤트 증가율 30%는 기존 전통시장 활성화 사례에서 관측되는 단기 방문 증가 범위를 참고하여 설정한 보수적 가정값이다. 다만 이는 문헌에서 인용된 엄밀한 수치가 아닌 탐색적 가정이므로, 실제 정책 적용 시에는 대상 시장의 특성에 맞는 실증 검증이 선행되어야 한다.

탐색적 시뮬레이션 결과, 복합 정책 시나리오(③)에서 가장 높은 유동인구 증가 효과가 나타났으며, 주차 여건 개선 없이 이벤트만 적용할 경우 특정 시간대 혼잡이 증가하며 방문 효율이 저하되는

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

현상이 관찰되었다. 다만 시나리오 수치(+20%, +30%)는 탐색적 가정값으로, 실제 정책 효과의 정량적 예측보다는 가능성 탐색으로 해석하는 것이 적절하다.

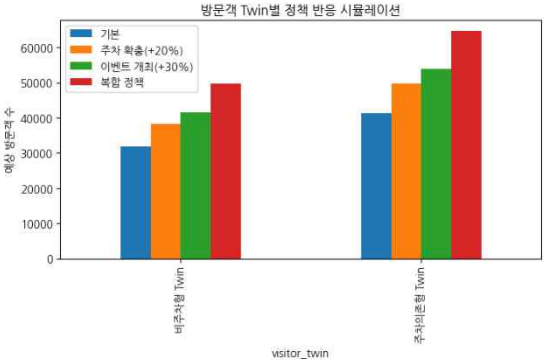


그림 6. 방문객 Twin별 정책 반응 시뮬레이션

V. 결론 및 정책 제언

5.1 연구 결과 요약

본 연구는 실측 유동인구·주차 데이터를 중심으로 기상 조건을 탐색적으로 검토한 Digital Visitor Twin 모델을 통해 오색시장 방문객 행태를 분석하였다. 분석 결과, 공영주차장 이용량은 방문객 수와 약한 양의 상관관계(r=0.272)를 보여 차량 접근성이 방문객 유입과 관련될 가능성을 시사하였으며, 장날에는 기상 조건의 악화에도 불구하고 유동인구가 유지·증가하는 환경 저항성이 관찰되었다.

5.2 정책적 시사점

첫째, 기상 연동형 장날 운영 전략이 필요하다. 한파나 폭염 시에도 장날 유동인구는 안정적으로 유지되므로, 해당 시점의 주차 관리·쉼터 제공 등 인프라 대응이 중요하다.

둘째, Digital Visitor Twin 기반 시뮬레이션은 정책 시행 전 혼잡도와 효과를 예측하는 의사결정 지원 도구로 활용될 수 있다. 단, 모델의 예측 정확도를 높이기 위해서는 보다 많은 연도의 데이터와 다양한 변수 통합이 필요하다.

셋째, 단일 정책보다는 접근성 개선과 방문 유인 정책을 결합한 복합 전략이 효과적임을 시뮬레이션 결과가 보여준다.

5.3 연구의 한계 및 향후 연구

본 연구는 2022년 상반기(1~6월) 데이터와 오산시 오색시장이라는 특정 지역에 한정된 분석으로, 계절적 편향(하계 미포함) 및 지역 특수성으로 인해 결과의 일반화에는 한계가 있다. 또한 r=0.272의 상관계수는 약한 양의 상관 수준으로, 주차 접근성의 영향력을 과도하게 해석하지 않도록 주의가 필요하다. 향후 연구에서는 연간 전체 데이터 및 강수량·미세먼지·대중교통 변수를 추가하고, 다양한 전통시장을 대상으로 모델의 범용성을 검증할 필요가 있다.

참 고 문 헌

[1] Digital Twin 개념 및 정책 활용 사례, 한국지능정보사회진흥원(NIA), https://www.nia.or.kr/common/board/Download.do?bcIdx=27643&cbIdx=66361&fileNo=1

[2] 김미향, 유재원. 빅데이터를 이용한 전통시장 활성화 분석. 전문경영인연구, 24(1), 331-351. 2021

[3] 오산시, "오색시장 및 인근 공영주차장 입출차 실측 데이터셋(2022-2023)", [2025 ABC 해커톤] 주최측 제공 데이터, 2024.

[4] 오산시 빅데이터 포털, "전통시장 유동인구 분석 서비스", 접속일: 2026.06.06. https://www.osan.go.kr/bigdata/gis/analysis/market.do

※본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과 입니다.

실내 포인트 클라우드 노이즈에 강건한 클라우드 서버 기반 3D LiDAR 공간 인식 시스템

박성재*, 김태훈**, 이세훈**, 김남기*, 안준호*
*경기대학교 AI컴퓨터공학부
**경기대학교 컴퓨터과학과

e-mail:{sungjae1357, ngkim, jha}@kyonggi.ac.kr*
e-mail:{201910538, redkye2}@kyonggi.ac.kr**

Cloud Server-Based 3D LiDAR Spatial Recognition System Robust to Indoor Point Cloud Noise

Park Sungjae*, Kim Taehoon**, Lee Sehun**, Kim Namgi*, Ahn Junho*
*Dept. of AI Computer Engineering, Kyonggi University
**Dept. of Computer Science, Kyonggi University

요 약

본 논문은 클라우드 서버 기반 처리와 클라우드 노이즈 강건성을 고려한 LiDAR 기반 실내 공간 인식 시스템을 제안한다. 실내 환경은 반복 구조, 협소 공간, 동적 객체, 누적 드리프트 등으로 인해 안정적인 공간 인식에 제약이 존재하며, 카메라 기반 실내 공간 인식은 개인정보 침해 가능성을 내포한다. 이에 본 연구는 LiDAR 포인트 클라우드와 IMU 데이터를 클라우드 서버에서 통합 처리하고, FAST-LIO2 기반 오도메트리와 노이즈 강건화 절차를 결합한 3차원 공간 인식 구조를 설계하였다. 제안 시스템은 정합 불안정, 누적 오차, 지도 품질 저하 문제에 대응하면서 프라이버시 보호형 센싱 구조를 유지하는 통합적 3D 공간 인식 프레임워크를 제시한다.

키워드: 3D LiDAR, Indoor environment, Spatial recognition, Point cloud, Noise robustness

I. 서론

배송, 안내, 순찰, 돌봄 등 다양한 서비스 분야로 실내 로봇의 활용 범위가 확대되고 있다. 이에 따라 주변 환경을 정확하게 인식하고 안정적으로

※ This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ICAN(ICT Challenge and Advanced Network of HRD) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2024-00436954)
※ This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(RS-2026-25489409)

모델링할 수 있는 공간 인식 기술이 요구된다.

특히 실내 환경은 복잡한 구조와 빈번한 동적 객체 개입으로 인해 신뢰성 있는 공간 정보 확보에 한계가 존재한다. 또한 카메라 기반 인식 방식은 풍부한 시각 정보를 제공하지만, 얼굴과 행동 정보 등 식별 가능한 정보를 포함할 수 있어 개인정보 보호 측면의 고려가 필요하다. 반면 LiDAR는 거리 기반 포인트 클라우드 데이터를 생성하므로 카메라 기반 방식에 비해 직접 식별 정보의 포함 가능성이 억제되며, 3차원 기하 정보 표현과 조명 변화에 대한 낮은 민감도를 바탕으로 실내 공간 인식에 유리하다 [1].

본 연구는 LiDAR 포인트 클라우드와 IMU 데이터를 클라우드 서버에서 통합 처리하여, 개인정보 보호를 고려하면서도 노이즈에 강건한 실내 공간 인식 시스템을 설계하는 것을 목적으로 한다. 이를 위해 LiDAR 기반 실내 공간 인식의 필요성을 제시하고, 실내 노이즈 요인에 대응하는 강건화 절차를 포함한 3차원 공간 인식 시스템을 제안하며, 제안 시스템의 통합적 의의를 고찰한다.

II. 관련 연구

2.1 실내 3D LiDAR 활용 연구

실내 환경에서의 LiDAR 활용 연구는 초기의 사람 감지·추적에서 시작되어, 최근에는 구조 변화 탐지와 실내 3차원 모델링 등 다양한 응용으로 확장되고 있다. LiDAR는 거리 기반 포인트 클라우드를 생성하므로 카메라 기반 센서보다 직접 식별 정보의 포함 가능성이 억제되어 프라이버시 친화적 실내 인식 수단으로 평가된다. 기존 연구는 이러한 특성을 바탕으로 사람 감지, 구조 변화 탐지, 실내 지도화 등 다양한 실내 응용에서 LiDAR의 활용 가능성을 보여주었다 [2]. 그러나 상당수 연구는 정적 센서나 제한된 공간에서의 국소 인식에 초점을 두고 있어, 실내 전체 공간을 연속적이고 통합적으로 인식하는 데에는 한계가 있다.

2.2 3D LiDAR SLAM 및 강건화 연구

이러한 한계를 보완하기 위해 LiDAR 기반 SLAM은 LOAM [3] 이후 LiDAR-IMU 결합과 직접 정합 기반 구조로 발전해 왔다. 특히 LIO-SAM은 팩터 그래프와 루프 클로저를 통해 전역 일관성을 강화하였고 [4], FAST-LIO2는 직접 정합과 반복 확장 칼만 필터를 바탕으로 실시간성과 강건성을 높였다 [5]. 또한 반복 구조와 협소 공간에서의 정합 불안정을 완화하기 위한 방법이 제시되었으며 [6], 동적 객체 제거를 통해 지도 품질 저하를 억제하는 효과가 검증되었다 [7]. 그러나 기존 연구는 주로 개별

그림 1. 클라우드 기반 LiDAR 실내 공간 인식

문제 해결에 초점을 두고 있으며, 프라이버시 보호형 LiDAR 센싱 및 클라우드 기반 처리 구조를 전제로 실내 환경의 복합적 노이즈 요인을 통합적으로 처리하는 시스템 수준의 접근은 아직 충분히 다루어지지 않고 있다.

、

III. 제안방법

3.1 시스템 개요

본 연구는 LiDAR 센서와 IMU로부터 획득한

※ This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ICAN(ICT Challenge and Advanced Network of HRD) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2024-00436954)
※ This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(RS-2026-25489409)

데이터를 클라우드 서버에서 통합 처리하여, 개인정보를 보호하면서도 실내 환경의 주요 노이즈 요인에 강건한 3차원 공간 인식 시스템을 제안한다.

그림 1은 제안 시스템의 전체 아키텍처를 나타낸다. 제안 시스템은 데이터 획득, 클라우드 전송, 전처리, 전단 오도메트리, 노이즈 강건화, 전역 공간 모델링 및 공간 인식 단계로 구성된다. 이동 플랫폼은 실내 주행 과정에서 LiDAR를 통해 거리 정보를, IMU를 통해 자세 및 운동 정보를 수집하며, 이를 서버로 전송한다. 서버에서는 시간 정렬, 모션 왜곡 보정, 포인트 클라우드 필터링을 수행하여 입력 데이터를 정제한다. 이후 FAST-LIO2 기반 전단 오도메트리를 적용하여 실시간 위치를 추정하고 지역 지도를 생성한다. 그 다음 단계에서는 반복 구조, 협소 공간, 누적 드리프트, 동적 객체와 같이 실내 환경에서 빈번하게 발생하는 노이즈 요인에 대응하기 위한 강건화 모듈을 결합하여 전역 3차원 지도를 구축한다. 최종적으로 생성된 전역 포인트 클라우드 맵을 기반으로 복도, 개방 공간, 협소 구간, 가구 배치와 같은 실내 공간의 구조적 특성을 인식한다.

3.2 노이즈 강건형 클라우드 공간 인식 시스템

제안 시스템의 공간 인식 시스템은 다음과 같다. 첫째, LiDAR와 IMU로부터 수집된 데이터를 클라우드 서버로 전송한 후, 시간 동기화와 모션 왜곡 보정을 수행하여 센서 간 불일치와 이동 중 발생하는 스캔 왜곡을 최소화한다. 둘째, 전처리된 데이터를 기반으로 FAST-LIO2를 적용하여 실시간 오도메트리와 맵을 생성한다.

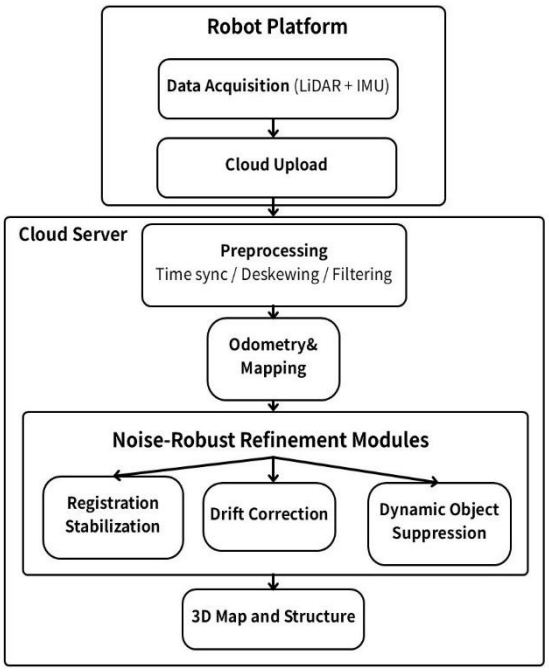


그림 1. 클라우드 기반 LiDAR 실내 공간 인식 시스템 아키텍처

FAST-LIO2는 직접 정합과 IMU 결합을 기반으로 하므로, 특징점이 부족하거나 구조가 단조로운 실내 환경에서도 비교적 안정적인 위치 추정을 수행한다. 셋째, 반복 구조와 협소 공간에서 발생하는 오정합 문제를 완화하기 위해 정합 안정화 전략을 반영한다. 넷째, 장거리 이동에 따라 누적되는 오차는 전역 최적화와 루프 클로저 보완 절차를 적용하여 보정한다. 이를 통해 지역적으로 축적된 오차를 전역 수준에서 완화하고 지도의 일관성을 유지한다. 다섯째, 사람이나 이동 물체와 같은 동적 객체로 인해

※ This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ICAN(ICT Challenge and Advanced Network of HRD) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2024-00436954)
※ This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(RS-2026-25489409)

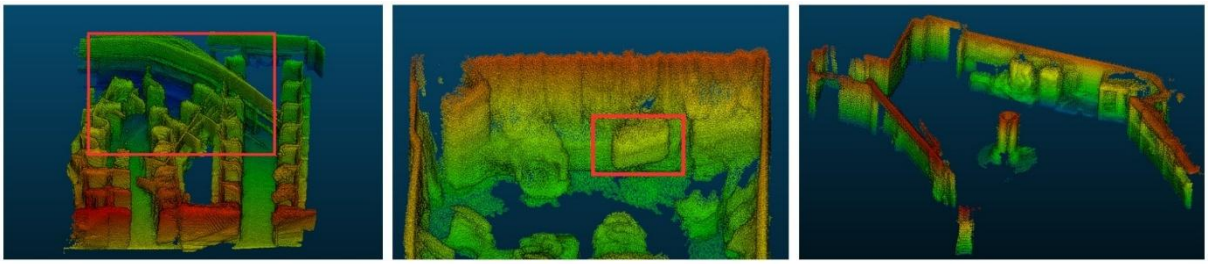


그림 2. 실내 환경에서 발생하는 노이즈 사례(고스트 현상, 누적 드리프트, 저고도 객체 불안정)

발생하는 지도 품질 저하를 줄이기 위해 동적 객체 억제 전략을 적용한다. 이를 통해 정적 구조 중심의 지도를 유지하고, 일시적 객체로 인한 정합 불안정과 지도 왜곡을 완화한다.

그림 2는 실내 환경에서 관찰되는 노이즈 사례를 예시적으로 보여준다. 제안 시스템은 실내 환경에서 관찰되는 고스트 현상, 정합 불안정, 누적 드리프트, 지도 품질 저하 문제를 유형별로 분리하여 대응한다는 점에서, 단일 SLAM 구조만을 사용하는 방식보다 실내 적용성이 높다.

IV. 결론

본 논문에서는 클라우드 기반 처리와 노이즈 강건성을 고려한 LiDAR 기반 실내 공간 인식 시스템을 제안하였다. 이를 위해 실내 LiDAR 활용 연구와 LiDAR SLAM 관련 연구를 검토하고, 반복 구조, 협소 공간, 동적 객체, 누적 드리프트 등 실내 환경의 주요 노이즈 요인을 분석하였다. 이를 바탕으로 LiDAR와 IMU 데이터를 클라우드 서버에서 통합 처리하고, FAST-LIO2 기반 오도메트리와 노이즈 강건화 절차를 결합한 3차원 공간 인식 구조를 설계하였다. 제안 시스템은 프라이버시 보호형 센싱 구조를 유지하면서도 실내 환경의 주요 노이즈 요인에 대응할 수 있는 통합적 설계 방향을 제시하였다는 점에서 의의가 있다.

참고문헌

[1] X. Huang, S. Pan, and M. Bennewitz, "Privacy Risks of Robot Vision: A User Study on Image Modalities and Resolution," arXiv:2505.07766, May 2025.

[2] C. Wen, J. Tan, F. Li, C. Wu, Y. Lin, Z. Wang, and C. Wang, "Cooperative indoor 3D mapping and modeling using LiDAR data," Information Sciences, Vol. 574, No. 2, pp. 192-209, Oct. 2021.

[3] J. Zhang and S. Singh, "LOAM: Lidar Odometry and Mapping in Real-time," Robotics: Science and Systems (RSS), pp. 2-8, July 2014.

[4] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus, "LIO-SAM: Tightly-coupled Lidar Inertial Odometry via Smoothing and Mapping," 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5135-5142, Oct. 2020.

[5] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, "FAST-LIO2: Fast Direct LiDAR-Inertial Odometry,"

※ This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ICAN(ICT Challenge and Advanced Network of HRD) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2024-00436954)

※ This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(RS-2026-25489409)

IEEE Transactions on Robotics, Vol. 38, No. 4, pp. 2053-2073, Aug. 2022.

[6] D. Chung and J. Kim, "NV-LIOM: LiDAR-Inertial Odometry and Mapping Using Normal Vectors Towards Robust SLAM in Multifloor Environments," IEEE Robotics and Automation Letters, Vol. 9, No. 11, pp. 9375-9382, Nov. 2024.

[7] C. Li, W. Li, W. Liu, Y. Shu, and Y. Lou, "FreeDOM: Online Dynamic Object Removal Framework for Static Map Construction Based on Conservative Free Space Estimation," IEEE Robotics and Automation Letters, Vol. 10, No. 6, pp. 5577-5584, June 2025.

※ This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ICAN(ICT Challenge and Advanced Network of HRD) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2024-00436954)

※ This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(RS-2026-25489409)

다기관 연합학습 클라이언트 환경 배포를 위한 가상화 방법

조성진, 정민혁, 이강윤*
가천대학교 컴퓨터공학과
e-mail: lmsj122@gachon.ac.kr, bvnm0121@gachon.ac.kr, keylee@gachon.ac.kr*

A Virtualization Approach for Deploying Client Environments in Multi-Institutional Federated Learning

Seong Jin Jo, Min Hyuk Jung, KangYoon Lee*
Dept. of Computer Engineering, Gachon University

요 약

연합학습은 로컬 데이터를 중앙으로 전송하지 않고도 모델을 학습할 수 있는 방식으로, 연합학습 환경에서는 클라이언트마다 학습 환경이 서로 다를 수 있다. 이러한 환경 이질성은 클라이언트의 학습 참여를 어렵게 하고, 연합학습의 안정적인 수행을 저해한다. 이에 본 연구는 클라이언트 환경의 이질성을 완화하기 위해 가상화를 통한 연합학습 클라이언트 환경 표준화 방법을 제안한다. 본 연구에서는 제안 방법을 구현하기 위해 연합학습 클라이언트 환경과 의존성을 Docker 이미지로 빌드하고, 각 클라이언트가 해당 이미지 기반의 컨테이너에서 학습을 수행하도록 설계하였다. 구현 및 실험 결과, 제안 방식을 통하여 클라이언트가 동일한 컨테이너 환경에서 연합학습 절차를 수행 가능함이 확인되었으며 본 연구는 실제 연합학습 환경에서 클라이언트 실행환경의 이질성을 완화하여 안정적인 학습 수행이 가능한 연합학습 클라이언트 환경을 제시한다.

키워드: 연합학습, 컨테이너 가상화, 클라이언트 실행환경

I. 서론

연합학습(Federated Learning, FL)은 데이터를 중앙 서버로 수집하지 않고, 각 클라이언트의 로컬 데이터에서 모델을 학습한 뒤 데이터 공유 없이 모델 파라미터만을 서버에서 집계하는 분산 학습 방식이다[1][2]. 이러한 특성으로 인해 개인정보와 의료정보처럼 외부 반출이 어려운 데이터를 활용할 수 있는 방법으로 주목받고 있다[3][4].

그러나 실제 환경에서는 클라이언트 간 운영체제, 라이브러리, 하드웨어 차이로 인해 실행 오류가 발

* Corresponding Author

※이 논문은 서울시 산학연 협력사업 2025년도 인공지능 기술사업화 지원사업(CY250165)의 지원을 받아 수행된 연구임

생할 수 있으므로, 다양한 환경에서도 일관되게 동작하는 클라이언트 실행환경이 필요하다.

본 논문에서는 Docker 기반 컨테이너 가상화를 활용한 연합학습 클라이언트 실행환경을 제안한다. 제안 방식은 클라이언트 코드와 실행 의존성을 Docker 이미지로 빌드하고 배포하여 로컬 환경 의존성을 줄이고, 동일한 환경에서 클라이언트가 연합학습에 참여할 수 있도록 한다.

II. 관련 연구

2.1 연합 학습과 클라이언트 환경의 이질성

연합학습은 여러 클라이언트가 로컬 데이터로 모델을 학습하고, 중앙 서버가 각 클라이언트의 모델 파라미터를 집계하여 글로벌 모델을 구성한다. 프라이버시 보호 측면에서 원본 데이터를 중앙 서버로 전송하지 않는다는 이점이 있으나, 실제 환경에서는 클라이언트 간 데이터 분포의 이질성(non-IID)과 이기종 실행 환경, 네트워크 상태, 로컬 연산 자원 등의 요인으로 운영 복잡도가 증가한다[5][6]. 특히 연합학습은 클라이언트 등록, 모델 배포, 로컬 학습, 파라미터 전송, 서버 집계 등 여러 단계가 반복적으로 수행되므로, 특정 클라이언트의 실행 실패나 환경 설정 문제는 전체 학습 과정에 영향을 줄 수 있다. 따라서 실제 연합학습에서는 모델 성능뿐 아니라 안정적인 클라이언트 실행 인프라와 운영 관리 체계가 함께 고려되어야 한다[7].

2.2 가상화 기술과 도커

가상화는 물리적 자원을 소프트웨어적으로 추상화함으로써 단일 하드웨어 위에서 복수의 독립적인 실행 환경을 구성할 수 있게 하는 기술이다. 이를 통해 각 실행환경은 CPU, 메모리, 스토리지, 네트워크 인터페이스 등의 자원을 논리적으로 할당받아 상호 간섭 없이 격리된 상태로 운용된다[8].

VM 기반 가상화는 운영체제 전체를 가상화하여 높은 격리성을 제공하지만 자원 사용량이 크고, 컨테이너 기반 가상화는 호스트 운영체제를 공유하면서 애플리케이션 실행환경만을 격리하므로 상대적으로 가볍다[9]. Docker는 애플리케이션과 실행 의존성을 이미지 형태로 빌드하고 배포하는 방식에 있어 널리 활용되는 컨테이너 플랫폼으로, 서로 다른 시스템에서도 동일한 실행환경을 구성하는 데 활용된다[10].

III. 제안방법

3.1 시스템 개요

본 연구는 클라이언트 실행환경의 이질성을 완화하기 위해 Docker 기반 컨테이너 실행 방식을 제안한다. 본 연구에서는 연합학습 클라이언트 코드와 실행 의존성을 Docker 이미지로 빌드하여, 각 클라이언트가 동일한 컨테이너 환경에서 연합학습에 참여하도록 한다.

3.2 기존 방식의 문제점

기존 방식에서는 각 참여자가 자신의 로컬 시스템에 실행환경을 직접 구성한 뒤 클라이언트로서 연합학습 참여를 시도한다. [그림 1]과 같이 클라이언트들이 서로 다른 이질적인 환경에서 실행될 경우, 동일한 코드라도 환경에 따라 정상적으로 동작하지 않을 수 있으며, 참여자가 제공된 매뉴얼을 정확히 수행하지 못하면 전체 연합학습의 지연 또는 중단으로 이어질 수 있다.

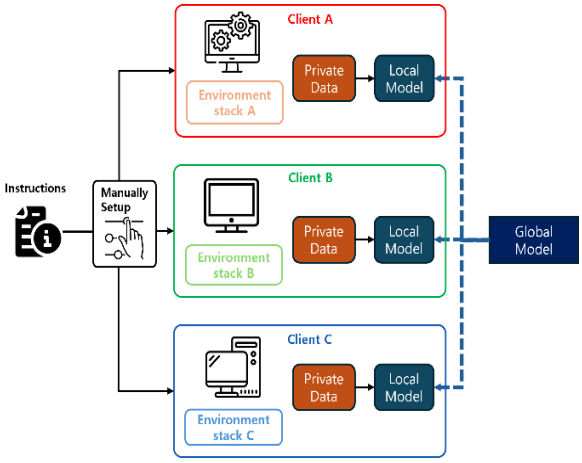


그림 1. 기존 클라이언트 실행 방식

3.3 가상화 기반 클라이언트 실행 방식

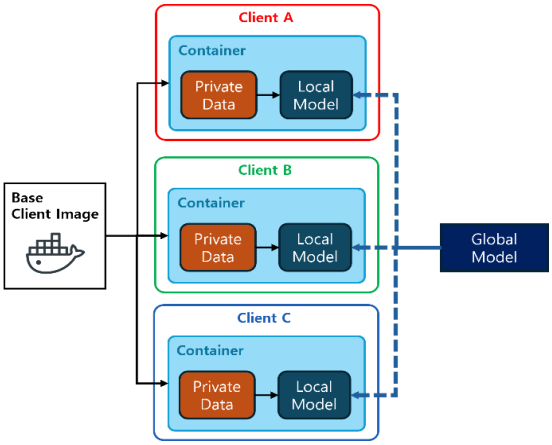


그림 2. 가상화 기반 클라이언트 실행 방식

본 연구에서는 연합학습 클라이언트 실행에 필요한 환경을 Docker 이미지로 구성하고, 각 참여자가 해당 이미지를 기반으로 컨테이너를 실행하도록 하였다. [그림 2]의 Base Client Image는 모든 클라이언트가 공통으로 사용하는 실행 환경을 포함한다. 각 참여자는 호스트 환경과 완전히 격리된 독립적인 컨테이너 내부에서 연합학습 클라이언트를 구동함으로서, 로컬 시스템의 소프트웨어 구성과 무관하게 일관된 실행 환경을 보장받는다.

IV. 실험 및 구현

본 장에서는 호스트에서 직접 환경을 구축하는 기존 방식과 본 연구에서 제안한 가상화 기반 방식을 각각 사용하여 연합학습을 수행하고, 두 방식의 학습 결과를 비교한다. 실험의 목적은 가상화한 실행환경이 기존 방식과 비교하여 학습 성능에 큰 영향을 주지 않으면서도, 클라이언트 실행환경을 일관되게 구성할 수 있음을 확인하는 것이다.

4.1 실험 환경

기존 방식과 가상화 기반 방식의 실험 환경 구성은 각각 그림 3과 그림 4에 나타내었다. 그림 3의 기존 방식에서는 서버 1, 2, 3의 호스트 환경에 연합학습 실행환경을 직접 구성하고, 각 서버에서 클라이언트를 실행한다. 반면 그림 4의 컨테이너 기반 방식에서는 동일한 Docker 이미지를 기반으로 생성된 컨테이너 내부에서 각 클라이언트를 실행한다.

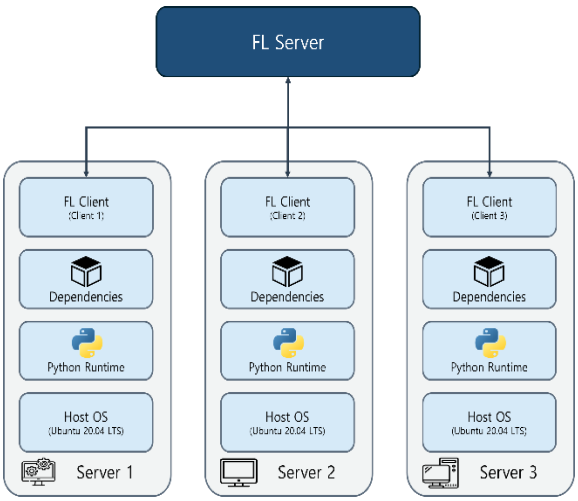


그림 3. 기존 방식의 실험 환경 구성

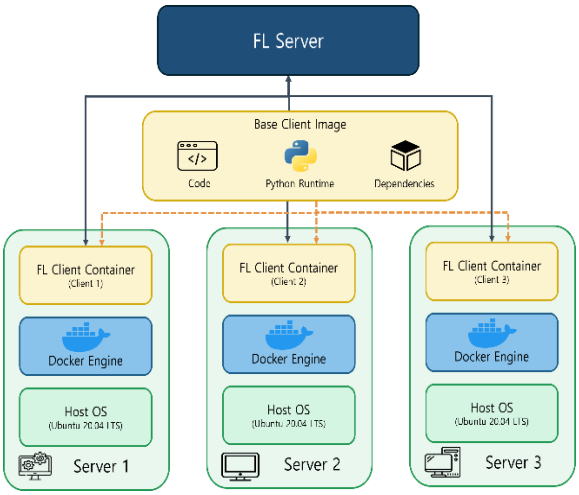


그림 4. 가상화 기반 방식의 실험 환경 구성

본 실험은 세 개의 클라이언트 환경에서 수행하였으며, 각 클라이언트는 서버 1, 2, 3에서 실행하였다. 실험에 사용한 하드웨어 및 운영체제 환경은 [표 1]과 같다.

실험에는 MNIST 데이터셋을 사용하였다. MNIST는 0부터 9까지의 손글씨 숫자 이미지로 구성되며, 60,000개의 학습 데이터와 10,000개의 테스트 데이터로 이루어진다[11]. 본 구현에서는 MNIST 학습 데이터를 학습, 검증, 로컬 테스트 용도로 각각 70%, 10%, 20% 비율로 분할하였다. 글로벌 모델의

최종 성능 평가는 별도로 제공되는 MNIST 테스트 데이터 10,000개를 이용하였다. 모델은 MNIST Classifier를 사용하였으며, 실험은 총 5 round, 각 round당 3 local epoch, batch size 32로 수행하였다. 연합학습 구현은 Flower 프레임워크를 기반으로 하였다[12].

표 1. 실험 환경

Client	GPU	CPU	RAM	OS
Server 1	NVIDIA GeForce RTX 2080 Ti	Intel Xeon Silver 4214R @ 2.40GHz	64GB	Ubuntu 20.04 LTS
Server 2	NVIDIA GeForce RTX 3090	Intel Core i9-10980XE @ 3.00GHz	64GB	Ubuntu 20.04 LTS
Server 3	NVIDIA GeForce RTX 3070 Ti	Intel Core i9-10980XE @ 3.00GHz	128GB	Ubuntu 20.04 LTS

4.2 연합학습 수행

두 방식 모두 동일한 연합학습 절차를 수행하였다. 각 클라이언트는 서버로부터 글로벌 모델을 배포 받은 뒤, 자신에게 할당된 MNIST 로컬 데이터로 학습을 수행하고 모델 파라미터를 서버로 전송한다. 서버는 이를 집계하여 글로벌 모델을 갱신하며, 이 과정은 설정된 round 동안 반복된다. 두 방식의 비교를 위해 각 round마다 글로벌 모델의 accuracy와 loss를 측정하였다.

4.3 수행 결과 및 비교

표 2. 기존 방식과 가상화 기반 방식의 모델 성능 비교

	Conventional Way		Virtualized Way	
Round	Accuracy	Loss	Accuracy	Loss
Round 0	0.1146	2.3034	0.1146	2.3034
Round 1	0.9897	0.0326	0.9896	0.0325
Round 2	0.9910	0.0301	0.9897	0.0336
Round 3	0.9928	0.0319	0.9921	0.0294
Round 4	0.9919	0.0381	0.9915	0.0399
Round 5	0.9919	0.0417	0.9925	0.0469

표 2는 기존 방식과 가상화 기반 방식의 round별 accuracy와 loss를 나타낸다. 두 방식 모두 Round 1 이후 accuracy가 0.99 수준으로 크게 증가하였고, loss는 초기값에 비해 크게 감소하였다. 최종 round 기준 accuracy는 기존 방식 0.9919, 가상화 기반 방식 0.9925로 거의 동일했으며, loss 역시 각각 0.0417과 0.0469로 유사하였다. 이는 가상화된 Docker 컨테이너 내부에서 클라이언트를 실행하더라도 기존 방식과 거의 동일한 학습 성능을 유지할 수 있음을 보여준다.

Round 1 이후 loss가 일부 증가하는 현상이 나타났으나 두 방식에서 유사한 경향이 나타났으므로, 이는 가상화로 인한 성능 저하로 볼 수 없다. 다만 본 실험은 세 개의 클라이언트를 사용한 제한적인 환경에서 수행되었으므로, 향후 대규모 환경에서의 추가 검증이 필요하다.

V. 결론

본 연구에서는 컨테이너 가상화를 Docker 기반으로 활용하여 연합학습 클라이언트 실행환경을 일관되게 구성하는 방법을 제안하였다.

실험 결과, 컨테이너 기반 실행 방식은 기존 로컬 실행 방식과 비교하여 accuracy와 loss 측면에서 동등한 수준의 학습 성능을 나타냈으며, 가상화 클라이언트 방식이 모델 수렴에 부정적인 영향을 미치지 않음을 실증적으로 확인하였다. 이는 제안 방식이 학습 성능을 유지하면서도 클라이언트로 하여금 별도의 실행환경 구성 없이 동일한 컨테이너 이미지를 통해 연합학습에 참여할 수 있도록 함을 의미한다.

본 연구는 가상화 기반 실행 환경이 연합학습의 모델 성능을 저해하지 않으면서도 클라이언트 측 환경 이질성 문제를 완화할 수 있음을 확인하였다. 이는 실제 배포 환경에서 참여자의 연합학습 참여 진입 장벽을 낮추고 운영 안정성을 높일 수 있음을 나타내며, 가상화 기반 접근이 실제 연합학습 인프라 구축에 효과적인 방향임을 시사한다.

참고문헌

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data," Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Vol. 54, pp. 1273-1282, Apr. 2017.

[2] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," ACM Transactions on Intelligent Systems and Technology, Vol. 10, No. 2, Article 12, pp. 1-19, Jan. 2019.

[3] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated Learning for Internet of Things: A Comprehensive Survey," IEEE Communications Surveys & Tutorials, Vol. 23, No. 3, pp. 1622-1658, Third Quarter 2021.

[4] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A Survey on Security and Privacy of Federated Learning," Future Generation Computer Systems, Vol. 115, pp. 619-640, Feb. 2021.

[5] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," IEEE Signal Processing Magazine, Vol. 37, No. 3, pp. 50-60, May 2020.

[6] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y. C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated Learning in Mobile Edge Networks: A Comprehensive Survey," IEEE Communications Surveys & Tutorials, Vol. 22, No. 3, pp. 2031-2063, Third Quarter 2020.

[7] J. Moon, S. Yang, and K. Lee, "FedOps: A Platform of Federated Learning Operations With Heterogeneity Management," IEEE Access, Vol. 12, pp. 4301-4314, Jan. 2024.

[8] G. J. Popek and R. P. Goldberg, "Formal Requirements for Virtualizable Third Generation Architectures," Communications of the ACM, Vol. 17, No. 7, pp. 412-421, Jul. 1974.

[9] W. Felter, A. Ferreira, R. Rajamony, and J. Rubio, "An Updated Performance Comparison of Virtual Machines and Linux Containers," 2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), pp. 171-172, Mar. 2015.

[10] C. Boettiger, "An Introduction to Docker for Reproducible Research," ACM SIGOPS Operating Systems Review, Vol. 49, No. 1, pp. 71-79, Jan. 2015.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, Vol. 86, No. 11, pp. 2278-2324, Nov. 1998.

[12] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. de Gusmão, and N. D. Lane, "Flower: A Friendly Federated Learning Research Framework," arXiv preprint arXiv:2007.14390, Jul. 2020.

기사 친화적 배차를 위한 AI 시뮬레이터 설계 : 인간 중심 알고리즘의 진화

김형주 *, 정재영O, 류승택, 강민구, 이주성, 오준희, 신태환
*한신대학교 (정보통신학부), O 경희대학교(프랑스어학과, 외국어대학)
e-mail: hyeongju7111@gmail.com, jjy716@khu.ac.kr

A Design on an AI simulator for Driver-Friendly DIspatching : Evolution and of Human-Centered Algorithms

Kim Hyeongju*, Joung Jaeyoung O, Ryu Seungtaek, Lee Juseong, Oh Junhee, Shin Taehwan
* Dept. of IT Transmedia Contents, Hanshin University
College of Computing and Artificial Intelligence, Hanshin University
O Dept. of French Language, Kyung Hee University
College of Foreign Language & Literature, Kyung Hee University

요 약

본 연구는 전세버스 운행 데이터를 기반으로 기사 친화형 AI 배차 시뮬레이터 모델을 제안한다. 전세버스 배차는 차량과 노선을 단순 연결하는 문제가 아니라 기사 피로도, 출퇴근 동선, 노선 난이도, 근무시간 편차, 반복 배차에 따른 불공정성 인식까지 고려해야 하는 복합 의사결정 문제이다. 제안 모델은 노선별 소요시간, 출발·도착지 거리, 공차거리, 근무시간, 운행 빈도, 기사 선호도, 배차 이력, 민원·불만 기록 등을 변수화하여 노선별 난이도와 기사별 부담도를 산출한다. 또한 도심형·외곽형·단거리 반복형·장거리 통근형·행사형 등 노선군별로 서로 다른 정상 범위와 공정성 기준을 설정하고, 좌절지수 파라미터를 도입하여 특정 기사에게 부담이 집중되는 현상을 완화한다. 제안 프레임워크는 경험 중심 배차를 설명 가능한 데이터 기반 의사결정 체계로 전환하고, 향후 TRP 및 IBMS와 연계한 자동 배차, 분쟁 예방, 성과평가 시스템으로 확장될 수 있다.

키워드: TRP·IBMS 통합 의사결정, 이벤트 로그 기반 데이터 통합, Kanban 업무흐름 관리, 규칙 기반 자동 검증 시스템, 전세버스 모빌리티 데이터 플랫폼

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심대학지원사업(경기 RISE사업)의 연구결과입니다.

그림 1. 기사 친화형 AI 배차 시뮬레이터 개념도

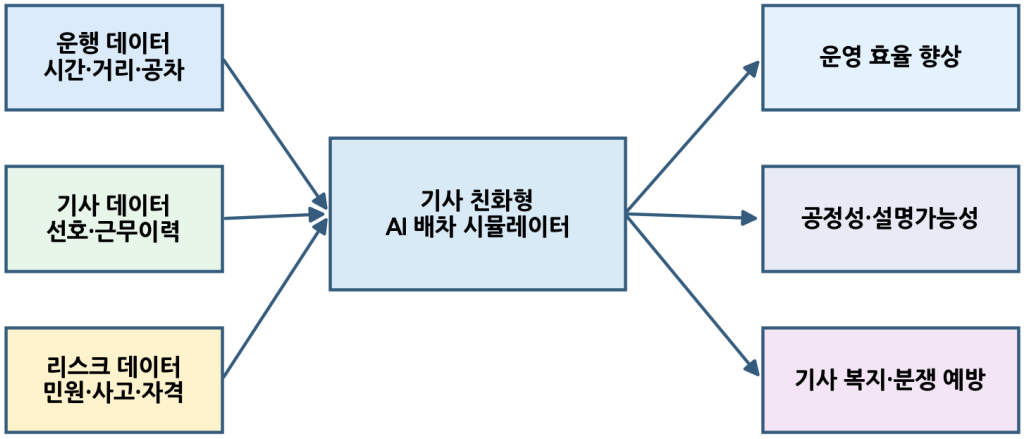


그림 1. 기사 친화형 AI 배차 시뮬레이터 개념도

I. 서론

전세버스 산업의 배차 업무는 일반적인 차량 배정 문제보다 훨씬 복잡하다. 통근·통학·행사·수시 운행이 혼재하고, 각 운행은 출발지와 도착지, 운행 시간대, 대기 시간, 공차 이동, 차량 크기, 기사 숙련도, 고객사 요구조건을 동시에 반영해야 한다. 특히 전세버스 회사에서 배차 결과는 기사 개인의 근무강도, 수입 안정성, 피로도, 조직 내 공정성 인식과 직접 연결되므로 배차는 단순한 운영 효율 문제가 아니라 노무관리, 안전관리, 고객서비스, 분쟁 예방을 동시에 포함하는 핵심 의사결정 영역이다.

그러나 현장의 배차는 아직도 숙련 배차 담당자의 경험, 전화·메신저 기반 조율, 일부 엑셀 기록에 의존하는 경우가 많다. 이러한 방식은 단기적으로 빠른 판단을 가능하게 하지만, 배차 이유를 사후에 설명하기 어렵고 특정 기사에게 좋은 노선 또는 어려운 노선이 반복 배정되었다는 인식이 발생할 수 있다. 또한 노동시간, 대기시간, 공차거리, 회차

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심대학지원사업(경기 RISE사업)의 연구결과입니다.

난이도 등이 누적 데이터로 관리되지 않으면 회사와 기사 간 분쟁이 발생했을 때 객관적 근거를 제시하기 어렵다.

최근 전세버스 산업은 기사 고령화, 신규 인력 유입 감소, 고객사의 시간준수 요구 강화, 안전관리 기준 고도화라는 구조적 변화를 동시에 겪고 있다. 이러한 환경에서는 배차의 효율성만이 아니라 배차 기준의 일관성, 설명 가능성, 그리고 기사 수용성이 중요해진다. 다시 말해 같은 운행을 처리하더라도 어떤 기사에게 왜 배정되었는지, 다른 대안과 비교해 어떤 점에서 합리적인지 기록으로 남길 수 있어야 한다.

본 연구는 전세버스 운행 데이터를 기반으로 한 기사 친화형 배차 최적화 알고리즘을 제안한다. 제안 모델은 단순 비용 최소화나 차량 회전율 극대화가 아니라 기사 복지와 운영 효율을 동시에 고려하는 다목적 최적화 구조를 목표로 한다. 이를 위해 노선 특성, 근무부담, 배차 이력, 기사 선호도, 불만·민원 기록을 정량화하고, 노선군별로 서로 다른 정상 범위와 평가 기준을 설정한다.

또한 본 연구는 단순한 아이디어 제시에 머물지 않고, 실제 업무 시스템과 연계할 수 있는 데이터 구조, 방어 조건, 평가 수식, 운영 절차, 실증 검증 방향까지 함께 제시한다. 이를 통해 학술적 논의뿐 아니라 현장 적용 가능성이 높은 실무 프레임워크로서의 타당성을 확보하고자 한다.

II. 이론적 배경 및 문제 정의

2.1 전세버스 배차의 데이터 구조

전세버스 배차 데이터는 크게 운행 데이터, 차량 데이터, 기사 데이터, 고객사 데이터, 정산 데이터로 구분할 수 있다. 운행 데이터에는 노선명, 출발지, 도착지, 출발시각, 도착시각, 대기시간, 운행거리, 공차거리, 반복주기 등이 포함된다. 차량 데이터에는 차량 연식, 좌석 수, 연료 유형, 정비 상태, 장비 장착 여부가 포함되고, 기사 데이터에는 거주지 또는 출근 기준지, 근무 가능 시간, 면허·교육 이력, 선호 노선, 최근 근무량, 휴무 이력 등이 포함된다.

고객사 데이터에는 운행 시작 시각 준수 여부, 정류장별 민감도, 담당자 응답 방식, 추가 요청 빈도, 운행 변경 가능성이 포함될 수 있다. 정산 데이터에는 운행 단가, 기사 수당, 공차 정산 방식, 추가 대기 발생 여부, 벌점·민원에 따른 비용 변동 등이 포함된다. 이들 데이터는 배차 이후의 성과와 만족도를 설명하는 핵심 근거가 된다.

예를 들어 동일한 2시간 운행이라도 도심 정체 구간인지, 외곽 장거리인지, 출근 시간대인지, 회차 후 다음 운행과 연결되는지에 따라 체감 부담은 크게 달라진다. 따라서 본 연구는 운행 단위의 물리적 변수뿐 아니라 기사 관점의 체감 부담 변수를 함께 고려한다.

2.2 기존 배차 방식의 한계

기존 배차 방식의 핵심 한계는 세 가지이다. 첫째, 배차 기준이 명시적으로 문서화되지 않아 결과의 설명 가능성이 낮다. 둘째, 노선별 난이도와 기사별 누적 부담을 비교할 수 있는 표준 지표가 부족하다. 셋째, 공정성을 동일 배분으로 오해하여 모든 노선을 같은 기준으로 평가하는 문제가 발생한다.

실제로 도심형 노선, 외곽형 노선, 단거리 셔틀, 장거리 통근 노선은 서로 다른 분포와 부담 구조를 가진다. 따라서 공정한 배차란 모든 기사에게 동일한 노선을 배정하는 것이 아니라, 서로 다른 노선 특성과 기사 조건을 반영하여 누적 부담이 지나치게 한쪽으로 치우치지 않도록 관리하는 것이다.

추가로 기존 방식은 예외 상황 처리에 취약하다. 갑작스러운 결근, 차량 고장, 고객사 시간 변경, 기사 교대 필요 상황이 발생하면 배차 담당자의 경험 의존도가 더 커진다. 이때 경험은 중요하지만, 데이터가 축적되지 않으면 비슷한 상황이 재발해도 조직 차원의 학습이 이루어지기 어렵다.

III. 연구모형: 기사 친화형 AI 배차 시뮬레이터

본 연구는 다음 네 가지 연구문제를 설정한다. 첫째, 전세버스 운행의 물리적 특성과 기사 체감 부담을 동시에 반영하는 배차 평가 구조를 어떻게 설계할 것인가. 둘째, 서로 다른 노선군에 대해 동일 기준이 아닌 차등적 공정성 기준을 어떻게 정의할 것인가. 셋째, 누적된 불만 가능성이나 배차 기피도를 반영하는 기사 친화형 보조 지표를 어떻게 구축할 것인가. 넷째, 실제 기업 시스템에 적용 가능한 설명 가능한 추천 절차를 어떻게 구현할 것인가.

이러한 연구문제는 단일 목적의 비용 최소화 모델만으로는 충분히 해결되기 어렵다. 따라서 본 연구는 다목적 평가, 제약 조건 기반 필터링, 누적 지표 관리, 운영자 승인 절차를 통합한 하이브리드 의사결정 구조를 제안한다.

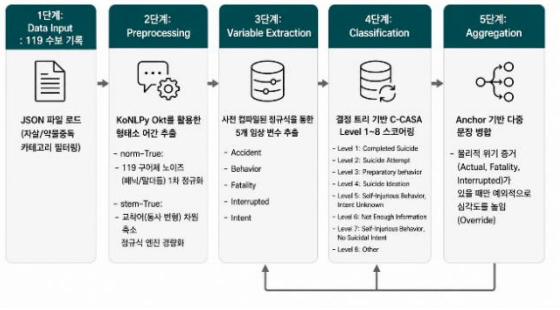


그림 2. 제안하는 C-CASA 기반 119 정신과적 응급신고 맥락분류 자동화 시스템 구조도 제시

본 연구에서 다루는 인공지능(AI)은 규칙 기반 의사결정 규칙과 데이터 기반 학습을 결합한 하이브리드 시스템으로 정의한다. 이는 인간 전문가의 경험적 지식과 관찰 데이터를 바탕으로, 노선 특성, 기사 부담, 공정성, 리스크를 동시에 반영하는 의사결정 보조 도구를 의미한다. 목적은 배차 결정의 자동화 수준을 높이되, 현장 운영자의 판단을

대체하기보다 체계적으로 보조하는 설명 가능한 AI를 구현하는 것이다. 이를 통해 배차의 근거를 기록하고, 책임 있는 의사소통과 분쟁 예방에 기여하고자 한다.

3.1 변수 정의

제안 모델은 배차 후보를 평가하기 위해 운행 효율 점수, 기사 부담 점수, 공정성 점수, 리스크 점수의 네 가지 축을 사용한다. 운행 효율 점수는 차량 회전율, 공차거리, 다음 운행 연결 가능성, 고객사 시간 준수 가능성을 반영한다. 기사 부담 점수는 총 근무시간, 운전시간, 대기시간, 출근·퇴근 동선, 야간·새벽 운행 여부, 연속 근무일 수를 반영한다.

공정성 점수는 최근 일정 기간 동안의 좋은 노선·어려운 노선 배정 비율, 누적 운행시간, 누적 공차 부담, 선호 노선 반영률을 비교한다. 리스크 점수는 민원 발생 가능성, 사고·지각 이력, 면허·교육·음주측정 등 필수 조건 충족 여부를 포함한다.

각 변수는 절대값 자체보다 상대적 위치가 중요하다. 예를 들어 공차거리 20km가 항상 높은 부담을 의미하지는 않는다. 도심 밀집 지역에서는 10km만으로도 높은 스트레스를 유발할 수 있고, 외곽 장거리 운행에서는 20km가 정상 범위일 수 있다. 따라서 변수는 노선군 내부 분포와 기사별 최근 누적 상태를 함께 반영하는 방식으로 정규화될 필요가 있다.

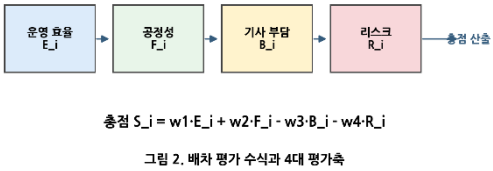


그림 2. 배차 평가 수식과 4 대 평가축

3.2 배차 평가 수식

배차 후보 i에 대한 총점 S_i는 다음과 같이 정의할 수 있다. S_i = w1E_i + w2F_i − w3B_i − w4R_i. 여기서 E_i는 운영 효율 점수, F_i는 공정성 점수, B_i는 기사 부담 점수, R_i는 리스크 점수이다. w1, w2, w3, w4는 회사의 운영 전략에 따라 조정되는 가중치이다.
예를 들어 성수기에는 운영 효율 가중치를 높일 수 있고, 노무 분쟁이 잦은 기간에는 공정성 및 부담 완화 가중치를 높일 수 있다. 또한 고객사 패널티가 큰 핵심 노선에는 시간 준수 관련 가중치를 높이고, 인력 피로가 누적된 시기에는 연속 운행 억제 가중치를 강화하는 방식도 가능하다.
본 수식은 단일 최적해를 기계적으로 확정하기보다 여러 대안의 상대적 우선순위를 계산하는 데 목적이 있다. 따라서 실무 적용 시에는 상위 n개의 후보안을 생성한 후, 담당자가 최종 승인하는 구조가 바람직하다. 이 방식은 알고리즘의 일관성과 현장 경험의 유연성을 동시에 확보할 수 있다는 장점이 있다.

3.3. 좌절지수 파라미터

본 연구는 기사 친화성을 반영하기 위해 좌절지수(Frustration Index)를 보조 지표로 제안한다. 좌절지수는 특정 기사가 최근 기간 동안 체감했을 가능성이 높은 배차 불만 요인을 정량화한 값으로, 반복 새벽운행, 장거리 공차, 비선호 노선 반복, 휴무 전후 과중 배차, 갑작스러운 배차 변경, 민원 위험 노선 반복 등을 반영한다. 좌절지수 FI_j은 j번 기사에 대해 최근 K주간의 특정 이벤트(야간 운행 비율, 강도 높은 노선 반복, 새벽 운행 집중도, 민원리스크 노선 배정 등)의 누적 정도를 정규화한 지표이며, FI_j = f(Δ근무시간, Δ새벽운행, Δ장거리노선배정, 민원리스크 노선 반복 등)로 정의한다. 좌절지수는 처벌 목적이 아니라 피로 누적 예측 및 회복 배차 우선순위 제어를 위한 보조지표이며, 특정 임계값을 초과하면 회복 배차를 우선하도록 설계한다.

좌절지수는 처벌을 위한 지표가 아니라 회복 배차와 부담 분산을 위한 조기 경보 지표로 해석되어야 한다. 같은 총 근무시간을 기록했더라도 어떤 기사는 안정적인 고정 노선을 소화했을 수 있고, 다른 기사는 시간대가 불규칙한 수시 운행을 연속적으로 처리했을 수 있다. 따라서 표면적 총량만으로는 체감 피로와 불만 가능성을 충분히 설명할 수 없다.

알고리즘은 좌절지수가 일정 기준을 초과한 기사에게 회복 배차, 선호 노선, 공차 부담이 낮은 운행을 우선 제안하도록 설계할 수 있다. 장기적으로는 좌절지수의 변화 추세를 추적함으로써 이직 위험이나 분쟁 가능성을 조기에 탐지하는 운영 보조지표로도 활용할 수 있다.

3.4. 제약조건 기반 후보 필터링

배차 추천 이전에는 필수 방어 조건을 먼저 점검해야 한다. 면허 유효성, 필수 교육 이수 여부, 음주측정 통과, 차량 정비 상태, 보험 및 검사 상태, 고객사 요구 자격 충족 여부 등은 점수 계산 이전의 필터 조건으로 처리해야 한다. 이 단계는 추천의 질뿐 아니라 법적·행정적 안전성을 확보하기 위해 필수적이다.

점수 기반 평가와 제약조건 기반 필터링을 분리하면 실무 운영이 훨씬 안정적이 된다. 즉, 자격상 배정이 불가능한 후보를 미리 제거한 뒤 그 안에서 가장 적합한 대안을 찾는 방식이 합리적이다. 이는 추천 결과에 대한 설명 가능성을 높이고, 담당자가 예외 승인한 경우에도 어떤 조건이 위배되었는지 명확히 기록할 수 있게 한다.

3.5 노선군별 맞춤 평가체계

전세버스 노선은 동일한 잣대로 평가하기 어렵다. 도심형 노선은 거리는 짧아도 정체, 정류장 접근성, 민원 가능성, 회차 복잡성이

높을 수 있고, 외곽형 노선은 거리가 길고 운전시간이 길지만 정체와 정류장 복잡도는 낮을 수 있다. 따라서 본 연구는 노선을 도심형, 외곽형, 단거리 반복형, 장거리 통근형, 행사·수시형 등으로 군집화한 뒤 각 군집의 정상 범위를 별도로 산정하는 방식을 제안한다.

노선군별 정상 범위는 누적 운행 데이터의 평균, 표준편차, 분위수 정보를 통해 산출한다. 이를 통해 특정 노선이 해당 군집 내부에서 얼마나 어려운지 판단할 수 있으며, 배차 담당자는 같은 시간의 운행이라도 실제 부담이 큰 노선을 구분할 수 있다.

예를 들어 도심형 노선에서는 평균 지연시간, 정류장 밀집도, 민원 빈도가 핵심 설명 변수로 작동할 수 있고, 외곽형 노선에서는 운전 연속시간, 회차 거리, 공차 복귀 부담이 더 중요할 수 있다. 따라서 공정성 역시 하나의 단일 지표가 아니라 노선군별 특성을 반영한 다층 구조로 설계되어야 한다.

그림 4. 노선군별 맞춤 평가체계

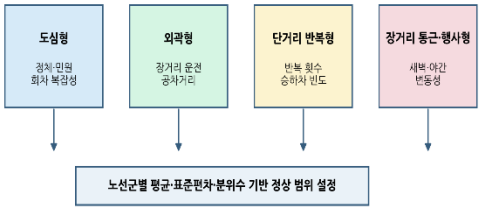


그림 4. 노선군별 맞춤 평가체계

표 1 노선군별 평가항목 예시

구분	주요 변수	평가 방향	활용 목적
도심형 노선	정체시간, 정류장 밀집도, 민원 빈도	동일 거리 대비 난이도 가중	민원·지연 리스크 완화
외곽형 노선	장거리 운전, 공차거리, 회차거리	장시간 운전 피로도 반영	피로 누적 방지
단거리 반복형	반복 횟수, 대기시간, 승하차 빈도	반복 스트레스 반영	기사 만족도 관리
장거리 통근형	새벽 출근, 야간 복귀, 대기시간	근무시간 편차 보정	근무량 공정성 확보
행사·수시형	변경 가능성, 고객 요청, 대기 불확실성	변동성 리스크 가중	돌발상황 대응

3.6 공정성 지표 산정 방식

본 연구에서의 공정성은 기사 간 누적 부담의 균형성과 노선군 간 불리함의 최소화를 목표로 하는 운영 원칙이다. 즉, 동일 수요 시나리오에서 특정 기사에게 과도한 근무시간, 운전시간, 대기시간이 집중되지 않도록 누적 부담의 분포를 균형 있게 관리하고, 도심형, 외곽형, 단거리 반복형 등 각 노선군의 특성에 따라 정상 범위를 다르게 설정하여 기사 간 편차를 최소화한다. 공정성은 노선군별 특성과 기사 상태를 반영하는 다층적 평가체계의 핵심 축으로 작용하며, 구성 요소는 다음과 같다. (1) 누적 부담의 균형성: 최근 일정 기간의 근무시간, 운전시간, 대기시간의 분포 차이를 최소화하는 정도. (2) 노선군별 적합도 반영의 적절성: 각 노선군의 정상 범위에 대한 기사 적합도의 반영 수준. (3) 선호도 반영의 편향 최소화: 기사 선호 노선 반영이 지나치게 편향되지 않도록 하는 조정 정도. (4) 기여도 다양성: 특정 기사에 배차가 집중되는 현상을 완화하고 다양한 기여를 촉진하는 정도. 이 네 가지 축의 종합은 공정성 점수로 표현되며, 모니터링 주기는 7일, 30일, 90일 단위로 설정되고

지속적인 불균형 시 회복 배차나 노선 분산 조정 등 운영 개입을 트리거한다. 공정성의 목표는 단순한 균등 배분이 아니라, 노선 특성과 기사 상태에 따른 차이를 합리적으로 조정하되 누적 부담의 불균형이 축적되지 않도록 관리하는 데 있다.

공정성은 모든 기사가 동일한 건수의 운행을 수행했는지 여부만으로 판단할 수 없다. 본 연구는 최근 7일, 30일, 90일의 누적 근무시간, 새벽·야간 운행 횟수, 장거리 노선 횟수, 민원위험 노선 배정 횟수, 공차 누적 거리, 비선호 노선 반복 횟수 등을 함께 고려하는 다기간 누적 방식의 공정성 지표를 제안한다.

이 지표는 단기 편차와 장기 편차를 동시에 점검하는 데 의미가 있다. 예를 들어 최근 일주일간 특정 기사에게 부담이 집중되었더라도 지난 한 달 동안 전체적으로 균형이 유지되었는지, 반대로 최근에는 완화되었으나 지난 3개월간 반복적으로 특정 기사가 어려운 노선을 떠안았는지 등을 구분해 볼 수 있다.

실무적으로는 공정성 점수를 인사평가에 직접 연결하기보다, 배차설명 자료와 운영 개선회의의 기초자료로 활용하는 것이 바람직하다. 이는 지표에 대한 방어적 반응을 줄이고, 기사와 회사가 함께 개선 가능한 운영언어로 정착시키는 데 도움이 된다.

3.7 시스템 구현 구조

제안 모델은 TRP와 IBMS를 결합한 구조에서 구현될 수 있다. TRP는 배차, 운행일보, 정산, 업무지시, 결재, 민원 대응 등 사무 흐름을 관리하는 업무 플랫폼으로 기능하고, IBMS는 차량 내 장비, GPS, DTG, 음주측정기, 사원증 태그, 출입문 센서, 피플카운팅 장치 등 물리적 데이터를 수집하는 통합 버스 관리 시스템으로 기능한다. 두 시스템이 연결되면 배차 계획, 실제 운행, 기사 상태, 차량 상태, 고객사 요구조건을 하나의 데이터 흐름으로 통합할 수 있다.

시스템 흐름은 운행 요청 수집, 노선 특성 자동 분류, 배차 후보군 생성, 기사별 부담·공정성 점수 계산, 리스크 조건 필터링, 최적 배차안 추천, 담당자 승인, 운행 후 실제 데이터 반영의 순서로 구성된다. 이 과정에서 알고리즘은 사람을 대체하기보다 배차 담당자의 판단을 설명 가능하고 재현 가능한 형태로 보조한다.

또한 시스템은 예외 처리 이력을 반드시 남겨야 한다. 예를 들어 고객사 긴급 요청으로 공정성 측면에서 다소 불리한 배차가 이루어졌다면, 그 사유와 승인자를 기록해야 한다. 이러한 예외 로그는 나중에 분쟁이 발생했을 때 회사가 임의적으로 배차하지 않았음을 설명하는 증거가 된다.

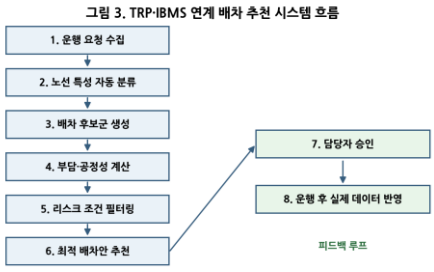


그림 3. TRP·IBMS 연계 배차 추천 시스템 흐름

3.8 운영 시나리오 예시

가령 동일한 시간대에 세 개의 통근 노선이 동시에 요청되었고, 배정 가능한 기사 네 명이 존재한다고 가정하자. 이때 알고리즘은 먼저 자격 및 방어 조건을 확인하고, 그 다음 각 기사별 최근 7일·30일 누적 부담, 노

선 선호도, 출근 동선 적합도, 민원위험 노선 경험, 좌절지수 수준을 반영하여 후보 순위를 계산한다.

이 과정에서 어떤 기사가 가장 높은 총점을 얻었다고 하더라도, 최근 동일 고객사 민원이 반복되었거나 새벽 운행이 과도하게 누적된 경우에는 우선순위가 조정될 수 있다. 반대로 운영 효율은 다소 낮더라도 기사 회복과 장기 공정성 관점에서 더 적합한 후보가 추천될 수 있다. 이러한 추천은 단기 최적화가 아닌 지속가능한 운영 최적화라는 점에서 의미가 있다.

3.9 실증 분석 설계 및 평가 프레임워크

본 연구는 개념모형 제안에 더해 향후 실증 검증을 위한 평가 프레임워크를 함께 제시한다. 실제 적용 시에는 일정 기간의 운행 데이터와 배차 이력 데이터를 수집하여, 기존 수기·경험 기반 배차와 제안 알고리즘 기반 추천 배차를 비교할 수 있다.

평가 지표는 크게 운영 성과, 기사 친화성, 공정성, 리스크 관리의 네 영역으로 구분된다. 운영 성과에는 공차거리 감소율, 배차 충돌 감소율, 시간준수율, 차량 회전율 향상 정도를 포함한다. 기사 친화성에는 근무시간 편차, 새벽·야간 배차 집중도, 비선호 노선 반복률, 회복 배차 반영률, 좌절지수 변화량을 포함할 수 있다. 공정성 영역에는 기사별 누적 부담 편차, 노선군별 어려운 노선 집중도, 선호 노선 반영률 차이 등을 포함한다. 리스크 관리 영역에는 민원 발생률, 지각 건수, 예외 승인 빈도, 자격 위반 사전 차단 건수 등이 포함된다.

실증 설계는 최소 3단계가 바람직하다. 첫째, 과거 데이터를 기반으로 노선군과 기사군의 기초 통계를 산출한다. 둘째, 동일한 운행 수요에 대해 기존 배차 방식과 제안 모델을 각각 시뮬레이션하여 비교한다. 셋째, 실제 파일럿 운영에서 현장 담당자와 기사 피드백을 수집해 정량 지표와 정성 평가를 함께 분석한다.

표 2 실증 평가 지표 프레임워크

영역	핵심 지표	측정 방식	기대 효과
운영 성과	공차거리, 시간준수율, 회전율	전월 대비 증감 및 시뮬레이션 비교	운행 효율 향상
기사 친화성	근무시간 편차, 좌절지수, 회복 배차율	기사별 누적 데이터 분석	피로 및 불만 완화
공정성	어려운 노선 집중도, 선호 반영률	7·30·90일 누적 비교	불공정성 인식 완화
리스크 관리	민원률, 지각률, 자격 위반 차단	사건 기록 및 예외 승인 로그	분쟁 및 사고 예방

3.10 기대 시뮬레이션 결과 예시

제안 모델이 적용될 경우 단기적으로는 공차거리 감소와 배차 충돌 완화가 가장 먼저 관찰될 가능성이 높다. 중기적으로는 특정 기사에게 집중되던 새벽·장거리·민원위험 노선의 반복 빈도가 완화되고, 장기적으로는 기사 만족도와 이직률, 민원 대응 비용에 긍정적인 효과가 나타날 수 있다.

특히 설명 가능한 배차 근거가 축적되면 배차 담당자와 기사 간 커뮤니케이션 비용이 감소한다. 이는 단순히 숫자상의 효율 향상보다 중요할 수 있다. 전세버스 산업에서 갈등의 상당 부분은 결과 자체보다도 결과가 형성된 과정의 불투명성에서 발생하기 때문이다.

3.11 기대효과와 정책적 시사점

첫째, 회사는 공차거리와 배차 충돌을 줄여 운영 효율을 높일 수 있다. 둘째, 기사는 자신의 배차가 어떤 기준으로 결정되었는지 설명받을 수 있어 불공정성 인식이 낮아진다. 셋째, 노무·민원·사고 리스크가 데이터로 관리되어 분쟁 발생 시 객관적 기록을 제시할 수 있다. 넷째, 장기적으로는 기사 만족도와 이직률 관리, 성과평가, 교육 우선순위 산정에도 활용할 수 있다.

정책적 측면에서 볼 때 본 모델은 단순 기업 내부 효율화 도구를 넘어 전세버스 산업의 디지털 전환 기반으로 해석할 수 있다. 운행 실적과 기사 부담을 동시에 고려하는 표준화된 데이터 구조가 정착되면, 향후 공공 지원사업, 안전관리 인증, 친환경 운행관리, 탄소배출 저감 정책과도 연계 가능성이 커진다.

또한 배차 알고리즘이 기사 친화성을 반영한다는 점은 기술 수용성 확보 측면에서도 중요하다. 현장 구성원이 기술을 감시와 통제의 도구로 인식하면 시스템 정착이 어렵다. 반대로 기술이 부담 완화와 공정성 확보의 도구로 기능한다는 인식이 확산되면 데이터 입력의 성실성과 운영 개선 참여도가 함께 높아질 수 있다.

3.12 한계 및 향후 연구

본 연구는 실무 시스템 설계를 위한 개념 모델과 평가체계를 제안하는 데 초점을 둔다. 실제 적용을 위해서는 충분한 기간의 운행 데이터, 기사별 선호도 데이터, 민원·지연·사고 데이터의 품질 확보가 필요하다. 특히 선호도와 불만 가능성은 정량화 과정에서 주관성이 개입될 수 있으므로, 설문 기반 데이터와 실제 운행 행동 데이터를 함께 참고하는 혼합 방식이 필요하다.

또한 좌절지수와 공정성 점수는 기사 개인의 민감한 평가로 오해될 수 있으므로, 처벌 목적이 아니라 부담 완화와 분쟁 예방 목적임을 제도적으로 명확히 해야 한다. 시스템 운영 규정, 지표 공개 범위, 예외 승인 절차, 설명 요청 절차 등을 함께 정비하지 않으면 오히려 불신이 증가할 가능성도 있다.

향후 연구에서는 실제 운행 데이터를 활용하여 노선군별 군집화 모델을 검증하고, 배차 추천 결과가 공차거리, 근무시간 편차, 기사 불만, 민원 발생률에 미치는 영향을 실증적으로 분석할 필요가 있다. 또한 강화학습, 제약최적화, 설명가능 AI 기법을 결합하여 배차 추천의 정확성과 수용성을 높일 수 있다.

IV. 결론

본 연구는 전세버스 배차를 단순한 차량 배정 문제가 아니라 기사 복지, 운영 효율, 공정성, 분쟁 예방을 함께 고려해야 하는 데이터사이언스 문제로 재정의하였다. 제안한 기사 친화형 AI 배차 시뮬레이터는 노선 특성, 근무부담, 공정성 지표, 좌절지수, 리스크 조건을 통합하여 배차 후보를 평가한다.

특히 도심형과 외곽형 등 노선군별로 정상 범위를 다르게 설정함으로써 현실에 맞는 공정성 기준을 제시한다. 이는 동일 배분이 아니라 합리적 부담 분산이라는 관점에서 전세버스 배차의 공정성을 새롭게 정의한다는 점에서 의미가 있다.

전세버스 산업은 인력 고령화, 기사 수급난, 고객사 요구 고도화, 노무 분쟁 증가라는 복합 문제에 직면해 있다. 이러한 환경에서 데이터 기반 배차 알고리즘은 회사가 비용을 줄이는 도구를 넘어 기사와 회사가 같은 방향을 바라보도록 만드는 운영 언어가 될 수 있다. 향후 TRP·IBMS와 연계된 실증 시스템

으로 발전시킨다면 전세버스 산업의 디지털 전환과 지속가능한 노사관계 형성에 기여할 수 있을 것이다.

참고문헌

I. 김형주. (2026). 전세버스 TRP 운영자료 및 배차 실무기록. 내부자료.

II. Pinedo, M. (2016). Scheduling: Theory, Algorithms, and Systems. Springer.

III. Bertsimas, D., & Tsitsiklis, J. (1997). Introduction to Linear Optimization. Athena Scientific.

IV. Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach. Pearson.

부록 : 현장 적용을 위한 데이터 수집 항목

기사 기본정보: 성명, 사번, 면허 종류, 면허 유효기간, 교육 이수 여부, 음주측정 기록, 근무 가능 시간대, 선호·비선호 노선.

노선 기본정보: 노선명, 고객사, 출발지, 도착지, 경유지, 예상 운행시간, 평균 지연 시간, 정류장 수, 회차 난이도.

운행 실적정보: 실제 출발·도착시각, 실제 운행거리, 실제 공차거리, 대기시간, 지연 여부, 민원 여부, 돌발상황 기록.

공정성 관리정보: 최근 7일·30일·90일 누적 근무시간, 새벽운행 횟수, 야간운행 횟수, 장거리노선 횟수, 민원위험 노선 배정 횟수.

시스템 방어정보: 배차 승인자, 변경 이력, 기사 통보 시간, 변경 사유, 예외 승인 사유, 운행 후 피드백.

위 항목은 배차 자동화를 위한 최소 데이터셋이 아니라, 배차 설명 가능성과 분쟁 예방을 함께 달성하기 위한 운영 데이터셋이다. 데이터는 처벌이 아니라 부담의 조기 발견, 위험의 사전 차단, 합리적 배차 기준의 형성을 위해 사용되어야 한다.

직무 역량 벡터와 강의평 감성 키워드 분석을 결합한 설명 가능한 교내 커리큘럼 추천 시스템

진수빈*, 왕희원, 박현진, 김나영, 이용걸
한신대학교 AI·SW 대학
e-mail:{vheh5053*, gmldnjs0321, guswls, sinamon500, pattern}@hs.ac.kr

An Explainable Campus Curriculum Recommendation System Combining Job-Competency Vectors and Sentiment Keyword Analysis of Course Reviews

Su been Jin*, Heewon Wang, Hyeon Jin Park, Nayoung Kim, Yong Geol Lee
School of Computing and Artificial Intelligence, Hanshin University

요 약

대학 교육 환경에서 학생들은 자신의 진로 목표에 적합한 교과목을 선택해야 하지만, 강의계획서, 이수체계도, 졸업요건, 수강 후기 등 관련 정보가 분산되어 있어 체계적인 수강 계획을 수립하는 데 어려움을 겪는다. 특히 소프트웨어·인공지능·데이터사이언스 분야에서는 희망 직무별 요구 역량이 빠르게 세분화되고 있으므로, 단순한 과목 검색이 아니라 학생의 진로 목표와 학습 이력을 반영한 개인화된 커리큘럼 추천이 필요하다. 본 연구에서는 학생의 희망 직무, 이수 과목, 학습 선호도, 교내 과목 정보, 강의평 데이터를 통합하여 개인 맞춤형 과목 추천과 학습 로드맵을 제공하는 CourseFit-AI 추천 구조를 제안한다. 제안 방법은 직무별 핵심 역량과 과목별 역량을 동일한 벡터 공간으로 표현하고, 코사인 유사도, 직무 연관성, 실습 비중을 결합하여 AI 매칭도를 산출한다. 또한 강의평 텍스트에서 전체 감성과 키워드 단위 감성을 분리하여 실습 풍부, 과제 많음, 환경설정 어려움 등 정형화된 수강 경험 키워드를 추출하고, 이를 추천 점수와 추천 사유 생성에 보조 지표로 반영한다. 본 시스템은 추천 결과와 함께 추천 이유를 제공함으로써 교육 추천 시스템에서 요구되는 설명 가능성을 강화하며, 추천 과목을 학기별 로드맵 및 교수자용 과목 수요 분석 기능으로 확장한다. 이를 통해 CourseFit-AI는 학생의 수강 의사결정뿐 아니라 학과 차원의 교육과정 운영에도 활용 가능한 데이터 기반 커리큘럼 추천 시스템으로 기능할 수 있다.

키워드 : 교과목 추천, 직무 기반 추천, 설명 가능한 AI, 강의평 분석, 학습 경로 설계

※ 본 연구는 2026년도 과학기술정보통신부 및 정보통신기획평가원의 'SW중심대학사업' 지원을 받아 수행되었음. (2024-0-00076)

I. 서론

최근 대학 교육에서는 학생의 진로 목표와 학습 이력을 반영한 맞춤형 수강 지원의 필요성이 높아지고 있다. 특히 인공지능, 소프트웨어, 데이터사이언스 분야에서는 동일 학과 학생이라도 희망 직무에 따라 필요한 역량이 달라진다. 예를 들어 AI 엔지니어에게는 머신러닝, 딥러닝 역량이 중요하고, 데이터 엔지니어에게는 데이터베이스, 빅데이터 처리 역량이 상대적으로 중요하다. 따라서 단순히 전공 과목이나 인기 과목을 안내하는 방식만으로는 학생의 진로에 맞는 수강 계획을 지원하기 어렵다.

기존 수강신청 과정에서는 학생이 강의계획서, 이수체계도, 졸업요건, 시간표, 강의평 등을 개별적으로 확인하고 스스로 종합해야 한다. 또한 과목명, 학점, 이수 구분과 같은 공식 정보만으로는 실제 수강생이 체감한 과제량, 실습 비중, 난이도, 프로젝트 경험 등을 파악하기 어렵다. 강의평은 이러한 정보를 포함하지만 자유 서술형 텍스트이므로 추천 시스템에서 그대로 활용하기 어렵다.

본 연구는 이러한 문제를 해결하기 위해 CourseFit-AI를 제안한다. CourseFit-AI는 학생의 희망 직무, 이수 과목, 학년, 선호 수업 방식, 과목 정보, 강의평 분석 데이터를 결합하여 개인 맞춤형 교과목 추천과 학습 로드맵을 제공하는 시스템이다. 본 연구의 핵심은 첫째, 직무와 과목의 역량 정보를 바탕으로 추천 점수를 산출하는 것, 둘째, 강의평에서 긍정·부정 키워드를 추출하여 추천 사유에 반영하는 것, 셋째, 추천 결과를 학생용 로드맵과 교수자용 수요 분석으로 확장하는 것이다.

II. 관련 연구

추천 시스템은 사용자 특성과 항목 정보를 분석하여 사용자에게 적합한 항목을 제안하는 기술이다. 일반적으로 콘텐츠 기반 추천, 협업 필터링, 하이브리드 추천 방식으로 구분된다[1][2]. 교육 분야의 추천 시스템은 일반 상품 추천과 달리 학습자의 목표, 현재 수준, 선수 지식, 학습 순서를 함께 고려해야 한다. 따라서 개별 과목 추천뿐 아니라 목표 역량에 도달하기 위한 학습 경로 추천이 중요하다[3].

또한 교육 추천 시스템에서는 추천 결과의 설명 가능성이 중요하다. 학생은 추천된 과목이 자신의 희망 직무와 어떤 관련이 있는지, 어떤 역량을 강화하는지 이해할 수 있어야 한다[4][5]. 한편 강의평과 같은 리뷰 데이터는 실제 수강 경험을 반영하지만, 하나의 문장 안에 긍정과 부정이 동시에 포함될 수 있다. 따라서 문장 전체 감성만 분류하는 방식보다, 특정 키워드 주변의 긍정·부정 맥락을 함께 분석하는 방식이 필요하다[6].

III. 제안방법

3.1 시스템 개요

CourseFit-AI는 학생 데이터, 직무 데이터, 과목 데이터, 강의평 분석 데이터를 입력으로 사용한다. 학생 데이터에는 희망 직무, 이수 과목, 학년, 선호 수업 방식이 포함된다. 직무 데이터는 직무별 필요 역량 키워드로 구성되며, 과목 데이터는 과목명, 학수번호, 과목 키워드, 난이도, 선수 과목, 개설 학기, 수업 유형을 포함한다. 강의평 분석 데이터는 과목별 긍정 키워드와 부정 키워드로 구성된다.

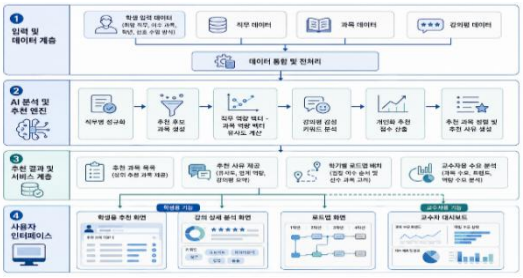


그림1. CourseFit-AI 추천 시스템 아키텍처

※ 본 연구는 2026년도 과학기술정보통신부 및 정보통신기획평가원의 'SW중심대학사업' 지원을 받아 수행되었음. (2024-0-00076)

그림 1은 CourseFit-AI의 전체 시스템 아키텍처를 나타낸다. 입력 및 데이터 계층에서는 학생 입력 데이터, 직무 데이터, 과목 데이터, 강의평 데이터를 통합하고 전처리한다. 이후 AI 분석 및 추천 엔진에서는 희망 직무를 표준 직무명으로 정규화하고, 이미 이수한 과목을 제외하여 추천 후보 과목을 생성한다. 다음으로 직무와 과목의 역량 유사도, 학생의 학습 단계, 선수 과목 충족 여부, 개설 가능성, 수업 방식 선호, 강의평 키워드를 반영하여 추천 점수를 산출한다. 최종 결과는 추천 과목 목록, 추천 사유, 학기별 로드맵, 교수자용 수요 분석 기능으로 제공된다.

3.2 직무 기반 추천 점수 산출

추천 과정에서는 먼저 전체 과목 중 학생이 이미 이수한 과목을 제외하여 후보 과목을 생성한다. 이후 학생의 희망 직무와 과목의 역량 키워드를 비교하여 직무 적합도를 계산한다. 예를 들어 데이터 엔지니어 직무의 필요 역량이 SQL, 데이터베이스, 데이터 파이프라인, 빅데이터, Spark이고, 특정 과목이 빅데이터와 Spark를 포함한다면 해당 과목은 데이터 엔지니어 직무와 일정 수준의 적합도를 가진다.

본 연구에서는 최종 추천 점수를 다음과 같이 정의한다.

$$\begin{aligned} Score(c,j,s) = {} & 0.45Sim(c,j) + 0.20Level(c,s) \\ & + 0.15Prereq(c,s) + 0.10Open(c) \\ & + 0.05Pref(c,s) + 0.05Review(c,s) \end{aligned}$$

여기서 $Sim(c,j)$는 과목 c와 희망 직무 j간의 직무 적합도, $Level(c,s)$는 학생 s의 현재 학습 단계와 과목 난이도의 적절성, $Prereq(c,s)$는 선수 과목 충족 여부, $Open(c)$는 과목의 개설 가능성, $Pref(c,s)$는 학생의 선호 수업 방식과 과목 유형의 일치도, $Review(c,s)$는 강의평 키워드와 학생 선호·회피 요소의 일치도를 의미한다.

이 점수식은 직무 기반 추천이라는 목적을 반영하여 직무 적합도에 가장 높은 비중을 두었다. 학습 단계와 선수 과목 충족 여부는 학생이 실제로 해당 과목을 수강할 수 있는지를 판단하는 기준으로 사용된다. 개설 여부, 수업 방식 선호, 강의평 키워드는 추천 결과를 보정하는 보조 요소로 활용된다. 최종적으로 후보 과목은 추천 점수 기준으로 정렬되며, 각 과목에는 추천 이유가 함께 제공된다.

3.3 강의평 감성 키워드 분석

CourseFit-AI는 공식 과목 정보만으로 파악하기 어려운 실제 수강 경험을 반영하기 위해 강의평 데이터를 분석한다. 강의평은 자유 서술형 텍스트이므로 먼저 불필요한 공백, 특수문자, 반복 표현을 정리한다. 이후 실습, 과제, 팀플, 환경설정, 난이도, 프로젝트, 피드백 등 수업 특성을 나타내는 도메인 키워드를 탐지한다.

본 연구에서는 리뷰 전체 감성과 키워드 단위 감성을 분리하였다. 예를 들어 “실습은 좋았는데 과제가 많아서 힘들었습니다.”라는 문장은 전체적으로 부정에 가깝지만, 세부적으로는 실습에 대한 긍정 평가와 과제에 대한 부정 평가가 함께 존재한다. 따라서 본 시스템은 리뷰 전체를 하나의 감성으로만 처리하지 않고, 키워드 주변 문맥을 분석하여 긍정 키워드와 부정 키워드를 각각 추출한다.

표 1. 강의평 키워드 분석 예시

입력 강의평	추출 키워드	추천 반영
실습이 많아서 도움이 되었다.	실습풍부, 도움됨	실습 선호 학생에게 긍정 반영
과제가 많아서 힘들었다.	과제많음, 힘듦	과제 부담 회피 학생에게 주의 정보 제공

※ 본 연구는 2026년도 과학기술정보통신부 및 정보통신기획평가원의 ‘SW중심대학사업’ 지원을 받아 수행되었음. (2024-0-00076)

환경설정이 어렵지만 프로젝트는 유익했다.	환경설정 어려움, 프로젝트경험	주의 요소와 장점을 함께 제시
설명이 친절하고 이해하기 쉬웠다.	설명친절, 이해쉬움	기초 단계 학생에게 긍정 반영

추출된 키워드는 추천 점수와 추천 사유 생성에 활용된다. 예를 들어 학생이 실습 중심 수업을 선호하면 '실습 풍부', '프로젝트 경험' 키워드가 있는 과목에 긍정적으로 반영할 수 있다. 반대로 과제 부담을 피하고 싶은 학생에게는 '과제 많음', '학습량 많음' 키워드를 주의 정보로 제공할 수 있다.

3.4 추천 결과 제공

추천 결과는 학생이 쉽게 이해할 수 있도록 과목 카드 형태로 제공된다. 각 과목에는 과목명, 학점, 이수 구분, 추천 점수, 핵심 키워드, 추천 이유가 함께 제시된다. 추천 이유는 "해당 과목이 희망 직무의 어떤 역량과 연결되는지"와 "강의평에서 어떤 수강 경험 키워드가 확인되었는지"를 중심으로 생성된다.

또한 추천 과목은 학기별 로드맵과 연계된다. 학생은 추천 과목을 로드맵에 저장하고, 이미 이수한 과목과 앞으로 수강할 과목을 함께 확인할 수 있다. 로드맵 구성 시에는 과목 난이도, 권장 학년, 선수 과목, 개설 학기를 고려하여 기초 과목이 심화 과목보다 먼저 배치되도록 한다. 교수자용 기능에서는 학생들의 희망 직무 분포와 과목 수요를 확인할 수 있어, 학생 상담이나 다음 학기 과목 개설 검토에 활용할 수 있다.

Ⅳ. 결론

본 연구에서는 학생의 희망 직무, 이수 과목, 학습 선호도, 과목 정보, 강의평 데이터를 통합하여 개인 맞춤형 교과목 추천과 학습 로드맵을 제공하는 CourseFit-AI를 제안하였다. 제안 시스템은 직무와 과목의 역량 정보를 비교하여 추천 점수를 산출하고, 학생의 학습 단계, 선수 과목, 개설 여부, 수업 방식 선호, 강의평 키워드를 함께 반영한다.

특히 본 연구는 강의평 분석에서 리뷰 전체 감성과 키워드 단위 감성을 분리함으로써 혼합 감성 리뷰를 처리할 수 있도록 하였다. 이를 통해 '실습 풍부', '과제 많음', '환경설정 어려움'과 같은 정형화된 수강 경험 키워드를 추출하고, 추천 사유에 반영할 수 있었다. CourseFit-AI는 학생에게 추천 결과의 근거를 제공하여 수강 의사결정을 지원하고, 교수자에게는 학생들의 직무 수요와 과목 수요를 파악할 수 있는 정보를 제공한다.

다만 본 연구는 프로토타입 설계와 규칙 기반 분석을 중심으로 수행되었기 때문에 대규모 실제 수강 이력과 강의평 데이터를 활용한 정량 검증에는 한계가 있다. 향후 연구에서는 실제 사용자 데이터를 확대 수집하고, 추천 만족도와 강의평 분석 정확도를 평가할 필요가 있다. 또한 학사 시스템과 연동하여 졸업요건, 시간표, 수강 정원 정보를 함께 반영한다면 실제 대학 환경에서 활용 가능한 커리큘럼 추천 시스템으로 발전할 수 있을 것이다.

참고문헌

[1] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data*

※ 본 연구는 2026년도 과학기술정보통신부 및 정보통신기획평가원의 'SW중심대학사업' 지원을 받아 수행되었음. (2024-0-00076)

Engineering, Vol. 17, No. 6, pp. 734-749, 2005.

[2] F. Ricci, L. Rokach, and B. Shapira, *Recommender Systems Handbook*, 3rd ed., Springer, 2022.

[3] X. Chen, J. Shen, W. Xia, J. Jin, Y. Song, W. Zhang, W. Liu, M. Zhu, R. Tang, K. Dong, D. Xia, and Y. Yu, "Set-to-sequence ranking-based concept-aware learning path recommendation," *arXiv preprint arXiv:2306.04234*, 2023.

[4] J. Frej, N. Shah, M. Knežević, T. Nazaretsky, and T. Käser, "Finding paths for explainable MOOC recommendation: A learner perspective," *arXiv preprint arXiv:2312.10082*, 2023.

[5] N. Tintarev and J. Masthoff, "A survey of explanations in recommender systems," *IEEE 23rd International Conference on Data Engineering Workshop*, pp. 801-810, 2007.

[6] M. Hu and B. Liu, "Mining and summarizing customer reviews," *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168-177, 2004.

※ 본 연구는 2026년도 과학기술정보통신부 및 정보통신기획평가원의 'SW중심대학사업' 지원을 받아 수행되었음. (2024-0-00076)

태그 참조 LLM 기반 AI 문서 보안 시스템

최은우 O, 박종원*, 김동호**, 민병석***
O,*세종대학교 데이터사이언스학과, **(주) 마크애니
***세종대학교 인공지능 데이터사이언스학과
e-mail: eunwooo0717@sju.ac.kr O, jason7815@sju.ac.kr*, kimdh@markany.com**,
bmin@sejong.ac.kr***

Tag-Reference LLM based AI Document Security System

Eunwoo Choi O, Jongwon Park*, Dong-ho Kim**, Byungseok Min***
O,*Department of Data Science, Sejong University, **MarkAny Co. Ltd
***Department of Artificial Intelligence and Data Science, Sejong University

요 약

기업, 공공기관, 연구기관의 문서 기반 업무가 증가하면서 문서 내 민감정보 유출 방지와 보안 등급 분류의 중요성이 커지고 있다. 업무 문서에는 개인정보, 식별정보, 개인별 계약·증권·고객 식별번호, 계약정보, 기술정보가 혼재하며, 표, 이미지, 캡처 화면과 같은 비텍스트 요소도 함께 포함된다. 따라서 단순 키워드 검색이나 고정 규칙만으로는 문맥에 따라 달라지는 민감성을 안정적으로 판단하기 어렵고, 외부 API 기반 대규모 언어모델에 원문 판단을 의존하는 방식도 보안 문서 처리 환경에서는 제약이 크다. 본 논문은 이러한 한계를 해결하기 위해 전처리 태그 생성, VLM 기반 이미지 내 문자 추출 및 설명 생성, LLM 기반 문맥 판단을 결합한 태그 참조 기반 AI 문서 보안 파이프라인을 제안한다. 제안 방법은 DOCX 문서에서 텍스트, 표, 이미지를 블록 단위로 정규화한 뒤 NER, 정규식, 보안 규칙, 이미지 내 문자 정보 및 VLM 설명을 결합하여 민감정보 가능성이 있는 문서 내 민감 구간 전처리 태그를 생성한다. 이후 LLM은 대상 문서 블록 전체 문맥을 읽고 전처리 태그, 주변 문맥, 이미지 내 문자 정보와 의미 설명을 참고하여 실제 보호 대상 여부를 판단하고, 민감 구간별 판단 근거를 생성한다. 이를 통해 민감정보 구간의 누락을 줄이고, 문맥 기반 필터링으로 불필요한 판단을 완화하며, 보안 담당자가 검토할 수 있는 설명을 포함한 판단 결과를 제공하는 것을 목표로 한다.

키워드: AI 문서 보안, 민감정보, 전처리 태그, 문맥 판단, LLM

I. 서론

최근 조직의 업무는 보고서, 계약서, 연구 문서, 내부 검토 자료와 같은 문서 단위 산출물을 중심으로 이루어진다. 이러한 문서에는 이름, 연락처, 주소, 식별번호와 같은 개인정보뿐 아니라 개인별 계약·증권·고객 식별번호, 고객명단, 계약 조건, 기술 사양, 시스템 화면과 같은 업무상 민감정보가 함께 포함될 수 있다. 특히 실제 문서는 순수 텍스트만으로 구성되지 않고 표, 삽입 이미지, 스캔 영역, 캡처 화면 등 다양한 시각 요소를 포함하므로, 문서 보안 판정은 텍스트와 비텍스트 정보를 함께 고려해야 한다. 기존 문서 보안 및 민감정보 탐지 방식은 정규식, 사전, 고정 규칙에 기반한 접근을 널리 사용해 왔다. 이 방식은 주민등록번호, 전화번호, 이메일처럼 형식이 명확한 정보에는 효과적이지만, 예시 데이터와 실제 민감정보의 구분, 공개 가능한 대표 연락처와 내부 담당자 정보의 구분, 문맥에 따라 민감도가 달라지는 기술·계약 정보 판정에는 한계가 있다. 통계 기반 머신러닝

또는 전통적 분류 모델은 문서 특징을 이용해 보안 등급을 예측할 수 있으나, 판단 근거를 충분히 설명하기 어렵고 문서 내 어느 항목을 마스킹해야 하는지 직접 제공하지 못한다.

LLM 을 이용한 문서 판단은 문맥 이해 측면에서 장점이 있으나, 보안 문서 처리 환경에서는 외부 API 기반 SOTA 모델 사용이 제한될 수 있다. 민감한 문서 원문을 외부 서버로 전송해야 하는 문제가 있으며, 제한된 실행 환경에서 사용하는 LLM 은 모델 규모와 자원 제약으로 인해 단독 사용 시 성능이 제한될 수 있다. 또한 LLM 에 문서 전체 판단을 의존하면 어떤 단어, 구문, 이미지 근거가 민감정보 판정에 기여했는지 설명하기 어렵고, 그 결과 과도한 마스킹이나 누락이 발생할 수 있다.

문서 내 이미지 처리에서도 단순 이미지 분류만으로는 충분하지 않다. 신분증 사진, 계정 화면 캡처, 계약서 일부 캡처, 명함 이미지, 내부 시스템 UI 화면은 단순히 민감 이미지로 분류하는 것보다 이미지 내부의 문자 정보와 시각적 의미를 함께 해석해야 한다. 따라서 현재 구현은 이미지에 포함된 이름, 조직명, 계정명, UI 텍스트와 같은 단서를 VLM 기반 이미지 내 문자 추출과 요약으로 확보하고, 이미지의 장면 및 문맥을 사전학습 VLM 기반 설명으로 구조화한다.

이에 본 연구는 텍스트 · 표 · 이미지에서 얻은 단서를 문서 블록 단위 후보 태그로 구조화하고, LLM 이 대상 블록과 주변 문맥을 함께 읽어 보호 대상 여부와 판단 근거를 생성하는 태그 참조 기반 문서 보안 파이프라인을 제안한다.

본 논문의 기여는 다음과 같다. 첫째, NER, 정규식, 보안 규칙, 표 구조, VLM 기반 이미지 내 문자 정보와 설명을 문서 내 민감 구간 전처리 태그로 통합하는 구조를 제안한다. 둘째, 대상 문서 블록과 주변 블록 문맥, 전처리 태그를 함께 활용하는 LLM 기반 문서 보안 판단 구조를 제시한다. 셋째, 이미지 내 문자 정보와 의미 설명을 텍스트 단서로 변환하여 LLM 기반 보안 판단에 통합하는 구조를 구현한다. 넷째, 영업 문서, KDPII, 도서 대조셋을 통해 태그 참조 구조의 탐지 성능과 오탐 억제 경향을 검증한다.

II. 관련 연구

2.1 Rule/Regex 및 NER 기반 탐지

기존 DLP 및 민감정보 탐지 시스템은 정규식, 사전, 내장 민감정보 유형, 정책 규칙을 기반으로 형식이 명확한 개인정보 패턴을 탐지하는 데 효과적이다. 그러나 규칙 기반 접근은 새로운 문서 유형이 추가될수록 규칙 유지보수 비용이 증가하며, 문맥에 따라 민감 여부가 달라지는 표현이나 비정형 업무 정보에 대한 일반화가 제한적이다.

NER 기반 접근은 이름, 기관명, 위치명 등 개체 단위 표현을 탐지할 수 있어 정규식으로 포착하기 어려운 텍스트 구간을 보완한다. 그러나 NER 결과 자체가 곧 민감정보 판정을 의미하지는 않는다. 예를 들어 기관명이나 사람명은 문서 목적과 공개 범위에 따라 보호 대상이 될 수도 있고 일반 정보가 될 수도 있다. 따라서 NER 은 최종 판정기라기보다 민감정보 가능성이 있는 표현을 전처리 태그로 기록하는 단계에 적합하다.

2.2 이미지 기반 문서 정보 추출

문서 이미지는 사진, 캡처 화면, 스캔 영역, 삽입 도표 등 다양한 형태로 존재한다. 이미지 단순 분류 방식은 이미지가 민감할 가능성을 빠르게 판단할 수 있지만, 이미지 내부에 포함된 이름, 계정명, 계약 정보와 같은 실제 단서를 설명하기 어렵다. 최근 VLM 기반 접근은 이미지 내부의 문자 정보와 장면 설명을 텍스트 단서로 변환할 수 있어, 이미지 처리를 단순 분류가 아니라 문서 보안 판단을 위한 정보 추출 문제로 다룰 수 있게 한다.

2.3 LLM 기반 판단과 설명 가능성

LLM 은 문서 내 문맥을 해석하고 모호한 표현의 의미를 판단하는 데 유용하다. 그러나 보안 문서 처리에서는 외부 API 사용이 제한될 수 있고, LLM 판단은 어떤 정보가 판단의 근거인지 명확하지 않은 블랙박스 문제를 가진다. 설명 가능한 문서 보안 시스템은 단순 등급 예측을 넘어 민감 구간, 근거, 위험 수준, 마스킹 필요성을 함께 제시해야 한다. 본 연구는 LLM 을 최종 문맥 판정기로 사용하되, 대상 문서 블록 문맥과

전처리 태그를 함께 입력하여 민감 구간 여부를 판단하고 그 판단 근거를 제시하도록 구성한다.

III. 제안 방법

3.1 전체 파이프라인

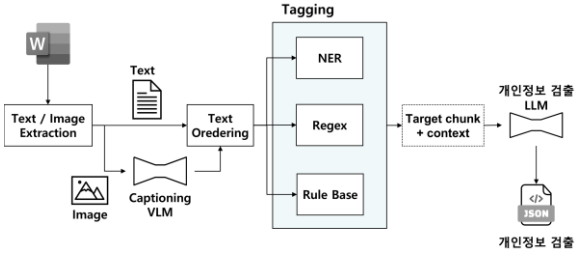


그림 1. 제안 태그 참조 기반 AI 문서 보안 판단 파이프라인

시스템은 DOCX 의 OOXML 구조를 분석하여 paragraph, table, image 블록과 이미지 asset 을 생성한다. 본 논문에서 청크는 단어 또는 문장 단위가 아니라 LLM 이 한 번에 판단하는 문서 블록 기반 입력 단위를 의미한다. 기본적으로 하나의 paragraph 또는 table 블록을 하나의 청크로 사용하되, 반복 행으로 구성된 표에서는 같은 행의 필드명과 값을 함께 보존하기 위해 행 단위 분할을 적용할 수 있다. 이후 각 청크에는 전처리 태그와 앞뒤 문맥을 연결하여 LLM 입력을 구성한다.

표 1. 제안 시스템의 단계별 중간 표현과 산출물

구분	역할
문서 블록	DOCX 문단, 표, 이미지를 공통 단위로 정규화한 기본 입력
판단 문서 블록(청크)	LLM 이 한 번에 판단하는 입력 단위로, paragraph 또는 table 블록을 기본으로 사용하고 반복 표는 행 단위로 분할
전처리 태그	NER, regex, 규칙, 표 구조, 이미지 분석에서 생성된 후보 구간 태그
문맥 판단 결과	후보 구간별 LLM 판정, 정보 유형, 위험도, 판단 이유

3.2 텍스트 기반 전처리 태그 생성

텍스트 기반 전처리 태그 생성 단계는 민감정보 후보 구간을 넓게 포착하기 위해 사용된다. NER 기반 태깅은 이름, 기관명, 위치명 등 개체 표현을 탐지하고, 정규식 기반 태깅은 전화번호, 이메일, 계좌번호, 주소, 식별번호, 내부 IP, API 주소, 인증키 등 형식적 패턴이 강한 항목을 탐지한다. 규칙 기반 태깅은 문서 보안 정책과 도메인 지식을 반영하여 업무상 민감할 수 있는 표현을 후보 구간으로 기록한다. 표 구조에서는 행과 셀 단위 텍스트를 보존하고, 헤더 행이 탐지되는 경우 header-cell 관계를 메타데이터로 기록한다. 본 실험에서는 표 블록을 줄 단위 텍스트로 정규화하거나 행 단위로 분할하여, LLM 이 같은 행 안의 필드명과 값을 함께 볼 수 있도록 하였다.

3.3 이미지 정보 추출 및 태그 통합

이미지 처리 단계는 단순 분류가 아니라 정보 추출 단계로 구성된다. 현재 구현은 별도 OCR 엔진이 아니라 DOCX 에 포함된 이미지 asset 을 VLM 에 입력하여 이미지 내 문자 정보와 요약, 상세 설명을 텍스트 단서로 변환한다. 이렇게 변환된 문자 정보와 설명은 대상 문서 블록, 주변 문맥, 전처리 태그와 함께 LLM 입력에 연결된다.

태그 통합 단계는 NER, regex, 규칙, 표 구조, 이미지 분석 결과에서 생성된 민감 구간 태그를 하나의 전처리 태그 집합으로 결합한다. 동일하거나 유사한 구간이 여러 생성원에서 탐지되면 중복을 정규화하고, 원문 블록 또는 이미지와의 연결 정보를 유지한다. 이 단계에서는 텍스트 · 표 · 이미지에서 얻은 정보를 동일한 LLM 입력 형식으로 정렬하여 이후 문맥 판단 단계에서 사용할 수 있도록 구조화한다.

3.4 LLM 문맥 판단과 산출물

LLM 은 후보 텍스트만 입력받는 것이 아니라 대상 문서 블록 전체 텍스트를 읽고, NER, regex, 규칙 기반 전처리 태그를 참고 신호로 함께 활용한다. 입력에는 대상 문서 블록 텍스트, 이전/다음 문서 블록 문맥, 대상 블록의 전처리 태그, 표 행 단위로 정규화된 블록 텍스트가 포함된다. 이미지가 포함된 경우에는 VLM 이 생성한 문자 정보와 요약 설명을 텍스트 단서로 함께 제공한다. 주변 문서 블록은 문맥 이해를 위한 참고 정보로만 사용되며, 최종 검출은 반드시 대상 문서 블록 내부의 텍스트 구간으로 제한된다. LLM 은 각 구간에 대해 실제 보호 대상 여부, 보안 라벨, 위험도, 마스킹 방식, 판단 이유를 생성한다.

이 구조에서 전처리 태깅은 민감 단서의 누락을 줄이는 역할을 하고, LLM 문맥 판단은 예시 데이터, 공개 정보, 일반

안내 문구, 단순 로고와 같이 보호 대상으로 보기 어려운 표현을 걸러내는 역할을 한다. 본 논문은 민감하다고 판단된 문서 내 구간과 판단 근거를 주요 산출물로 다룬다.

IV. 실험 및 평가

4.1 실험 데이터셋

실험 데이터셋은 원본 DOCX, 문서 블록, 전처리 태그, 정답 라벨을 연결할 수 있도록 구성한다. 라벨러는 파서가 생성한 문서 블록을 기준으로 민감정보 구간의 시작/끝 위치, 정보 유형, 민감 여부, 마스킹 필요 여부를 기록한다. 평가 단위는 문장이나 단어가 아니라 paragraph 또는 table 기반 문서 블록을 기본으로 하며, 영업 문서처럼 동일한 표 구조 안에 여러 개인정보 행이 반복되는 경우에는 행 단위 청크를 추가로 사용한다.

영업 문서 샘플은 고객·계약·담당자 정보가 포함된 업무 문서 양식을 모사한 10 개 문서로 구성하였다. 성능 평가는 실제 보호 대상 개인정보 152 개 구간만 정답으로 사용하고, 예시·더미 값과 공개·기관·회사성 정보는 오탐 진단 항목으로 분리하였다. 반복 표는 같은 행의 필드명과 값을 함께 판단할 수 있도록 행 단위로 분할하였으며, Tag-reference 조건에는 사전 구축된 고품질 후보 태그를 사용하였다. KDPII 데이터셋[1]에서는 source order 기준 4,700 문장을 추출하고, 이름, 별명/닉네임, ID/계정명, 전화번호, 이메일, 주소, 주민등록번호, 외국인등록번호, 여권번호, 운전면허번호, 계좌번호, 카드번호 등 13 종 1 차 개인정보를 정답 라벨로 사용하였다. KDPII 와 공개 도서 대조셋 실험에서는 NER, 정규식, 규칙 기반 전처리를 통해 생성한 태그를 사용하였다. 공개 문서에서의 오탐 억제 경향은 AI Hub 도서자료 기계독해 데이터셋[2]의 validation context 를 이용하여 확인하였다. 평가는 precision, recall, F1-score, false positive, false negative 를 중심으로 수행하였고, 이때 모델이 예측한 개인정보 유형명이 정답 라벨명과 다르더라도, 예측 구간이 정답 구간과 겹치면 정답 후보로 보았다. 단, 하나의 정답 구간에는 하나의 예측만 대응시켜 중복 예측이 성능을 부풀리지 않도록 하였다. 동일 문서 블록 또는 행 단위 청크 안에서 start/end/text 가 같은 반복 예측은 모든 조건에 동일하게 하나의 예측으로 정규화하였다. 본 연구에서 F1-score 는 precision 과 recall 의 조화평균으로 정의한다.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

실험은 영업 문서 통제 실험, KDPII 1 차 개인정보 평가, 공개 도서 본문 대조셋의 세 흐름으로 구성하였다. 모든 LLM 판단 조건은 대상 블록 앞뒤 1 개 블록을 참고하는 W=1, direction=both 를 기본 문맥 조건으로 사용하였다.

V. 결과 및 논의

표 2. 실험 1: 영업 문서 개인정보 탐지 성능 비교

모델	방법	Prec.	Rec.	F1
Qwen 2B	LLM-only	0.7500	0.2368	0.3600
	Tag-reference	0.6420	0.3421	0.4464
Qwen 9B	LLM-only	0.6963	0.9803	0.8142
	Tag-reference	0.7308	1.0000	0.8444
Qwen 27B	LLM-only	0.8352	1.0000	0.9102
	Tag-reference	0.8492	1.0000	0.9184

표 2 는 영업 문서 10 개에 대해 LLM-only 와 Tag-reference + LLM 을 겹침 구간 기준으로 비교한 결과이다. 이 실험은 실제 운영 태거의 자동 탐지 성능이 아니라, 고품질 후보 태그가 주어졌을 때 LLM 이 문맥적으로 보호 대상 여부를 선별할 수 있는지를 확인하는 데 초점을 둔다. Qwen3.5 계열 모델[3]을 사용한 결과, Qwen 2B 와 9B 에서는 Tag-reference 조건이 F1-score 를 각각 0.3600 에서 0.4464, 0.8142 에서 0.8444 로 개선하였다. Qwen 27B 에서도 동일한 recall 1.0000 을 유지하면서 precision 을 0.8352 에서 0.8492 로, F1-score 를 0.9102 에서 0.9184 로 개선하였다.

표 3. KDPII 1 차 개인정보 성능 비교

모델	방법	Prec.	Rec.	F1
Qwen 2B	LLM-only	0.6556	0.4987	0.5665
	Tag-reference	0.3473	0.7645	0.4776
Qwen 9B	LLM-only	0.4599	0.6940	0.5532
	Tag-reference	0.7637	0.8060	0.7843
Qwen 27B-FP8	LLM-only	0.7815	0.8287	0.8044
	Tag-reference	0.8869	0.8589	0.8727
Qwen 27B-FP16	LLM-only	0.7304	0.8325	0.7781
	Tag-reference	0.9092	0.8577	0.8827
XLM-R large	Fine-tuned token classifier	0.8097	0.9484	0.8735

표 3 은 KDPII 1 차 개인정보 문장에 대해 동일한 태그 참조 구조를 적용한 결과이다. KDPII 원논문은 14 개 한국어 LM 중 XLM-R large 계열의 Xroberta_L 이 가장 높은 F1-score 0.89 를 보였다고 보고하였다[1]. 본 연구에서는 학습형 LM 대비 성능 수준을 확인하기 위해 XLM-R large[4]를 직접 fine-tuning 하여 기준선으로 제시하였다. 재학습한 XLM-R large 는 F1-score 0.8735 로 원논문 보고치보다는 낮았지만, 프롬프트 기반 LLM 조건과 비교할 수 있는 학습형 기준선으로 기능한다. Qwen 9B 에서는 Tag-reference 조건이 LLM-only 대비 F1-score 를 0.5532 에서 0.7843 으로 크게 개선하였고, FP 는 647 개에서 198 개로 감소하였다. Qwen 27B-FP8 에서도 F1-score 가 0.8044 에서 0.8727 로 증가하여 재학습한 XLM-R 기준선과 거의 같은 수준에 도달하였으며, Qwen 27B-FP16 에서는 0.7781 에서 0.8827 로 증가하여 이를 소폭 상회하였다. 특히 27B 계열에서는 precision 과 recall 이 함께 개선되거나 precision 이 크게 향상되어, 전처리 태그가 단순히 더 많은 구간을 예측하게 만드는 것이 아니라 LLM 이 문맥 속에서 태그를 선별하는 데 도움을 주었음을 보여준다. 반면 Qwen 2B 에서는 recall 이 0.4987 에서 0.7645 로 증가했지만 FP 가 크게 늘어 F1-score 가 하락하였다. 이는 작은 모델이 태그를 참고 신호가 아니라 정답에 가까운 신호로 과신할 수 있음을 보여주며, 태그 참조 구조의 효과가 모델의 문맥 처리 능력과 태그 품질에 의해 좌우됨을 시사한다.

표 4. 도서 대조셋 민감 검출 수 비교

방법	문서	청크	검출 수	검출 문서 수
LLM-only	100	2,479	1	1
Tag-reference	100	2,479	0	0

표 5. 도서 대조셋의 전처리 태그 분포와 실제 민감 판단

전처리 태그	태그 수	청크 수	민감 판단
ADDRESS	565	293	0
ORG	463	281	0
PERSON	253	154	0
DATE	9	8	0
PHONE	1	1	0
RESIDENT_ID_LABEL	3	2	0

표 4 와 표 5 는 공개 도서 본문 대조셋에서 제안 방식의 오탐 억제 경향을 확인한 결과이다. AI Hub 도서자료 기계독해 validation 의 source order 기준 100 개 content_id 문서에서 추출한 2,479 개 문장 청크에서 LLM-only 조건은 공개 교육정책 문맥의 기관장 실명인 박민수 1 건을 민감정보로 판단하였다. 반면 Tag-reference 조건에서는 PERSON 태그가 생성된 동일 문맥에서도 최종 민감 검출이 발생하지 않았다. 해당 조건별 차이는 그림 2 와 그림 3 의 예시에서 확인할 수 있다.

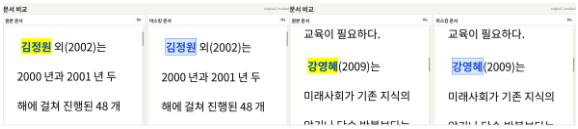


그림 2. 공개 도서 인물명 비민감 판단 사례

그림 3. 박민수 인물명 판단 차이 사례:
(좌) Tag-reference, (우) LLM-only

그림 2 와 그림 3 은 공개 도서 본문 대조셋에서 인물명이 포함된 문맥을 처리한 사례를 나타낸다. 그림 2 에서는 학술 인용 문맥에 등장한 인물명이 LLM-only 와 Tag-reference 조건 모두에서 민감정보로 검출되지 않았다. 반면 그림 3 의 박민수 사례에서는 LLM-only 조건이 공공기관 활동 설명문에 등장한 기관장 실명을 민감정보로 검출하였으나, Tag-reference 조건은 동일 인물명을 기관명 및 직책 정보와 결합된 공개 인물명 문맥으로 해석하여 최종 민감 검출을 억제하였다. 이는 전처리 태그가 단순히 후보 구간을 추가하는 역할에 그치지 않고, LLM 이 후보의 문맥적 성격을 구분하는 참고 신호로 작용할 수 있음을 보여준다.

영업 문서는 여러 사람의 이름, 연락처, 주소, 금액, 직급이 표 형태로 밀집되어 있어 LLM 이 일부 예시 표현이나 비개인정보 표현을 민감정보로 확장 판단할 가능성이 있다. 따라서 전처리 태그는 보호 대상 후보를 단순히 늘리는 장치가 아니라, 개인정보 유형과 문맥을 함께 검토하도록 LLM 입력을 구조화하는 역할을 한다. 또한 LLM 출력 구간과 정답 라벨 구간의 경계가 완전히 일치하지 않는 경우가 많아 정확 구간 기준에서는 더 낮은 점수가 나타날 수 있으므로, 본 논문에서는 마스킹 대상 영역을 포착했는지 확인하기 위한 겹침 구간 기준을 중심으로 결과를 제시한다.

제안 방법은 판단 근거를 민감 구간별로 기록하므로 보안 담당자의 사후 검토와 정책 개선에 활용될 수 있다. 다만 LLM 의 성능 한계, 문서 유형별 편차, 문자 추출 오류, 규칙 관리 비용은 여전히 남아 있다. 특히 이미지 내 작은 글자, 저해상도 캡처, 복잡한 표 구조는 태그 생성과 판단 성능을 낮출 수 있으므로 유형별 평가가 필요하다.

VI. 결론

본 논문은 문서 보안 환경에서 민감정보 탐지를 수행하기 위한 전처리 태그 참조 기반 LLM 문맥 판단 구조를 제안하고, 세 종류의 평가셋에서 이를 검증하였다. 영업 문서 통제 실험에서는 사전 구축된 고품질 후보 태그가 주어졌을 때 전반적으로 F1-score 가 개선되었으며, Qwen 27B 에서는 동일한 recall 을 유지하면서 precision 을 높였다. KDPII 1 차 개인정보 4,700 문장 실험에서는 실제 NER, regex, 규칙 기반 전처리 태그를 사용하였고, Qwen 9B 와 27B 계열에서 Tag-reference 조건이 LLM-only 대비 F1-score 와 FP 억제 측면에서 뚜렷한 개선을 보였다. 도서 대조셋에서는 실제 전처리 태그 1,294 개가 입력되었음에도 Tag-reference 조건의 민감 검출이 0 건으로 나타나, 공개 본문 속 일반 개체명을 곧바로 보호 대상으로 확장하지 않음을 확인하였다.

향후 연구에서는 다양한 문서 유형과 실제 보안 정책을 포함한 평가셋을 구축하고, 표/도표/스캔 문서 처리, PDF 입력 확장, 문자 추출 오류 보정, 정책 기반 판단 고도화를 수행할 계획이다.

참고문헌

[1] L. Fei et al., KDPII: A New Korean Dialogic Dataset for the Deidentification of Personally Identifiable Information, IEEE Access, vol. 12, pp. 135626-135641, 2024

[2] AI Hub, 도서자료 기계독해 데이터셋. https://aihub.or.kr/aihubdata/data/view.do?aihubDataSe=data&dataSetSn=92.

[3] Qwen Team, Qwen3.5 , Hugging Face, 2026. Available: https://huggingface.co/Qwen/Qwen3.5-27B.

[4] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440-8451,

합성이미지 기반 Downstream 분류 모델의 원본(Real) 샘플 멤버십 추론 위험 분석

김효주*, 이혁재O, 김고은**, 박서정***, 류봄****, 우서현*****
*서울여자대학교 정보보호학부 사이버보안전공
O명지대학교 스마트모빌리티공학전공
**중앙대학교 생명과학과/데이터프라이버시융합전공
***숙명여자대학교 컴퓨터과학전공
****이화여자대학교 정치외교학과
*****중앙대학교 산업보안학과
e-mail: agneshyoju@swu.ac.kr*, hjaelee0617@gmail.comO, oldsilver031@naver.com**, sm15168@sookmyung.ac.kr***, bomryu@ewha.ac.kr****, wooseo0304@gmail.com*****

Membership Inference Risk Analysis of Real Training Samples in Synthetic-Image-Based Downstream Classifiers

Hyoju Kim*, Hyukjae LeeO, Goeun Kim**, Seojeong Park***, Bom Ryu****, Seohyeon Woo*****
*Dept. of Cyber Security, Seoul Women's University
ODept. of Smart Mobility Engineering, Myongji University
**Dept. of Life Science/Data Privacy, Chung-Ang University
***Dept. of Computer Science, Sookmyung Women's University
****Dept. of Political Science and International Relations, Ewha Womans University
*****Dept. of Industrial Security, Chung-Ang University

요 약

본 연구는 FFHQ 실제 얼굴과 StyleGAN 합성 얼굴로 학습한 real/fake 분류기에 점수 기반 멤버십 추론 공격(Membership Inference Attack, MIA)을 적용해, 원본 real 학습 샘플의 멤버십 신호가 학습 규모에 따라 어떻게 달라지는지를 측정했다. 모든 학습 규모에서 무작위(AUC 0.5)를 상회하는 신호가 일관되게 관측되었고, 소규모일수록 강한 양상을 보였다. 다만 낮은 오탐 조건의 탐지율은 낮아, 강한 재식별 공격이라기보다 활용 파이프라인에 잔존하는 약한 위험에 가깝다. 이는 합성 이미지가 포함된 downstream 모델에서도 학습 샘플의 약한 멤버십 신호가 잔존할 수 있으며, 합성데이터 활용 단계의 후속 모델까지 공격 기반 검증 대상에 포함할 필요가 있음을 시사한다.

키워드: Membership Inference Attack (MIA), Synthetic Data, Privacy Risk, Data Anonymization, Downstream Classifier

I. 서론

합성데이터는 원본 개인정보를 직접 노출하지 않으면서도 모델 학습에 활용할 수 있다는 점에서 프라이버시 보호의 대안으로 주목받고 있다. 그러나 합성데이터라는 이유만으로 안전성이 보장되는 것은 아니다. 생성모델이 학습 데이터를 부분적으로 기억할 경우[1], 그 흔적은 합성 결과물뿐 아니라 합성데이터를 활용해 학습한 후속 모델에도 남을 수 있다. 국내외 규범은 합성데이터를 자동으로 익명정보로 보지 않고, 실제 재식별 가능성과 공격 가능성을 기준으로 위험을 평가하는 방향을 보인다. 멤버십 추론 공격(Membership Inference Attack, MIA)은 특정 데이터가 모델 학습에 사용되었는지를 모델의 출력 반응을 통해 추론하는 공격으로, 합성데이터의 안전성을 검증하기 위한 방법 중 하나로 논의되고 있다. 그러나 구체적인 정량 기준은 아직 충분히 마련되어 있지 않다. 예컨대 Shokri 등[2]과 Carlini 등[3]은 분류 모델에 대한 MIA 방법론과 평가 지표를 제시하였으나, 합성데이터 활용 환경에서의 임계값 기준은 별도로 논의되지 않았다. 또한 기존 MIA 연구는 주로 생성모델 단계에 집중해 왔으며, 합성데이터가 downstream 모델의 학습 데이터로 재사용되는 활용 단계의 위험에 관한 선행연구 논의는 아직 부족한 상황이다. 여기서 downstream 모델은 합성데이터를 직접 생성하는 모델이 아니라, 합성데이터를 학습 데이터로 활용해 특정 과업을 수행하도록 학습된 후속 모델을 의미한다.

본 연구는 플리커-페이스-HQ(Flickr-Faces-HQ, FFHQ) 실제 얼굴과 StyleGAN 합성 얼굴로 학습한 downstream real/fake 분류기에 점수 기반 MIA를 적용해, 원본 real 학습 샘플의 멤버십 신호가 분류기 출력에 남는지를 학습 규모별로 측정한다. FFHQ는 고해상도 실제 얼굴 이미지로 구성된 공개 데이터셋이며, StyleGAN은 실제 얼굴 이미지의 분포를 학습해 새로운 합성 얼굴 이미지를 생성하는 대표적 생성모델이다. 본 연구에서 말하는 real/fake 분류기는 입력 이미지가 실제 얼굴인지 합성 얼굴인지를 구분하도록 학습된 이진 분류 모델을 의미한다. 즉, 본 연구는 합성 이미지 자체가 원본 이미지를 재현하는지를 직접 분석하는 것이 아니라, 합성 이미지가 포함된 후속 분류 모델의 출력에 원본 real 학습 샘플의 학습 포함 여부를 구분할 수 있는 신호가 남는지를 분석한다.

본 연구의 핵심 가설은 합성 얼굴 이미지를 포함해 학습한 downstream real/fake 분류기에서도 원본 real 학습 샘플의 멤버십 신호가 출력 점수에 잔존하며, 이를 점수 기반 MIA로 탐지할 수 있다는 것이다. 이를 통해 합성데이터 기반 분류 모델의 멤버십 위험을 정량 평가하고, 합성데이터의 법적 지위 및 공격 기반 검증 기준의 공백과 연결해 논의하고자 한다. 이 가설이 확인된다면 합성데이터의 안전성 검증은 생성 단계의 데이터 자체에만 머무르지 않고, 합성데이터가 재사용되는 활용 단계의 후속 모델까지 포함해야 한다는 정책적 시사점을 제공한다.

II. 관련 연구

2.1 생성모델에 대한 멤버십 추론

MIA는 특정 데이터가 모델 학습에 사용되었는지를 모델의 반응을 통해 추론하는 공격이다. 생성모델을 대상으로 한 대표적 연구로 Chen 등[4]의 GAN-Leaks가 있다. 이들은 공격자의 접근 수준에 따라 생성모델 MIA를 분류하고, 학습 샘플이 생성 분포에 더 가깝게 재현될 수 있다는 점을 이용한 거리 기반 공격을 제시했다. van Breugel 등[5]은 생성모델 내부에 접근하지 않고 공개된 합성데이터만으로도 MIA가 가능함을 보였으며, 생성모델의 국소적 과적합이 합성데이터에 멤버십 신호로 남을 수 있음을 설명했다. 이들 연구는 공통적으로 합성데이터를 생성하는 단계의 위험을 평가한다. 국내에서도 생성모델 대상 MIA가 연구되어, 오윤주 등은 Stable Diffusion을 대상으로 한 멤버십 추론 공격을 제시한 바 있다.[6]

2.2 분류 모델에 대한 점수 기반 공격과 과적합

분류 모델을 대상으로 한 MIA는 Shokri 등[2]의 연구를 통해 체계화되었다. 이들은 shadow model을 이용해 target model의 학습 데이터 포함 여부를 추론하는 공격 방법을 제

시했으며, 이후 연구들은 손실이나 신뢰도와 같은 출력 점수만으로도 멤버십 여부를 구분할 수 있음을 보였다. Yeom 등[7]은 과적합과 MIA 취약성의 관계를 분석하여, 학습 손실과 검증 손실의 차이가 클수록 멤버십 정보가 더 쉽게 드러날 수 있음을 밝혔다. 국내에서도 분류 모델 대상 MIA와 과적합의 관계가 정리, 분석된 바 있다.[8] 이는 학습 데이터 규모가 작거나 모델이 학습 데이터에 과도하게 적합 될수록 MIA 위험이 커질 수 있음을 시사한다.

2.3 본 연구의 위치

본 연구와 유사한 흐름의 연구로는 Giomi 등[9]의 Anonymeter와 van Breugel 등[5]의 DOMIAS가 있다. Anonymeter는 유럽 일반 데이터 보호 규정(General Data Protection Regulation, GDPR)의 단일화·연결 가능성·추론 위험 기준에 따라 합성 표 형식 데이터에 대한 공격 기반 평가를 수행하며, DOMIAS는 합성데이터 분포에 남은 생성모델의 과적합 흔적을 이용해 멤버십을 추론한다. 두 연구는 합성데이터의 프라이버시 위험을 공격 기반으로 평가한다는 점에서 본 연구와 연결되지만, 주된 초점은 합성데이터 생성 단계에 있다.

그러나 실제 배포 환경에서는 생성모델의 내부 구조나 파라미터가 공개되지 않고, 합성데이터 또는 이를 활용해 학습된 후속 모델만 외부에 노출될 수 있다[4][5]. 따라서 생성모델 직접 접근을 전제로 한 MIA만으로는 합성데이터 공개 이후 활용 단계에서 발생하는 프라이버시 위험을 충분히 평가하기 어렵다. 본 연구는 이 공백을 보완하기 위해 합성 이미지를 포함해 학습된 downstream real/fake 분류기를 공격 대상으로 삼고, 출력 점수 기반 MIA를 활용 단계의 프라이버시 감사 도구로 적용한다. 또한 Carlini 등[3]이 제안한 위양성률(False Positive Rate, FPR) 1% 조건에서의 진양성률(True Positive Rate, TPR), 즉 TPR@1%FPR 기준을 사용해 낮은 오탐 조건에서의 실질적 탐지 가능성을 함께 해석한다.

III. 합성데이터 관련 법·정책 현황

3.1 국내: 합성데이터의 법적 지위와 검증 공백

현행 「개인정보 보호법」 제58조의2는 시간·비용·기술 등을 합리적으로 고려할 때 더 이상 개인을 알아볼 수 없는 정보에 대해서만 보호법 적용을 제외한다. 따라서 합성데이터가 규제 대상에서 제외되기 위해서는 합성으로 생성되었다는 사실만으로는 부족하고, 재식별이 사실상 어렵다는 점이 확인되어야 한다. 즉, 익명정보에 한하여 동법의 적용에서 제외된다. 개인정보보호위원회의 「합성데이터 생성 참조모델」[10]과 「합성데이터 생성·활용 안내서」[11] 역시 안전성 검증의 필요성을 제시하지만, MIA와 같은 공격 기반 검증의 구체적 지표나 임계값은 명시하지 않는다.

3.2 해외: 실질 위험 기반 판단과 공격 기반 평가

해외에서도 합성데이터는 실질적 식별 가능성을 기준으로 판단된다. GDPR Recital 26[12]은 식별 가능성 판단 시 합리적으로 사용 가능한 모든 수단을 고려하도록 하고, 유럽 데이터 보호 이사회(European Data Protection Board, EDPB) 의견서 28/2024[13]는 AI 모델의 익명성과 모델 쿼리를 통한 개인정보 추출 가능성을 사안별로 검토할 것을 요구한다. 싱가포르 개인정보보호위원회(Personal Data Protection Commission, PDPC) [14]도 합성데이터의 재식별 위험을 멤버십 추론과 속성 추론 등 공격 기반 평가로 점검하도록 권고한다.

국내외 논의는 합성데이터의 안전성을 생성 방식이 아니라 실제 식별 가능성과 공격 가능성을 기준으로 판단해야 한다는 점에서 일치한다. 본 연구는 합성 이미지를 활용해 학습한 downstream 분류기에 MIA를 적용하여, 합성데이터 활용 단계의 멤버십 위험을 정량적으로 확인하고 공격 기반 검증 기준 논의에 필요한 실증적 근거를 제공한다. 이는 합성데이터의 안전성 검증 범위를 생성 단계에서 활용 단계까지 확장해야 한다는 본 연구의 핵심 주장과 연결된다.

IV. 실험 설계 및 방법

본 연구는 실제 얼굴 이미지와 합성 얼굴 이미지 일부를 함

께 학습한 downstream real/fake 분류기를 대상으로, 모델 출력 점수를 이용해 특정 real 이미지가 해당 분류기의 학습 데이터에 포함되었는지를 추론한다.

4.1 데이터셋과 분류 모델

본 연구는 FFHQ(Flickr-Faces-HQ) 실제 얼굴과 StyleGAN 합성 얼굴로 구성된 공개 데이터셋 '140k Real and Fake Faces'[15]를 사용하였다. 이 가운데 실제 얼굴 50,000장과 합성 얼굴 50,000장을 각각 real 및 fake 범주로 정의하였다. 본 연구에서 말하는 '합성데이터 기반 분류 모델'은 생성모델(StyleGAN) 자체가 아니라, StyleGAN으로 생성된 합성 이미지와 FFHQ 실제 이미지를 함께 학습한 downstream real/fake 이진 분류기를 의미한다. 따라서 본 연구는 합성 이미지 자체의 원본 재현 여부가 아니라, 후속 분류기가 학습한 real 샘플에 대해 보이는 출력 signal을 분석한다.

분류기는 ImageNet 사전학습 ResNet34를 기반으로 하며, 기존 1,000-class 출력층을 2-class(real/fake) 출력층으로 교체한 뒤 real/fake 분류 데이터에 맞게 미세 조정하였다. 각 subset에서는 real과 fake 이미지 수를 동일하게 맞추었고, 손상 파일 제거와 ImageNet 통계 기반 정규화를 수행하였다. 소규모 학습에서의 배치 통계 불안정을 줄이기 위해 BatchNorm 층은 eval()과 requires_grad=False로 고정하였다. 최적화에는 Adam과 CosineAnnealingLR을 사용했으며, 검증 손실 기준 early stopping을 통해 best checkpoint를 복원하였다.

본 연구가 real/fake 분류기를 공격 대상으로 삼은 이유는 다음과 같다. 첫째, real/fake 분류기는 딥페이크 및 AI 생성 이미지 탐지 서비스에서 활용되는 대표적인 downstream classifier 형태이며, 일부 서비스는 이진 예측과 confidence score를 반환하는 블랙박스 모델로 제공된다. 따라서 공격자는 모델 내부에 접근하지 않더라도 출력 점수를 관찰할 수 있고, 본 연구의 설정은 실제 배포 환경을 고려한 프라이버시 감사 시나리오로 볼 수 있다. 둘째, 본 연구의 목적은 원본 real 학습 샘플의 학습 포함 여부 누출을 측정하는 것이다. DCR(Distance to Closest Record)은 원본과 합성 이미지 간 유사도를, Attribute Inference와 Model Inversion은 각각 속성 추론과 입력 복원 가능성을 평가한다. 반면 MIA는 특정 샘플의 학습 포함 여부를 직접 평가하는 공격이므로, 본 연구에서는 출력 점수 기반 MIA를 사용하였다.

4.2 멤버십 정의와 데이터 분할

real 이미지는 target model 학습 후보인 member_pool과 학습·검증에 사용하지 않는 holdout nonmember_pool로 50:50 분할하였다. 학습 규모에 따른 위험 변화를 보기 위해 subset 크기는 n = 250, 750, 1,250, 2,500으로 설정했으며, 동일한 data seed 안에서는 작은 subset이 큰 subset에 포함되는 nested 구조로 구성하였다. 각 subset의 90%는 target model 학습에 사용하고, 10%는 early stopping용 validation으로 사용하되 MIA 평가에서는 제외하였다(표 1).

표 1. Subset 크기별 데이터 구성
(real/fake 각 범주, 단위: 장)

n	real train (member)	real validation (평가제외)	fake train	fake validation
250	225	25	225	25
750	675	75	675	75
1,250	1,125	125	1,125	125
2,500	2,250	250	2,250	250

따라서 real 데이터는 분할상 다음 세 범주로 관리된다.

① member: target model 학습에 실제 포함된 real 이미지.

② excluded validation: target model 검증에만 사용된 real 이미지.

③ non-member: target model 학습 및 검증에 사용되지 않은 holdout real 이미지.

공격기의 임계값 선택 과정에서 발생할 수 있는 낙관적 편향을 줄이기 위해, member와 non-member 표본은 각각 attack validation과 attack test로 50:50 분리하였다. attack validation은 공격 학습과 임계값 선택에 사용하고, attack test는

최종 MIA 성능 평가에만 사용하였다. non-member 표본은 holdout에서 member와 같은 수만큼 추출해 두 집단의 크기를 맞췄다. 같은 data seed와 subset 안에서는 모든 분할을 고정했으며, 실험은 data seed 3개와 model run 3회로 반복하였다.

4.3 신호와 공격

각 평가 이미지를 분류기에 입력하여 손실, 신뢰도(정답 범주 확률), 엔트로피, logit margin(real − fake)을 계산하였다. 선행연구에서는 손실 기반 멤버십 추론에서 반복 손실 추정 및 입력 증강에 따른 손실 평균화가 공격 성능을 개선할 수 있음이 보고되었다[1]. 이에 본 연구에서는 입력 증강에 따른 신호 변화를 보기 위해 RandomResizedCrop, 수평 뒤집기, ColorJitter를 적용한 5회 증강 손실의 평균과 표준편차를 추가로 산출하였다. 본 연구의 real/fake 이진 분류 설정에서는 cross-entropy 손실, 정답 범주 신뢰도, logit margin이 서로 강하게 연관된다. 모델이 real class에 높은 신뢰도를 부여할수록 손실은 낮아지고 margin은 커지는 경향이 있기 때문이다. 따라서 손실 기반 공격을 기본 기준선으로 두고, 증강 기반 통계량을 결합한 XGBoost 공격과 비교하였다.

공격 방식은 두 가지이다. 첫째, 손실 임계값 공격은 −loss를 단일 신호로 사용하며, attack validation의 ROC에서 Youden's J를 최대화하는 임계값을 선택하였다. 둘째, XGBoost 학습형 공격은 −loss, −aug_loss_mean, loss − aug_loss_mean, aug_loss_std를 feature로 사용해 XGBClassifier를 학습하였다. 이 공격은 단일 임계값 공격이 포착하기 어려운 증강 기반 신호의 추가 효과를 확인하기 위한 비교 방법이다. 두 공격 모두 attack validation에서 학습 및 임계값 선택을 수행하고, attack test에서만 최종 성능을 평가하였다.

4.4 평가 지표

주 지표는 AUC와 TPR@1%FPR이고, 보조로 precision, recall, attack advantage(max(TPR − FPR))를 보고한다. 해석을 돕기 위해 member와 non-member의 평균 신호 차이(signal gap)와 generalization gap(검증 − 학습 손실)을 함께 계산했다. model run 3회를 평균한 뒤 data seed 3개를 평균하고, 95% 신뢰구간은 t-분포로 산출한다.

V. 실험 결과 및 분석

5.1 멤버십 신호의 존재

그림 1은 subset 크기별 member와 non-member의 손실 및 신뢰도 분포를 비교한 결과이다.

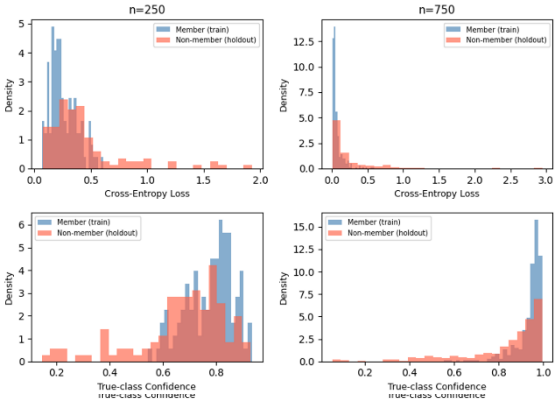


그림 1. Subset 크기별 member(파랑)와 non-member(빨강)의 손실 · 신뢰도 분포

모든 조건에서 member는 non-member보다 낮은 손실과 높은 신뢰도 구간에 더 많이 분포했고, 모든 subset에서 AUC가 0.5를 초과하였다. 이는 분류기 출력 점수에 학습 포함 여부를 구분할 수 있는 신호가 남아 있음을 의미한다. 다만 두 분포가 넓게 겹치고 손실 기반 AUC가 0.63~0.69 수준에 머물러, 강한 멤버십 신호로 보기는 어렵다.

5.2 학습 규모에 따른 변화

학습 규모가 커질수록 멤버십 신호는 전반적으로 감소하였다. 표 2와 그림 2에서 손실 기반 AUC는 n = 250과 n = 750에서 각각 0.688, 0.691로 0.69 안팎을 보였으나, n = 1,250에서 0.647, n = 2,500에서 0.628로 낮아졌다. XGBoost 기반 AUC도 유사한 흐름을 보였고, n = 750에서 0.704로 가장 높았다. 다만 n = 750의 95% 신뢰구간이 [0.558, 0.823]으로 넓어, 이를 안정적인 최고 성능으로 해석하기는 어렵다.

그림 3의 ROC 곡선에서도 모든 조건이 대각선 위에 위치했지만, n이 커질수록 좌상단에서 멀어지는 경향을 보였다. 이는 학습 데이터 증가로 non-member real 이미지에 대한 손실과 신뢰도도 개선되어, member와 non-member의 출력 차이가 줄어들기 때문으로 해석된다. 실제로 loss_sep_d는 0.676에서 0.406으로 감소했고, n = 2,500에서는 신뢰도 격차가 0.09 수준까지 좁혀졌다.

표 2. Subset 크기별 멤버십 추론 성능

(model run 3회 × data seed 3개 평균)

n	AUC (loss) [95% CI]	AUC (XGB)	TPR@1%FPR (loss)	TPR@1%FPR (XGB)	gen. gap
250	0.688 [0.678–0.697]	0.679	8.55%	7.96%	0.094
750	0.691 [0.558–0.823]	0.704	3.45%	4.67%	0.183
1,250	0.647 [0.623–0.672]	0.663	2.05%	3.30%	0.144
2,500	0.628 [0.593–0.663]	0.636	1.63%	2.64%	0.149

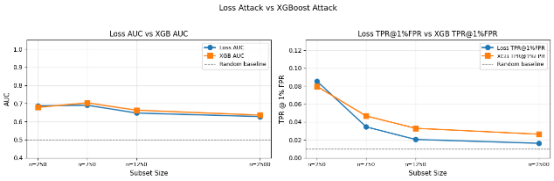


그림 2. Subset 크기에 따른

AUC(왼쪽)와 TPR@1%FPR(오른쪽)

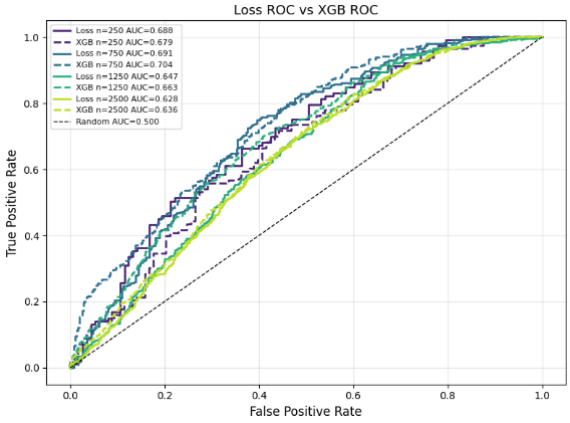


그림 3. Subset 크기별 ROC 곡선

(손실 임계값 공격 대 XGBoost 공격)

5.3 공격 방식 비교와 신호의 해석

두 공격 방식의 성능 차이는 크지 않았다. XGBoost 기반 AUC는 n = 250을 제외한 모든 구간에서 손실 임계값 공격보다 높았지만, 증가 폭은 ΔAUC +0.008~+0.016 수준에 그쳤다. 이는 본 실험에서 관측된 멤버십 신호의 상당 부분이 손실 계열 단일 신호만으로도 포착될 수 있음을 의미한다. 다만 공격의 절대적 탐지력은 제한적이었다. AUC는 0.63~0.70으로 무작위보다 높았지만, TPR@1%FPR은 손실 기준 1.63~8.55%, XGBoost 기준 2.64~7.96%에 그쳤다. 따라서 본 결과는 downstream 모델에 약한 멤버십 신호가 잔존함을 시사한다.

학습 규모별로는 n = 250에서 가장 안정적인 신호가 관측되었다. n = 250은 AUC가 비교적 높고 신뢰구간이 가장 좁았으며, TPR@1%FPR도 두 공격 방식 모두에서 가장 높았다. 반면 n = 750은 점추정 AUC가 근소하게 높았지만 신뢰구간이 넓어 안정적인 최고 성능으로 보기는 어렵다. 또한 n = 250은 generalization gap이 가장 낮음에도 높은 AUC와 TPR@1%FPR을 보여, 멤버십 신호가 전역적 과적합만으로 설명되지는 않음을 시사한다.

VI. 결론 및 정책적 시사점

6.1 주요 결과

본 연구의 실험 결과, 합성 이미지를 포함해 학습한 downstream real/fake 분류기에서도 원본 real 학습 샘플의 멤버십 신호가 관측되었다. 모든 학습 규모에서 AUC가 무작위 수준인 0.5를 초과했으며, 전반적으로 학습 규모가 작을수록 신호가 강하게 나타났다. 다만 신호의 절대적 강도는 제한적이었다. 가장 강한 신호를 보인 n = 250 조건에서도 TPR@1%FPR은 최대 8.55%에 그쳐, 낮은 오탐을 요구하는 현실적 조건에서는 대부분의 member를 탐지하지 못하였다. 따라서 본 연구의 결과는 강한 재식별 공격을 입증한다기보다, 합성데이터 활용 파이프라인에 무시하기 어려운 멤버십 신호가 잔존할 수 있음을 보여준다.

6.2 공격 기반 검증 절차의 필요성

이 결과는 합성데이터의 법적 지위를 판단할 때, 해당 데이터가 합성 방식으로 생성되었다는 사실만으로 익명성을 단정해서는 안 된다는 점을 시사한다. 합성데이터는 원본을 직접 복제하지 않더라도, 생성 과정이나 활용 과정에서 학습

데이터의 통계적 흔적을 일부 반영할 수 있다. 특히 본 연구처럼 합성 이미지를 포함해 학습한 후속 분류 모델에서도 멤버십 신호가 관측된다면, 합성데이터의 안전성은 생성 시점의 데이터 자체뿐 아니라 이를 활용해 학습된 모델까지 포함해 평가되어야 한다. 멤버십 추론이 곧 특정 개인의 재식별을 의미하는 것은 아니지만, 학습 포함 여부가 무작위보다 높은 수준으로 추론될 수 있다는 점은 익명정보 요건을 판단할 때 고려해야 할 위험 신호로 볼 수 있다.

따라서 합성데이터의 공개·활용 이전에는 공격 기반 안전성 검증 절차를 제도적으로 구체화할 필요가 있다. 현행 「합성데이터 생성·활용 안내서」와 「합성데이터 생성 참조모델」은 안전성 검증의 필요성을 제시하지만, 어떤 공격을 사용해 어느 수준까지 검증해야 하는지에 대한 정량 절차는 충분히 구체화되어 있지 않다. 본 연구는 멤버십 추론처럼 실제 공격을 가정한 평가가 이러한 공백을 보완하는 하나의 실증적 방법이 될 수 있음을 보였다. 특히 합성데이터나 이를 활용해 학습한 모델을 외부에 공개하는 경우, 공개 전 단계에서 멤버십 추론과 속성 추론 등 복수의 공격 시나리오를 적용하고, 그 결과를 안전성 판단의 근거로 활용하는 절차가 필요하다.

정책적으로는 단일 지표에 의존하기보다 복수 지표를 활용한 위험 기반 평가 체계를 마련할 필요가 있다. AUC는 전체적인 구분 가능성을 보여주지만, 실제 공격 환경에서는 낮은 오탐 조건에서 얼마나 많은 member를 탐지할 수 있는지가 더 중요할 수 있다. 따라서 AUC와 함께 TPR@low FPR, attack advantage, signal gap 등을 종합적으로 고려하는 평가 기준이 필요하다. 다만 데이터 유형, 공개 범위, 모델 구조, 공격자의 접근 수준에 따라 위험 수준이 달라질 수 있으므로, 획일적 임계값보다는 데이터셋별·활용 목적별 검증을 병행하는 위험 기반 접근이 적절하다.

공격 기반 평가 결과가 일정 수준 이상의 위험을 보이는 경우에는 후속 보완 조치로 이어지는 절차도 함께 설계되어야 한다. 예를 들어 멤버십 추론 위험이 높게 나타나는 경우에는 합성데이터를 재생성하거나, 과도하게 세부적인 패턴을 완화하는 일반화 처리를 적용하거나, 노이즈 추가와 같은 프라이버시 보호 기법을 도입할 수 있다. 또한 데이터나 모델을 완전 공개하기보다 접근 권한을 제한하거나, 연구 목적에 한정된 폐쇄형 분석 환경을 제공하는 방식도 고려할 수 있다. 즉, 합성데이터의 안전성 검증은 위험 수준에 따라 적절한 기술적·관리적 조치로 연결되는 관리 체계로 설계되어야 한다.

얼굴 이미지와 같은 고위험 데이터에 대해서는 더 엄격한 검토가 필요하다. 얼굴 이미지는 개인의 신원과 직접 연결될 수 있을 뿐 아니라, 외부 데이터와 결합될 경우 인구통계적 속성이나 민감한 특성의 추론 가능성으로 이어질 수 있다. 따라서 합성 얼굴 이미지가 원본과 시각적으로 다르더라도, 학습 포함 여부나 속성 추론 가능성이 확인된다면 이를 낮은 수준의 위험으로만 간주하기 어렵다. 본 연구에서 합성 얼굴 이미지를 포함해 학습된 downstream 분류기에서도 원본 학습 샘플의 멤버십 신호가 실증적으로 관측된다는 점에서, 합성데이터의 익명성 판단은 데이터가 합성되었는지 여부가 아니라, 식별 가능성, 연결 가능성, 추론 가능성을 종합적으로 검토하는 방식으로 이루어져야 한다.

6.3 한계 및 향후 연구

본 연구는 단일 데이터셋(FFHQ + StyleGAN)과 ResNet34 기반 분류기에 한정되므로 결과 일반화에 한계가 있다. 또한 공개된 합성 데이터셋을 사용했기 때문에 생성모델의 학습 데이터 기억 여부를 직접 측정하지는 못하였다. 실험 역시 제한된 subset 규모와 반복 횟수에 기반하므로 통계적 안정성에 한계가 있으며, 특히 n = 750 조건에서는 신뢰구간이 넓게 나타났다. 향후 연구에서는 생성모델을 직접 학습해 생성 단계의 memorization과 downstream 활용 단계의 잔여 위험을 함께 측정하고, 다양한 데이터셋, 모델 구조, 공격 강도에서 결과의 일반화 가능성을 검증할 필요가 있다. 이를 통해 합성데이터의 안전성 판단을 생성 시점의 일회성 검토가 아니라, 활용 단계까지 포함하는 지속적 프라이버시 위험 평가 체계로 확장할 수 있을 것이다.

참고문헌

[1] N. Carlini et al., “Extracting Training Data from Diffusion Models,” 32nd USENIX Security

Symposium (USENIX Security 23), pp. 5253–5270, 2023.

[2] R. Shokri et al., "Membership Inference Attacks against Machine Learning Models," 2017 IEEE Symposium on Security and Privacy (S&P), IEEE, pp. 3–18, 2017.

[3] N. Carlini et al., "Membership Inference Attacks From First Principles," 2022 IEEE Symposium on Security and Privacy (S&P), IEEE, pp. 1897–1914, 2022.

[4] D. Chen et al., "GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models," Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, ACM, pp. 343–362, 2020.

[5] B. van Breugel et al., "Membership Inference Attacks against Synthetic Data through Overfitting Detection," Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR, Vol. 206, pp. 3493–3514, 2023.

[6] 오윤주, 박소희, 최대선, "Textual Inversion을 활용한 Adversarial Prompt 생성 기반 Text-to-Image 모델에 대한 멤버십 추론 공격," 정보보호학회논문지, Vol. 33, No. 6, pp. 1111–1123, Dec. 2023.

[7] S. Yeom et al., "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting," 2018 IEEE 31st Computer Security Foundations Symposium (CSF), IEEE, pp. 268–282, 2018.

[8] 오원준, 허준범, "기계학습 모델에 대한 멤버십 추론 공격 분석," 한국정보과학회 학술발표논문집, pp. 1173–1175, 2021.

[9] M. Giomi et al., "A Unified Framework for Quantifying Privacy Risk in Synthetic Data," Proceedings on Privacy Enhancing Technologies (PoPETs), Vol. 2023, No. 2, pp. 312–328, 2023.

[10] 개인정보보호위원회, 「합성데이터 생성 참조모델」, 2024.05.

[11] 개인정보보호위원회, 「합성데이터 생성·활용 안내서」, 2024.12.

[12] European Union, "Regulation (EU) 2016/679 (General Data Protection Regulation, GDPR), Recital 26," 2016.04.

[13] European Data Protection Board (EDPB), "Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models," pp. 1–42, Dec. 2024.

[14] Personal Data Protection Commission (PDPC) Singapore, "Advisory Guidelines on the Handling of Re-identification Risks in Anonymised Data," pp. 1–24, 2018.

[15] Xhlulu, "140k Real and Fake Faces," Kaggle Dataset, 2020. https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces (2026. 05. 23. 방문).

합성 행동 데이터와 이상 탐지 모델 기반 블록체인 스마트 컨트랙트를 활용한 청소년 게임 과몰입 통합 관리 시스템 연구

김지환*(주저자/발표자), 홍지수(숙명여자고등학교), 임경현(The Rectory School),
설순욱**(교신저자)(한국기술교육대학교 전기전자통신공학부)
suseol@koreatech.ac.kr

A Study on an Integrated Adolescent Gaming Overuse Management System Using Synthetic Behavioral Data, Anomaly Detection, and Blockchain Smart Contracts

Jihan Kim*, Jisoo Hong (Sookmyung Women's High School), Kyunghyun Lim(The Rectory School), Soonwook Seol**(Dept. of Electrical, Electronics and Communication Engineering, Korea University of Technology and Education)

요 약 본 연구는 청소년 게임 과몰입 문제를 데이터 기반으로 해결하기 위해 합성 행동 데이터 생성, 머신러닝 기반 이상 탐지, 블록체인 스마트 컨트랙트 자동 제어를 결합한 BGMS(Blockchain-based Game Management System)를 제안한다. 여성가족부·한국콘텐츠진흥원 등 공식 실태조사를 참고하여 설정한 모수로 N=1,000 합성 행동 데이터(정상군 850·과몰입군 150)를 생성하고, Isolation Forest, One-Class SVM, LSTM Autoencoder 세 모델을 비교 평가하였다. 탐지 결과는 Hyperledger Fabric 기반 세 체인코드(게임시간·결제·알림)와 연동되어 청소년·보호자 대상 차등 알림 및 이용 제한이 자동 트리거되도록 설계하였다. 합성 데이터의 분포 가정 하에서 LSTM Autoencoder 는 시계열 패턴 학습을 통해 심야 급증(burst) 및 충동 결제 구간에서 비교 모델 대비 상대적으로 높은 탐지 성능(F1=0.868, AUC-ROC=0.943)을 보였다. 또한 탐지에서 스마트 컨트랙트 트리거까지의 end-to-end 지연이 평균 약 230ms 로 측정되어 실시간 적용 가능성을 점검할 수 있었다. 다만 본 수치는 합성 데이터 환경에 기반한 것으로, 실사용 데이터를 통한 외부 검증이 후속 과제로 남는다. 본 연구는 규제 중심 접근을 보완하는 데이터 기반 보호 모델의 설계 가능성을 제시한다는 데 의의가 있다.

키워드: 청소년 게임 과몰입, 이상 탐지, LSTM Autoencoder, 블록체인 스마트 컨트랙트, 합성 행동 데이터

I. 서 론

세계보건기구(WHO)는 2019 년 ICD-11 에 게임이용장애(Gaming Disorder)를 공식 질병코드 6C51 로 등재하였다[1]. 이는 게임 과몰입이 단순한 생활 습관의 문제가 아닌 의료·공공 정책 차원에서 대응이 필요한 사회적 문제임을 국제적으로 공인한 것이다. 국내에서도 과학기술정보통신부의 스마트폰 과의존 실태조사(2023)에 따르면 10 대 청소년의 과의존 위험군 비율이 성인층보다 유의미하게 높게 나타났으며[2], 여성가족부의 청소년 인터넷·스마트폰 이용습관 진단조사(2023)에서는 과의존 위험군 청소년이 매년 20 만 명 이상으로 확인되었다[5].

기존의 청소년 게임 과몰입 규제 정책은 주로 셧다운제, 플레이타임 제한, 나이 인증 등 제도적 수단에 의존해 왔다. 그러나 이러한 규제는 VPN 우회, 타인 명의 계정 사용 등을 통해 손쉽게 우회될 수 있고, 청소년의 실제 행동 패턴을 실시간으로 감지하고 대응하는 능동적 메커니즘이 부재하다는 한계를 갖는다[3]. 또한 기존 연구들은 설문 기반 데이터를 주로 활용하고 있어 실시간 행동 데이터 분석에 적용하기 어렵다.

본 연구는 이러한 한계를 극복하기 위해 (1) 공식 통계 기반 합성 행동 데이터 생성, (2) 머신러닝 기반 비지도 이상 탐지 모델 비교·평가, (3) 블록체인 스마트 컨트랙트를 활용한 자동 제어를 결합한 BGMS(Blockchain-based Game Management System)를 제안한다. 본 논문의 구성은 다음과 같다. 2 장 관련 연구, 3 장 기존 시스템 문제점 분석, 4 장 연구 방법 및 BGMS 설계, 5 장 실험 결과 및 기대효과, 6 장 결론 순으로 기술한다.

II. 관련 연구

2.1 청소년 게임 과몰입 현황 및 기존 정책

한국콘텐츠진흥원의「게임이용자 실태조사」등 관련 조사에 따르면, 청소년 게임 이용자는 주당 다소 높은 게임 이용 시간을 보이며, 과몰입 위험군으로 분류된 청소년은 정상군에 비해 심야 시간대 게임 이용 빈도가 상대적으로 높은 경향을 보이는 것으로 보고되고 있다[4]. 여성가족부의「청소년 인터넷·스마트폰 이용습관 진단조사」에 따르면, 최근 연도(2023·2024 년) 기준 청소년 인터넷·스마트폰 과의존 위험군은 전체 진단 대상의 약 18% 수준으로 보고되고 있다[5]. 이는 청소년 디지털 과의존 문제가 일부 집단의 예외적 현상이 아니라 상시적 모니터링이 필요한 사회적 이슈임을 시사한다.

기존 규제 정책인 셧다운제(강제적 야간 게임 이용 제한)는 2011 년 도입 이후 실효성 논란이 지속되었으며, 2021 년 청소년보호법 개정으로 폐지가 확정되어 2022 년 1 월 시행이 종료되었다. 대안으로 도입된 게임시간 선택제 역시 보호자 동의 기반의 자율 규제 형태이므로, 실제 청소년의 게임 행동을 실시간으로 감지하고 대응하는 기술적 메커니즘과는 거리가 있다[3]. 이에 행동 데이터 기반의 능동적·기술적 접근의 필요성이 제기되고 있다.

2.2 이상 탐지 모델 관련 연구

이상 탐지(Anomaly Detection)는 정상 패턴에서 벗어난 데이터 포인트를 탐지하는 머신러닝의 하위 분야로, 금융 사기 탐지, 네트워크 침입 탐지, 의료 이상 신호 감지 등 다양한 분야에 적용되어 왔다[6]. 비지도 학습 기반의 대표적 모델로는 Isolation Forest[7], One-Class SVM[8], Autoencoder 계열[9] 등이 있다.

Liu et al.(2008)이 제안한 Isolation Forest 는 이진 트리를 이용하여 정상 데이터와 이상 데이터를 분리하는 방식으로, 고차원 데이터에서도 효율적으로 동작하며 계산 복잡도가 낮다[7]. One-Class SVM 은 정상 데이터의 분포를 초평면으로 정의하고 이를 벗어나는 데이터를 이상으로 분류하며, 소규모 데이터셋에서 안정적인 성능을 보인다[8]. LSTM Autoencoder 는 시계열 데이터를 인코딩하고 재구성하는 과정에서 재구성 오차(MSE)가 큰 구간을 이상으로 탐지하는 방식으로, 시간적 의존성이 있는 행동 데이터에 특히 적합하다[9].

게임 행동 이상 탐지에 관한 선행 연구로는 이은조(2016)의 머신러닝 기반 온라인 게임 보안 시스템 연동 연구[10]와 Kim & Lee(2020)의 머신러닝 기반 스마트폰 과의존 예측 연구[11]가 있으나, 청소년 게임 과몰입 도메인에 특화된 비지도 모델 비교 연구는 아직 미흡하다.

2.3 블록체인 스마트 컨트랙트의 활용

블록체인 스마트 컨트랙트는 조건이 충족될 때 자동으로 실행되는 자기 집행 코드로, 중개자 없이 신뢰할 수 있는 자동화된 거래를 가능하게 한다[12]. Hyperledger Fabric 은 허가형(permissioned) 블록체인 플랫폼으로, 엔터프라이즈 환경에서의 프라이버시 보호와 성능 요구사항을 충족하며, 체인코드(chaincode)를 통한 스마트 컨트랙트 구현을 지원한다[13].

헬스케어, 공급망 관리, 디지털 저작권 관리 등 다양한 분야에서 스마트 컨트랙트 기반 자동화 시스템이 연구·적용되어 왔으나[14], 청소년 게임 관리 분야에의 적용은 국내외에서 거의 연구된 바 없다. 본 연구는 확인된 범위에서 이상 탐지 모델의 출력과 블록체인 스마트 컨트랙트를 연동하여 청소년 게임 과몰입 관리에 적용하는 end-to-end 프레임워크를 제시한다는 점에서 의의가 있다.

III. 기존 시스템 문제점 분석

3.1 규제 중심 접근의 한계

국내 청소년 게임 과몰입 대응 정책은 지난 10 여 년간 주로 규제 입법에 의존해 왔다. 그러나 기술 환경의 급격한 변화와 청소년의 디지털 역량 향상으로 인해 규제 실효성이 지속적으로 저하되고 있다. 표 1 은 기존 규제 시스템의 주요 문제점을 유형별로 정리한 것이다.

표 1. 기존 청소년 게임 과몰입 대응 시스템의 문제점 분석

문제 유형	세부 문제점	영향
규제 우회 용이성	VPN 사용, 타인 명의 계정, 가상번호 인증 등 기술적 우회 수단이 다양하여 실질적 차단 불가	기술적 감지 불가
실시간 대응 부재	행동 데이터를 실시간으로 분석·대응하는 시스템이 없어 사후 조치만 가능	능동적 개입 불가
데이터 부재	개인정보 규제로 실제 청소년 행동 데이터 수집·활용이 극히 제한적	연구 기반 취약
개인 맞춤 부재	일률적 시간 제한으로 개인별 행동 패턴 차이를 반영하지 못함	실효성 저하
투명성·신뢰성 부족	관리 이력 조작 가능성 및 보호자-자녀 간 신뢰 기반 부재	분쟁 해결 곤란

3.2 데이터 기반 접근의 필요성

첫째, 기술적 우회 문제이다. 청소년의 인터넷·스마트폰 이용이 보편화된 환경에서[15], VPN, 가상번호 인증, 타인 명의 계정 사용 등 제도적 규제를 회피하는 수단은 널리 알려져 있고 점차 정교해지고 있다. 따라서 규제 일변도의 대응만으로는 이러한 기술적 우회를 근본적으로 차단하기 어렵다.

둘째, 실시간 대응 부재 문제이다. 현행 게임시간 선택제는 보호자가 사전에 설정한 시간대에만 접속을 제한하는 방식으로, 청소년의 실시간 행동 패턴 변화(심야 급증, 결제 급증(burst, 단시간 다회·고액 결제 집중 현상) 등)에 즉각적으로 대응하지 못한다. 이는 과몰입 초기 단계에서의 조기 개입 기회를 놓치게 하는 구조적 결함이다.

셋째, 데이터 기반 분석의 부재이다. 청소년 개인별 게임 행동 데이터를 수집·분석하여 과몰입 위험도를 정량적으로 산출하고 맞춤형 개입을 실시하는 시스템은 국내에 존재하지 않는다. 개인정보 보호 규제로 인해 실제 행동 데이터 활용이 제한적인 상황에서, 합성 데이터 기반 접근법이 대안으로 부상하고 있다.

넷째, 투명성 및 신뢰성 문제이다. 현행 관리 이력은 중앙 서버에 저장되어 조작 가능성이 존재하며, 보호자-자녀 간 분쟁 발생 시 객관적 증거로 활용하기 어렵다. 불변(immutable) 기록을 보장하는 블록체인 기술의 도입이 이 문제를 근본적으로 해결할 수 있다.

3.3 BGMS 의 차별화 지점

상기 문제점 분석을 토대로, BGMS 는 세 가지 핵심 차별화 전략을 채택하였다. (1) 규제가 아닌 탐지: 행동 데이터를 실시간으로 분석하여 과몰입 패턴을 자동 탐지함으로써 우회 가능성을 낮추는 기술적 보호 메커니즘을 구현한다. (2) 사후 조치가 아닌 선제적 개입: LSTM Autoencoder 의 시계열 패턴 학습을 통해 과몰입 징조를 조기에 포착하고, 스마트 컨트랙트를 통해 즉각적인 보호 조치를 자동 실행한다. (3) 중앙 집중이 아닌 분산 신뢰: Hyperledger Fabric 기반 블록체인에 모든 탐지·조치 이력을 불변 기록으로 저장함으로써 투명성과 신뢰성을 보장한다.

IV. 연구 방법 및 BGMS 시스템 설계

4.1 합성 행동 데이터 생성

실제 청소년 게임 행동 데이터는 개인정보 보호 규제로 인해 접근이 극히 제한적이다. 본 연구는 여성가족부(2023) 및 한국콘텐츠진흥원(2023) 등 공식 실태조사가 보고하는 청소년 디지털 이용 현황을 참고하여, 저자가 설정한 모수(population parameter)를 바탕으로 합성 행동 데이터를 생성하였다.

데이터셋은 N=1,000 명의 가상 청소년으로 구성되며, 정상군(Normal) 850 명(85%)과 과몰입군(Overuse) 150 명(15%)으로 구분하였다. 각 개체에 대해 30 일간의 일별 행동 로그를 생성하였으며, 주요 변수는 표 2 와 같다.

표 2. 합성 데이터 주요 변수 정의

변수명	설명	이상 기준 (과몰입군)
daily_playtime	일별 플레이 시간 (분)	180 분 초과 (정상군 평균 62 분)
daily_payment	일별 인앱 결제액 (원)	50,000 원 이상
late_night_login	자정 이후 접속 여부 (0/1)	주 3 회 이상
payment_burst	단시간 다회 결제 횟수	30 분 내 3 회 이상
weekend_ratio	주말 플레이 비율	0.65 초과
session_count	일별 접속 횟수	10 회 초과
high_value_payment	단일 결제 고액 여부 (원)	200,000 원 이상 (단일 결제)

주) 표 2 의 임계값은 여성가족부·한국콘텐츠진흥원 등 공식 실태조사가 보고하는 청소년 디지털 이용 현황을 참고하되, 합성 데이터 생성을 위해 연구자가 설정한 시뮬레이션 기준값이다.

합성 데이터 생성 시, 정상군은 각 변수를 공식 통계의 이용 현황을 참고하여 저자가 설정한 평균 및 표준편차 기반의 정규분포(normal distribution)로 샘플링하였으며, 과몰입군은 이상 기준을 초과하는 값이 일정 비율 이상 발생하도록 별도의 분포를 설정하였다. 이를 통해 공식 통계의 이용 현황을 참고한 시뮬레이션 데이터셋을 구성하였다.

4.2 이상 탐지 모델 설계 및 평가

본 연구는 세 가지 비지도 이상 탐지 모델을 비교 평가하였다.

(1) Isolation Forest: 무작위 분할을 통한 이진 트리 앙상블 방식으로, 과몰입군의 이상 행동 패턴을 격리하는 데 활용하였다. 오염률(contamination) 파라미터는 실제 과몰입 비율인 0.15 로 설정하였다.

(2) One-Class SVM: RBF 커널을 사용하여 정상군의 데이터 분포를 초평면으로 정의하고, 이를 벗어나는 데이터를 이상으로 분류하였다. 하이퍼파라미터 nu=0.15 로 설정하였다.

(3) LSTM Autoencoder: 30일 시계열 데이터를 입력으로 받아 인코더-디코더 구조를 통해 재구성하며, 재구성 오차(MSE)가 임계값을 초과하는 시점을 이상으로 탐지한다. 인코더는 LSTM(64 units)→LSTM(32 units), 디코더는 RepeatVector→LSTM(32 units)→LSTM(64 units)→TimeDistributed(Dense) 구조로 구성하였다. 학습 시 정상군 데이터만 사용하는 one-class 기반(준지도적 비지도) 접근으로 훈련하였다.

모델 평가는 Precision, Recall, F1-score, AUC-ROC 네 가지 지표를 사용하였으며, 5-fold 교차 검증을 적용하였다. 다만 본 평가에 사용된 합성 데이터는 과몰입군의 이상 기준(예: daily_playtime 180 분 초과, 결제 burst 등)을 명시적으로 사용하여 생성되었으므로, 동일 기준이 모델의 의사결정 임계값 설정에 직접 반영될 경우 자기참조적 평가(self-referential evaluation) 및 라벨 누수(label leakage) 위험이 존재한다. 이에 본 연구는 합성 데이터의 생성 기준과 모델의 학습·임계값 결정 절차를 분리하여, 학습 단계에서는 정상군 데이터만을 사용하고 임계값은 정상군 검증 분할의 95th percentile 통계만으로 결정함으로써 누수 위험을 완화하였다. 다만 합성 데이터 환경의 구조적 한계상 누수 위험을 완전히 배제할 수는 없으며, 실사용 데이터를 통한 외부 검증이 필요하다.

4.3 BGMS 시스템 아키텍처

그림 1 은 BGMS 의 전체 시스템 아키텍처를 도식화한 것이다. BGMS 는 ① 데이터 수집 레이어, ② 이상 탐지 레이어, ③ 자동 제어 레이어(Hyperledger Fabric), ④ 보호 조치 출력의 4 단계 파이프라인으로 구성된다.

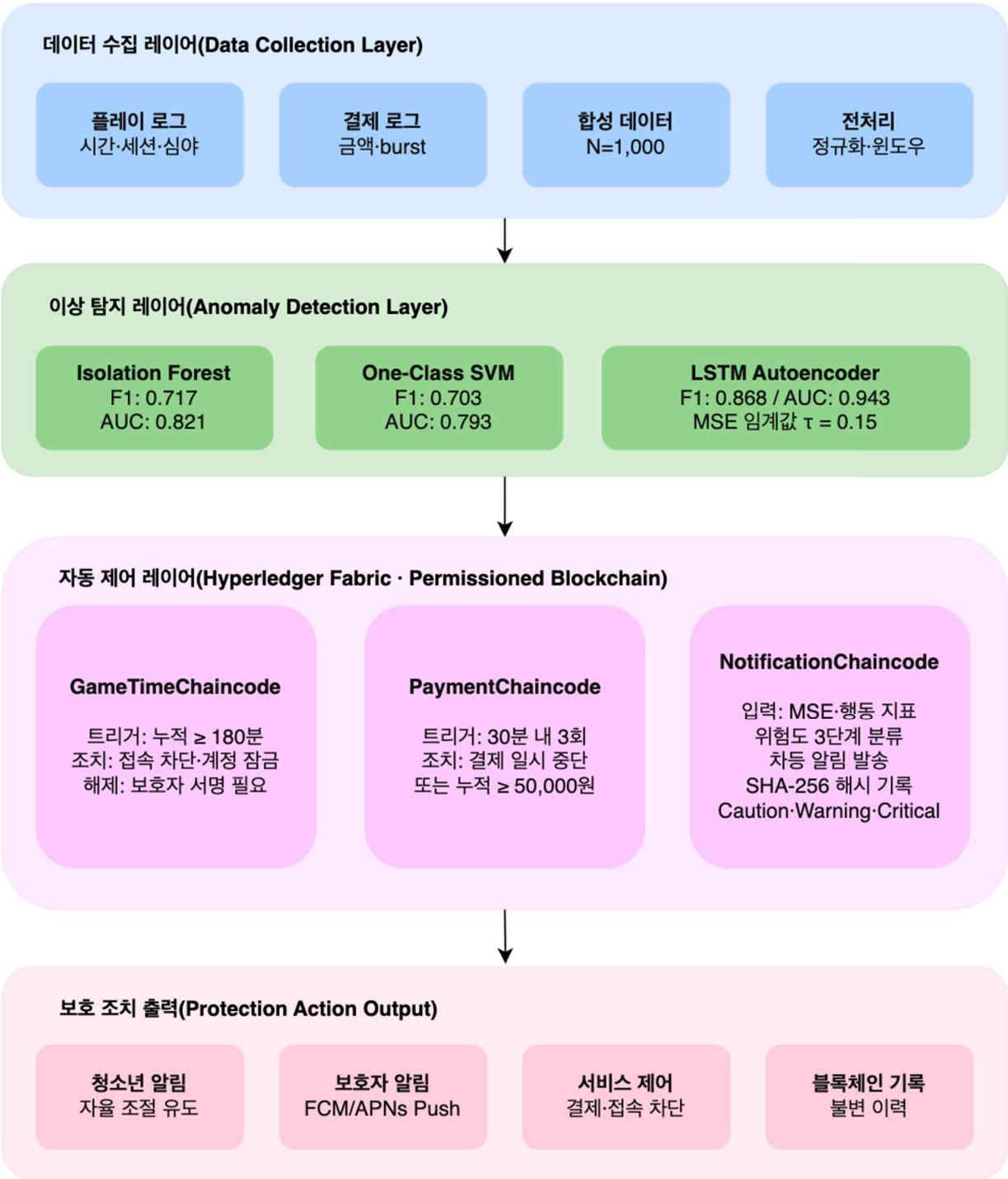


그림 1. BGMS 시스템 전체 아키텍처

4.4 BGMS 상세 동작 메커니즘

■ 1 단계: 데이터 수집 및 전처리. 게임 플랫폼의 플레이 로그(접속 시간·플레이 시간·세션 수), 결제 시스템 로그(결제액·결제 시각·burst 여부), 행동 로그(심야 접속·주말 비율)를 실시간으로 수집한다. 수집된 원시 데이터는 결측값 처리, Min-Max 정규화, 30 일 슬라이딩 윈도우 변환을 거쳐 이상 탐지 모델의 입력 형식으로 전처리된다. 전처리 소요 시간은 평균 10~15ms 이다.

■ 2 단계: LSTM Autoencoder 이상 탐지. 전처리된 30 일 시계열 데이터(shape: [1, 30, 6])가 LSTM Autoencoder 에 입력된다. 인코더는 시계열의 잠재 표현(latent representation)을 32 차원 벡터로 압축하고, 디코더는 이를 다시 원래 시계열로 재구성한다.

재구성 오차(MSE)를 계산하여 사전에 결정된 임계값(threshold=0.15, 검증 데이터 95th percentile 기반)을 초과하면 이상(과몰입 의심)으로 판정한다. LSTM 추론 소요 시간은 평균 45~80ms 이다.

■ 3 단계: 스마트 컨트랙트 트리거. 이상 판정 결과가 Hyperledger Fabric 네트워크의 체인코드 호출 API 로 전달된다. Raft 기반 Ordering Service 를 통해 트랜잭션이 검증·정렬된 후, 피어 노드의 World State DB 에 커밋된다. 트랜잭션 제출-커밋 소요 시간은 평균 120~200ms 이다.

■ 4 단계: 보호 조치 실행 및 기록. 각 체인코드가 해당 보호 조치를 실행한다. GameTimeChaincode 는 게임 서버에 접속 차단 API 를 호출하고, PaymentChaincode 는 결제 게이트웨이에 잠금 요청을 전송하며, NotificationChaincode 는 이상 탐지 결과와 두 체인코드의 실행 이벤트를 종합하여 위험도를 3 단계(주의·경고·위험)로 분류하고, 청소년 및 보호자에게 단계별로 차등화된 메시지를 FCM/APNs 를 통해 실시간 전송한다(상세 프로토콜은 4.5 절 참조). 모든 조치의 실행 이력은 블록체인에 불변(immutable) 기록으로 저장된다. 보호자 앱 알림 전송까지의 전체 end-to-end 평균 지연 시간은 약 230ms 이다.

4.5 위험 등급별 알림 프로토콜 및 공지 메시지 설계

본 시스템의 위험 등급별 알림 프로토콜은 단순한 사후 규제 통보에 머물지 않고, 청소년의 능동적 행동 변화를 유도하는 관점에서 설계되었다. NotificationChaincode 는 4.1 절에서 전처리된 행동 지표 데이터와 이상 탐지 모델이 산출한 재구성 오차(MSE)를 종합하여 위험도를 3 단계로 분류(주의, 경고, 위험; Caution, Warning, Critical)하고, 이를 기반으로 청소년과 보호자 양측에 메시지를 전송한다. 각 단계별 발동 조건과 스마트 컨트랙트 제어 조치는 다음과 같다.

- **Level 1(주의, Caution: 인지 및 자율 조절)**
  - 발동 조건: daily_playtime 이 한계치의 약 60% 수준인 108 분에 도달하였거나, 이상 탐지 모델의 재구성 오차(MSE)가 경계 구간(0.10 이상 0.15 미만)에 진입한 경우
  - 공지 대상: 청소년 클라이언트
  - 메시지 템플릿: **[청소년용]** "현재 게임 이용 시간이 정상군 평균을 상회하고 있습니다. 건강한 게임 이용을 위해 휴식을 권장합니다."
  - 시스템 조치: 계정 잠금과 같은 강제적 스마트 컨트랙트 제어는 보류하며, 청소년이 스스로 과몰입 상태를 인지하고 행동을 조절할 수 있도록 인게임 환경에서 팝업 알림만을 제공한다.
- **Level 2(경고, Warning: 능동적 개입 및 부분 제어)**
  - 발동 조건: daily_playtime 이 한계치의 약 80% 수준인 144 분에 도달하였거나, 이상 탐지 모델의 재구성 오차(MSE)가 확정적 구간(0.15 이상 0.30 미만)에 진입하였거나, payment_burst 가 30 분 간 3 회 이상 감지된 경우
  - 공지 대상: 보호자 애플리케이션
  - 메시지 템플릿: **[보호자용]** "자녀의 계정에서 이상 이용 패턴(장시간 플레이, 충동 결제)이 감지되었습니다. 앱에서 상세 로그를 확인하시고 능동적으로 중재해 주시기 바랍니다."
  - 시스템 조치: PaymentChaincode 가 결제 게이트웨이에 일시 중단 신호를 전송하여 이상 행동을 부분적으로 제어한다. 이와 동시에 FCM/APNs 를 통해 보호자 앱으로 알림을 발송하여 부모의 대화 및 중재 행위를 유도한다.
- **Level 3(위험, Critical: 시스템 보호 및 강제 잠금)**
  - 발동 조건: daily_playtime 이 한계치인 180 분에 도달하였거나, 이상 탐지 모델의 재구성 오차(MSE)가 0.30 이상이거나, 단일 결제 200,000 원 이상의 고액 충동 결제가 발생한 경우
  - 공지 대상: 보호자 애플리케이션
  - 메시지 템플릿: **[보호자용]** "자녀의 일일 이용 시간(180 분) 초과로 스마트 컨트랙트에 의해 계정이 즉시 잠금 처리되었습니다. 본 조치는 보호자의 명시적인 디지털 서명 승인이 있기 전까지 해제되지 않습니다."
  - 시스템 조치: GameTimeChaincode 가 게임 서버에서 접속 차단 API 를 호출하여 계정을 즉시 잠금 처리한다. 잠금 해제는 보호자의 명시적 디지털 서명 승인 전까지 불가능하며, 모든 제어 이벤트는 블록체인에 영구 기록된다.

4.6 체인코드 상세 설계

GameTimeChaincode는 일별 누적 플레이 시간을 실시간으로 모니터링하며, 180분 임계값 초과 시 게임 서버에 계정 잠금

트랜잭션을 자동 발행한다. 잠금 해제는 보호자의 명시적 승인(디지털 서명) 없이는 수행되지 않도록 설계하여 우회 가능성을 낮춘다. PaymentChaincode는 30분 슬라이딩 윈도우 내 결제 횟수 및 누적 금액을 추적하여, burst(3회 이상 또는 50,000원 이상) 발생 시 결제 게이트웨이에 즉시 중단 신호를 전송한다. NotificationChaincode는 이상 탐지 결과와 앞선 두 체인코드의 실행 이벤트를 구독·종합하여 4.5절에 기술한 3단계 위험도 프로토콜에 따라 청소년 및 보호자에게 차등화된 알림을 발송하며, 모든 탐지·조치 이벤트를 SHA-256 해시와 함께 블록체인에 영구 기록한다.

V. 실험 결과 및 기대효과

5.1 이상 탐지 모델 성능 비교

합성 데이터 기반 실험 환경에서 세 가지 이상 탐지 모델의 평가 결과는 표 3 에 제시하였다. LSTM Autoencoder 가 Precision, Recall, F1-score, AUC-ROC 모든 지표에서 비교 모델 대비 가장 높은 성능을 보였으며, 이는 합성 데이터의 분포 가정 하에서의 결과로 한정하여 해석되어야 한다.

표 3. 이상 탐지 모델 성능 비교 (5-fold 교차 검증) (출처: 본 연구 합성 데이터)

모델	Precision	Recall	F1-score	AUC-ROC
Isolation Forest	0.742	0.693	0.717	0.821
One-Class SVM	0.681	0.727	0.703	0.793
LSTM Autoencoder	0.891	0.847	0.868	0.943

LSTM Autoencoder 의 우수한 성능은 30 일 시계열 데이터에 내재된 시간적 의존성(temporal dependency)을 효과적으로 학습하기 때문으로 분석된다. 특히 심야 burst 패턴과 충동 결제 패턴처럼 단기간에 집중되는 이상 행동은 시계열 모델이 정적 모델에 비해 월등히 잘 포착하였다. Isolation Forest 와 One-Class SVM 은 계산 효율성이 높으나, 시간적 맥락 정보를 활용하지 못하는 한계로 인해 낮은 성능을 보였다.

5.2 End-to-End 시스템 지연 시간 분석

합성 데이터 기반 실험 환경에서 BGMS의 실시간 적용 가능성을 점검하기 위해 탐지→스마트 컨트랙트 트리거 end-to-end latency를 측정하였다. 데이터 전처리(10~15ms), LSTM 추론(45~80ms), 체인코드 트랜잭션 제출 및 커밋(120~200ms), 부모 앱 푸시 알림 전송(30~50ms)으로 구성되며, 전체 평균 지연 시간은 약 230ms로 측정되었다. 이는 게임 서비스의 실시간성 요구를 충족하는 수준으로, Hyperledger Fabric의 허가형 합의 메커니즘(Raft ordering)이 낮은 지연 시간을 가능하게 하는 핵심 요인이다.

5.3 스마트 컨트랙트 동작 검증

세 가지 체인코드의 동작을 Hyperledger Fabric 테스트 네트워크 상의 기능 시뮬레이션을 통해 검증하였다. GameTimeChaincode 는 일일 플레이 시간이 180 분을 초과하는 시점에 기능 동작 시뮬레이션 검증에서 정의된 조건에 따라 접속 차단 트랜잭션이 정상적으로 트리거됨을 확인하였다. PaymentChaincode 또한 30 분 내 3 회 이상의 결제 burst 가 발생한 시뮬레이션 시나리오에서 결제 중지 및 알림 트랜잭션이 정의된 조건에 따라 정상적으로 트리거됨을 확인하였다. NotificationChaincode 는 이상 탐지 결과와 앞선 두 체인코드의 실행 이벤트를 종합하여 3 단계 위험도 분류 및 청소년·보호자 대상 차등 알림 발송 로직이 시뮬레이션 시나리오에서 의도대로 동작하였으며, 모든 이벤트가 블록체인 원장에 불변 기록으로 저장되는 기능을 확인하였다. 본 결과는 정형 시나리오 기반의 기능 동작 검증으로서, 실서비스 트래픽·이상 케이스에 대한 추가 검증이 필요하다.

5.4 BGMS 의 기대효과

BGMS 도입을 통해 기대되는 효과는 사회적, 기술적, 정책적 측면에서 다음과 같이 구분할 수 있다.

(1) 사회적 기대효과: 청소년 게임 과몰입 위험군의 조기 탐지 및 개입을 통해 과몰입 진행을 억제하고, 학업·수면·대인관계 등 생활 전반에 미치는 부정적 영향을 감소시킬 수 있다. 부모-자녀 간 신뢰 기반의 디지털 건강 관리 문화를 조성하고, 게임이용장애 치료·상담 수요를 선제적으로 감소시킴으로써 사회적 비용을 절감할 수 있다. 매년 20 만 명 이상으로 추산되는 과의존 위험군[5] 가운데, 본 연구의 가정에 따라 10%만 조기 개입에 성공하더라도 연간 2 만 명 이상의 청소년에게 긍정적 영향을 미칠 수 있다.

(2) 기술적 기대효과: LSTM Autoencoder 기반 시계열 이상 탐지의 실용성을 청소년 행동 분석 도메인에서 실증함으로써, 유사 도메인(스마트폰 과의존, SNS 중독 등) 확장 적용을 위한 방법론적 기반을 제공한다. 블록체인과 AI 이상 탐지를 결합한 end-to-end 프레임워크는 디지털 헬스케어, 금융 사기 탐지, 스마트 계약 기반 자동화 시스템 등 다양한 분야에 응용 가능하다. 평균 230ms 의 처리 지연은 실시간성이 요구되는 게임 서비스에 부합하는 수준으로, 합성 데이터 환경에서의 상용화 가능성을 시사한다.

(3) 정책적 기대효과: 기존 규제 입법 중심의 사후적 접근에서 벗어나, 데이터 기반·기술 중심의 능동적 보호 체계로의 정책 패러다임 전환을 촉진할 수 있다. BGMS 의 블록체인 기반 불변 이력 관리는 개인정보 보호와 투명한 관리 이력을 동시에 달성함으로써, 향후 청소년 디지털 건강 관리 관련 법제 설계에 기술적 근거를 제공할 수 있다. 또한, 게임 플랫폼 사업자·보호자·정부 기관 간의 데이터 공유 및 협력 체계 구축을 위한 표준 프레임워크로 발전시킬 수 있다.

5.5 연구의 한계 및 향후 연구

본 연구는 합성 데이터 기반 실험 환경에서 BGMS의 설계 타당성과 기술적 실현 가능성을 점검한 것으로, 다음과 같은 한계가 존재한다.

(1) 합성 데이터 한계: 본 연구의 모든 정량적 성능 지표(F1=0.868, AUCROC=0.943, end-to-end latency 약 230ms 등)는 공식 통계를 참고하여 설정한 모수 기반 합성 데이터에 기반하므로, 실제 청소년 게임 이용·결제 행동의 분포와 노이즈를 완전히 대표한다고 보기 어렵다. 따라서 본 결과는 실세계 일반화 성능이 아닌 합성 데이터 분포 가정 하에서의 비교 우위로 해석되어야 한다.

(2) 실사용 데이터 및 외부 검증의 필요성: 향후 연구에서는 게임 사업자·교육기관과의 협력을 통해 실사용 행동 로그를 비식별화하여 확보하고, 외부 데이터셋·교차 기관 검증을 통해 모델의 일반화 가능성을 평가할 필요가 있다.

(3) IRB 및 보호자 동의·프라이버시 검토 필요성: 본 시스템은 미성년자의 행동 데이터를 다루므로, 실사용 검증 단계에서는 기관생명윤리위원회(IRB) 심의, 보호자 및 본인 동의 절차, 개인정보 영향평가, 가명·익명 처리 등 「개인정보 보호법」·「청소년 보호법」 및 관련 가이드라인에 따른 사전 검토가 반드시 선행되어야 한다.

(4) 오탐 및 과잉 개입 위험: 이상 탐지 결과에 따른 자동 계정 잠금·결제 중단은 정상 이용자에 대한 오탐(false positive) 시 학습·여가·결제 활동을 부당하게 제약할 수 있다. 따라서 단계별 알림(주의·경고·위험)과 보호자 승인 게이트, 사용자 이의 제기 절차, 정기적 모델 재학습 및 임계값 재조정과 같은 안전장치가 함께 설계되어야 한다.

(5) 블록체인 기반 기록 및 결제 제한의 운영·법적 제약: Hyperledger Fabric 등 허가형 블록체인에 미성년자 행동·결제 이력을 기록하는 경우, 정보주체 권리(열람·삭제·정정) 보장과 불변성(immutability) 간의 충돌, 결제 제한 조치의 약관·결제대행사 책임 분담, 그리고 게임산업법·전자상거래법·전자금융거래법 등과의 정합성에 관한 추가적인 검토가 필요하다.

향후 연구에서는 이상의 한계를 보완하기 위해 (i) 실제 게임 플랫폼·결제 시스템과의 PoC 연동을 통한 실사용 검증, (ii) 강화학습 기반 동적 임계값 조정과 사용자별 맞춤 임계값 학습, (iii) 연합 학습·차분 프라이버시 등 프라이버시 보존 기술과의 결합, (iv) 보호자·청소년 대상 사용성 평가 및 윤리·법적 영향 평가를 단계적으로 수행할 계획이다.

VI. 결 론

본 연구는 청소년 게임 과몰입에 대한 데이터 기반 대응으로 합성 행동 데이터 생성, 비지도 이상 탐지, 블록체인 스마트 컨트랙트 자동 제어를 결합한 BGMS 를 제안하고, 합성 데이터 기반 실험 환경에서 그 설계 타당성과 기술적 실현 가능성을 점검하였다. 기존 규제 중심 접근의 다섯 가지 구조적 한계(우회 용이성, 실시간 대응 부재, 데이터 부재, 개인 맞춤 부재, 투명성 부족)를 분석하고, 각 한계에 대응하는 기술 설계를 BGMS 에 반영하였다.

주요 기여는 다음 세 가지이다. 첫째, 공식 통계를 참고하여 모수를 설정하는 청소년 행동 합성 데이터 설계 방법론을 제시하여, 개인정보 규제 환경에서 실데이터 확보가 어려운 문제를 완화할 수 있는 대안을 마련하였다. 둘째, 세 가지 비지도·시계열 이상 탐지 모델을 게임 과몰입 도메인에 적용·비교하여, 합성 데이터의 분포 가정 하에서 LSTM Autoencoder 가 시계열 행동 패턴 탐지에 상대적으로 우수함(F1=0.868, AUC-ROC=0.943)을 확인하였다. 셋째, 이상 탐지 출력과 블록체인 스마트 컨트랙트를 연동하는 end-to-end 관리 프레임워크를 설계하고, 동일 환경에서 평균 약 230ms 의 처리 지연을 측정함으로써 실시간 적용의 기술적 가능성을 점검하였다. 본 연구의 한계와 후속 과제는 5.5 절에 상세히 정리한 바와 같이, 실사용 데이터 기반 외부 검증, IRB 및 보호자 동의 절차, 오탐·과잉 개입에 대한 안전장치 설계, 블록체인 기록·결제 제한과 관련된 법·제도적 검토 등이 핵심

과제이다. 향후 연구에서는 이러한 과제들을 단계적으로 해결하여 BGMS 의 실용성과 안전성을 함께 높이는 방향으로 발전시킬 계획이다.

참 고 문 헌

[1] World Health Organization, International Classification of Diseases 11th Revision (ICD-11), WHO, 2019.

[2] 과학기술정보통신부·한국지능정보사회진흥원, "2023 년 스마트폰 과의존 실태조사", 2024.

[3] 문화체육관광부·여성가족부, "셧다운제도 폐지 및 청소년의 건강한 게임이용 환경 조성방안", 관계부처 합동 보도자료, 2021. 8. 25.

[4] 한국콘텐츠진흥원, "2023 게임이용자 실태조사", 한국콘텐츠진흥원, 2023.

[5] 여성가족부, "2023 년 청소년 인터넷·스마트폰 이용습관 진단조사", 여성가족부, 2023.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Surveys, Vol. 41, No. 3, pp. 1-58, 2009.

[7] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," IEEE International Conference on Data Mining (ICDM), pp. 413-422, 2008.

[8] B. Schölkopf et al., "Estimating the Support of a High-Dimensional Distribution," Neural Computation, Vol. 13, No. 7, pp. 1443-1471, 2001.

[9] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long Short Term Memory Networks for Anomaly Detection in Time Series," ESANN 2015, pp. 89-94, 2015.

[10] 이은조, "온라인 게임 보안을 위한 머신러닝 시스템 연동 방안", 정보보호학회지, Vol. 26, No. 3, pp. 30-37, 2016.

[11] Y. J. Kim and D. S. Lee, "Predictive Analysis of Problematic Smartphone Use by Machine Learning Technique," Journal of the Korea Society of Computer and Information, Vol. 25, No. 2, pp. 213-219, 2020.

[12] N. Szabo, "Smart Contracts: Building Blocks for Digital Markets," Extropy Journal, 1996.

[13] E. Androulaki et al., "Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains," EuroSys'18, pp. 1-15, 2018.

[14] A. Ekblaw, A. Azaria, J. D. Halamka, and A. Lippman, "A Case Study for Blockchain in Healthcare: 'MedRec' Prototype for Electronic Health Records and Medical Research Data," Proc. IEEE Open & Big Data Conference, 2016.

[15] 한국인터넷진흥원, "2022 인터넷 이용 실태조사", 한국인터넷진흥원, 2022.

설문 기반 개인정보 수집·암호화·자동 파기 보조 시스템 설계 및 구현※

유승우O, 김수현*
O순천향대학교 컴퓨터소프트웨어공학과(학부생)
*순천향대학교 컴퓨터소프트웨어공학과(교수)
e-mail:20233532@sch.ac.krO, kimsh@sch.ac.kr*

Design and Implementation of a Survey-Based Personal Information Collection, Encryption, and Automated Deletion Assistant System

Seung-Woo YuO, Su-Hyun Kim*
ODept. of Computer Software Engineering, SoonChunHyang University
(Undergraduate student)
*Dept. of Computer Software Engineering, SoonChunHyang University
(Professor)

요 약

개인정보 보호법 제21조에서는 보유 기간이 경과하거나 처리 목적이 달성된 개인정보를 지체 없이 파기하도록 규정하고 있다. 그러나 Google Forms, SurveyMonkey 등 기존 설문 플랫폼은 자동 파기 기능을 제공하지 않아 법적 의무 이행이 관리자의 수작업에 의존하고 있다. 또한 응답 데이터에 대한 암호화가 미흡하고 열람·복사 행위에 대한 감사 기록이 존재하지 않아 내부 유출 및 무단 접근에 취약하다. 이러한 한계는 응답자가 자신의 정보가 언제, 어떻게, 어디에 사용되고 파기되었는지 확인할 수 없는 투명성 부재 문제로도 이어진다. 본 연구는 이러한 문제를 해결하기 위해 설문 기반 개인정보 수집·암호화·자동 파기 보조 시스템을 설계하고 구현하였다. 제안 시스템은 AWS KMS와 AES-256-GCM을 결합한 Envelope Encryption으로 응답 데이터를 보호하고, NIST SP 800-88 Crypto Erase 원칙에 따라 설문 만료 시 복호화 키를 자동으로 소거하여 물리적 미디어 파괴 없이 복구 불가능한 파기를 실현한다. 또한 SHA-256 해시체인 구조의 감사 로그를 통해 데이터 생성·열람·복사·파기 전 과정의 무결성을 보장하며, 응답자가 본인 데이터의 사용 이력을 직접 확인할 수 있는 투명성을 제공한다.

키워드: 개인정보 보호법, 설문 플랫폼, Envelope Encryption, 암호학적 파기, 해시체인 감사 로그

※ "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지원을 받아 2026년도 AI중심대학(2026-0-00032)의 결과로 수행되었음"

I. 서론

디지털 환경의 확산과 함께 설문조사는 학술 연구, 고객 만족도 조사, 인사 관리 등 다양한 분야에서 개인정보를 수집하는 주요 수단으로 자리 잡고 있다. 그러나 수집된 개인정보의 보호와 파기에 관한 기술적 조치는 여전히 미흡한 실정이다.

개인정보 보호법 제21조는 보유 기간이 경과하거나 처리 목적이 달성된 개인정보를 지체 없이 파기하도록 규정하고 있으며, 2023년 개정을 통해 영리 목적 여부와 무관하게 업무상 타인의 개인정보를 처리하는 개인도 파기 의무 대상에 포함되었다[1]. 그러나 실제로는 법적 의무 이행이 제대로 이루어지지 않는 경우가 빈번하다. 2023년에는 탈퇴 회원 및 보유 목적이 종료된 개인정보를 파기하지 않은 3개 기업에서 개인정보가 유출되어 제재를 받은 사례가 보고된 바 있다[2].

이러한 문제의 근본적인 원인은 기존 설문 플랫폼의 구조적인 한계에 있다. Google Forms, SurveyMonkey, Qualtrics 등 널리 사용되는 설문 플랫폼은 자동 파기 기능을 제공하지 않아 법적 의미 이행이 관리자의 수작업에 전적으로 의존하고, 응답 데이터에 대한 암호화가 미흡하여 DB(DataBase) 탈취 시 데이터가 즉시 노출된다. 또한 데이터 열람·복사·파기 행위에 대한 감사 기록이 존재하지 않아 내부 유출 발생 시 책임 소재를 확인하기 어려우며, 응답자가 자신의 정보가 언제 어떻게 사용되고 파기되는지 확인할 수 없어 개인정보 자기 결정권이 보장되지 않는다.

본 연구에서는 이러한 문제를 네 가지 문제를 해결하기 위해 설문 기반 개인정보 수집·암호화·자동 파기 보조 시스템을 설계하고 구현하였다.

제안 시스템은 세 가지 핵심 기술을 중심으로 구성된다. 첫째, AWS KMS(Amazon Web Services Key Management Service)와 AES-256-GCM(Advanced Encryption Standard 256-bit Galois/Counter Mode)을 결합한 Envelope Encryption을 적용하여 응답 데이터를 저장 구간에서 완전히 보호한다. 둘째, NIST SP (National Institute of Standards and Technology Special Publication) 800-88 Crypto Erase 원칙에 따라 설문 만료 시 복호화 키를 자동으로 소거함으로써 물리적 미디어 파괴 없이 즉각적이고 복구 불가능한 파기를 실현한다. 셋째, SHA-256(Secure Hash Algorithm 256-bit) 해시체인 구조 감사 로그를 통해 데이터 생명주기 전반의 무결성을 보장하고 응답자에게 데이터 사용 이력의 투명한 열람 수단을 제공한다.

본 연구의 구성은 다음과 같다. 2장에서 관련 법령 및 기존 연구를 검토하고, 3장에서는 제안 시스템의 전체 구조와 핵심 기능을 기술한다. 4장에서는 구현 결과를 분석하고, 5장에서 결론 및 향후 연구 방향을 제시한다.

II. 배경 및 관련 연구

2.1 법적 배경

개인정보 보호법 제21조는 개인정보처리자가 보유 기간의 경과, 개인정보의 처리 목적 달성 등 개인정보가 불필요하게 된 경우 즉시 해당 개인정보를 파기하도록 규정하고 있다. 파기 방법은 동법 시행령 제16조 제1항에 따라 전자적 파일 형태의 경우 복원이 불가능한 기술적 방법으로 영구 삭제해야 하며, 이를 위반할 경우 규모나 목적에 관계 없이 법적 책임이 발생할 수 있어 기술적 파

※ "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지원을 받아 2026년도 AI중심대학(2026-0-00032)의 결과로 수행되었음"

기 조치의 구비가 필요하다[3].

2.2 기존 설문 플랫폼의 한계

기존 설문 플랫폼은 직관적인 인터페이스와 높은 접근성을 바탕으로 널리 활용되고 있으나, 개인정보 보호 측면에서 공통적인 한계를 지닌다.

자동 파기와 파기 증명 기능이 전무하여 개인정보 보호법상 파기 의무를 기술적으로 이행할 수 없으며, 데이터 열람·복사 행위에 대한 감사 로그가 존재하지 않아 내부 유출 발생 시 책임 소재 확인이 불가능하다. 또한 응답자가 본인 데이터의 사용 이력을 확인할 수 있는 수단이 제공되지 않아 개인정보 자기 결정권의 실질적 보장이 어렵다.

2.3 관련 기술 연구

2.3.1 Envelope Encrption

Envelope Encryption은 데이터 암호화 키 DEK(Data Encryption Key)로 실제 데이터를 암호화하고, 별도의 키 암호화 키 KEK(Key Encryption Key)로 DEK를 암호화하여 보관하는 계층적 키 관리 기법이다. 이 방식은 대량의 데이터를 효율적으로 암호화하면서도 키 관리를 중앙화할 수 있다는 장점이 있다. AWS KMS는 HSM(Hardware Security Module)기반의 CMK(Customer Master Key)를 통해 DEK의 생성과 복호화를 관리하며, CMK는 외부로 유출되지 않는다. Envelope Encryption의 보안성은 DEK의 무작위성과 IV(Initialization Vector)의 유일성에 의존하며, 특히 AES-256-GCM 모드에서는 동일한 키와 IV 조합의 재사용이 인증 태그 유출로 이어질 수 있어 매 암호화마다 독립적인 무작위 IV 생성이 필수적이다[4].

2.3.2 Crypto Erase

NIST SP 800-88 Rev.1은 미디어 삭제 방법으로 Clear, Purge, Destroy 세 가지를 정의하며, 그 중 Crypto Erase는 암호화된 데이터의 복호화 키를 소거하여 데이터를 복구 불가능한 상태로 만드는 기법이다. 물리적 미디어 파괴 없이도 즉각적인 파기 효과를 달성할 수 있어 클라우드 환경에서 특히 유효하다. 암호화 키가 소거된 이후 암호문이 저장 매체에 잔존하더라도 복호화가 물리적으로 불가능하므로, 이를 개인정보 보호법상 파기에 준하는 기술적 조치로 적용할 수 있다[5].

2.3.3 해시체인 기반 감사 로그

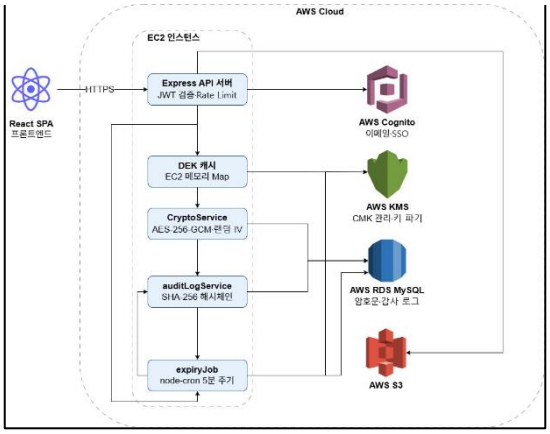
해시체인은 각 레코드가 직전 레코드의 해시값을 포함하도록 구성하여 사후 위변조를 탐지할 수 있는 데이터 구조이다. 블록체인의 핵심 원리와 동일한 단방향 연결 구조로, 특정 레코드를 수정하면 이후 모든 레코드의 해시값이 연쇄적으로 불일치하게 되어 위변조 사실이 드러난다. 감사 로그에 해시체인을 적용하면 관리자에 의한 로그 조작을 탐지할 수 있어 내부 위협 대응에 효과적이다. SHA-256은 NIST FIPS(Federal Information Processing Standard) 180-4에서 표준화된 해시 함수로, 충돌 저항성과 역상 저항성이 검증되어 로그 무결성 보장에 적합하다[6].

III. 제안 시스템

3.1 시스템 개요

본 연구에서 제안하는 시스템은 설문 응답 데이터의 수집·암호화·자동 파기·감사 기록을 하나의 플랫폼에서 처리하는 개인정보 보호 중심의 설문 보조 시스템이다. [그림 1]은 제안 시스템의 전체 아키텍처를 나타낸다.

※ "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지원을 받아 2026년도 AI중심대학(2026-0-00032)의 결과로 수행되었음"



[그림 1. 시스템 전체 아키텍처]

시스템은 React 기반 프론트엔드, Node.js/Express 백엔드, AWS RDS MySQL(Relational Database Service) 데이터베이스, AWS Cognito 인증, AWS KMS 키 관리, AWS S3 파일 저장소로 구성된다. 프론트엔드는 Cloudeflare Pages를 통해 CDN(Content Delivery Network)으로 배포되며, 백엔드와의 통신은 Cloudflare Tunnel을 통해 HTTPS(HyperText Transfer Protocol Secure)로 라우팅 된다.

3.2 Envelope Encryption 기반 암호화

3.2.1 키 계층 구조

제안 시스템은 AWS KMS의 CMK와 설문별 DEK로 구성되는 2계층 키 구조를 채택한다. CMK는 AWS KMS 내부 HSM에만 존재하며 외부로 유출되지 않는다. DEK는 설문 생성 시 *KMS GenerateDataKeyCommand*를 통해 AES-256 키로 생성되며, 평문 DEK는 서버 RAM(Random Access Memory)의 dekCache에 캐싱 되고 암호화된 DEK는 DB의 surveys.encrypted_dek 컬럼에 저장된다. [그림 2]는 키 계층 구조를 나타낸다.

[그림 2. Envelope Encryption 키 계층 구조]

3.3 NIST SP 800-88 기반 자동 파기

3.3.1 파기 원리

제안 시스템의 파기 방식은 NIST SP 800-88 Rev.1의 Crypto Erase 원칙을 따른다. 암호문은 DB에 잔존하지만 복호화에 필요한 키를 소거함으로써 물리적 미디어 파괴 없이 즉각적이고 복구 불가능한 파기를 실현한다. 키가 존재하지 않는 암호문은 현실적인 시간 내에 복호화가 불가능하므로 개인정보 보호법상 파기에 준하는 기술적 조치로 간주할 수 있다.

3.3.2 자동 파기 실행 순서

node-cron 스케줄러가 1분 주기로 만료 설문을 탐지하여 파기를 수행한다. *expires_at ≤ NOW() AND status='active'* 조건을 만족하는 설문이 탐지되면 DB 트랜잭션 내에서 *status='destroyed', encrypted_dek=NULL*으로 업데이트하고 감사 로그를 기록하고 dekCache에서 해당 설문의 평문 DEK를 제거하며, *KMS ScheduleKeyDeletion*은 트랜잭션 외부에서 비동기로 호출하여 트랜잭션 타임아웃을 방지한다.

3.4 SHA-256 해시체인 감사 로그

3.4.1 해시 계산 구조

각 감사 로그는 id, surveyId, actorId, action, detail, prevHash, createdAt의 7개 필드를 JSON(JavaScript Object Notation) 형식으로 SHA-256으로 해싱하여 log_hash를 생성한다. 각 로그는 직전 로그의 log_hash를 prevHash로 포함하는 단방향 체인 구조를 형성하며, 최초 로그

※ "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지원을 받아 2026년도 AI중심대학(2026-0-00032)의 결과로 수행되었음"

는 'GENESIS" 문자열을 prevHash로 사용한다. 해시 계산식은 다음과 같다.

$$H_0 = SHA256(id \parallel surveyId \parallel actorId \parallel action \parallel detail \parallel "GENESIS" \parallel createdAt)$$

$$H_n = SHA256(id \parallel surveyId \parallel actorId \parallel action \parallel detail \parallel H_{(n-1)} \parallel createdAt)$$

3.4.2 위변조 탐지

verifyChain 함수는 전체 감사 로그를 순차적으로 순회하며 hash_mismatch, chain_break 두 가지 유형의 위변조를 탐지한다. hash_mismatch는 로그 필드가 수정되어 재계산한 해시값이 저장된 log_hash와 불일치하는 경우를 나타내며, chain_break는 로그가 삭제되어 다음 로그의 prev_hash가 직전 로그의 log_hash와 불일치 하는 경우를 나타낸다. 탐지 결과는 {valid, brokenAt, reason} 형태로 반환되며, 설문 소유자와 응답자 모두 API(Application Programming Interface)를 통해 무결성 검증을 수행할 수 있다.

3.5 사용자 정보 암호화 및 자동 채움

사용자의 이름, 전화번호, 주민등록번호는 설문 DEK와 분리된 전용 DEK로 AES-256_GCM 암호화하여 "iv:authTag:ciphertext" 형식으로 DB에 저장된다. 설문 문항에 *piType(name/phone/residentNumber)* 속성이 설정된 경우, 응답자가 해당 문항에 접근하면 복호화된 개인정보가 자동으로 채워진다. 설문 DEK와 분리하여 관리하는 이유는 설문 파기 시 사용자 개인정보가 함께 소거되지 않도록 파기 범위를 설문 응답으로 한정하기 위함이다.

3.6 인증 및 권한 모델

인증은 AWS Cognito 기반 이메일/비밀번호 로그인과 Google OAuth 2.0(Open Authorization) 소셜 로그인을 지원한다. JWT(JSON Web Token) 검증 미들웨어는 Cognito 공개키로 서명을 검증한 후 *req.user={sub, email}*을 주입한다. 권한 모델은 creator_id 일치 여부로 설문 소유자를 판단하며, <표1>과 같이 구성된다.

<표1. 권한 모델>

대상	허용 행위
인증된 모든 사용자	설문 생성, 응답 제출
설문 소유자	마감, 응답 열람, 증빙 업로드, 설문 삭제, 무결성 검증
설문 응답자	본인 응답 설문의 감사 로그 열람, 무결성 검증, 증빙 열람

IV. 제안 시스템 분석

4.1 Envelope Encryption 분석

제안 시스템은 AES-256-GCM 모드를 적용하여 응답 데이터의 기밀성과 무결성을 동시에 보장한다. 매 응답마다 *crypto.randomBytes(12)*로 독립적인 IV를 생성하여 동일 DEK 재사용 환경에서도 IV 재사용 취약점을 방지한다. CMK는 AWS KMS 내부 HSM에만 존재하여 외부 유출이 불가능하며, DEK는 설문마다 독립 생성되므로 특정 설문의 DEK가 노출되더라도 다른 설문 데이터는 영향을 받지 않는다.

4.2 자동 파기 분석

파기는 *encrypted_dek=NULL* 처리, dekCache 제거, *KMS ScheduleKeyDeletion*의 세 단계로 순차적으로 수행된다. 첫 번째 단계 완료 시점부터 복호화가 불가능한 상태가 확보되므로 KMS 장애

※ "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지원을 받아 2026년도 AI중심대학(2026-0-00032)의 결과로 수행되었음"

등으로 세 번째 단계가 지연되더라도 파기의 실질적 효과는 즉시 발생한다. 파기 실행 시 감사 로그에 destroyed 액션과 함께 *status='destroyed'*로 상태가 기록되어 외부 감사 시 파기 일시와 행위자를 증명할 수 있다.

4.3 해시체인 감사 로그 분석

SHA-256의 충돌 저항성에 의해 특정 로그 레코드를 수정하면 이후 모든 해시값이 연쇄적으로 불일치하여 위변조가 즉시 탐지된다. 로그 필드 수정 시 hash_mismatch, 레코드 삭제 시 chain_break로 구분하여 탐지 결과를 반환하며, 설문 소유자와 응답자 모두 API를 통해 무결성 검증을 수행할 수 있다. 응답자는 본인 데이터의 생성·열람·복사·파기 이력을 확인할 수 있어 기존 플랫폼에서 제공하지 않는 투명성을 제공한다.

4.4 기존 플랫폼 대비 비교

본 연구는 서론에서 제기한 네 가지 문제를 다음과 같이 해결하였다. 첫째, 자동 파기 부재 문제는 node-cron 기반 만료 탐지와 Crypto Erase를 결합하여 관리자 개입 없이 파기 의무를 자동으로 이행함으로써 해결하였다. 둘째, 암호화 미흡 문제는 Envelope Encryption과 AES-256-GCM을 적용하여 DB 탈취 시에도 복호화 키 없이는 데이터 접근이 불가능하게 함으로써 해결하였다. 셋째, 감사 기록 부재 문제는 SHA-256 해시체인 감사 로그를 통해 데이터 생성·열람·복사·파기 전 과정을 기록하고 위변조 탐지를 가능하게 함으로써 해결하였다. 넷째, 투명성 부재 문제는 응답자가 본인 데이터의 사용 및 파기 이력을 직접 확인할 수 있도록 함으로써 해결하였다.

V. 결론

본 연구는 기존 설문 플랫폼이 개인정보 보호법 제21조의 파기 의무를 기술적으로 이행하기 어렵다는 문제를 해결하기 위해 설문 기반 개인정보 수집·암호화·자동 파기 보조 시스템을 설계하고 구현하였다. AWS KMS와 AES-256-GCM 기반 Envelope Encryption으로 응답 데이터를 보호하고, NIST SP 800-88 Crypto Erase 원칙에 따른 키 자동 소거로 복구 불가능한 파기를 실현하였으며, SHA-256 해시체인 감사 로그로 무결성과 투명성을 보장하였다. 향후 연구로는 DEK 캐시 TTL(Time To Live) 도입, 파기 E2E(End-to-End) 및 위변조 탐지에 대한 테스트 자동화, DB 관리자의 로그 전체 재작성 공격에 대응하기 위한 외부 불변 저장소 연동을 제안한다.

참고문헌

[1] 국가법령정보센터, "개인정보 보호법 제21조," https://www.law.go.kr/, 2023.

[2] 머니투데이, “부실관리·개인정보 미파기... 3개사서 120만명 개인정보 유출,” https://www.mt.co.kr/tech/2023/10/26/2023102609161591945, Oct. 26, 2023.

[3] 국가법령정보센터, “개인정보 보호법 시행령 제16조,” https://www.law.go.kr/, 2023.

[4] Amazon Web Services, “AWS Key Management Service Developer Guide,” https://docs.aws.amazon.com/kms/, 2024.

[5] NIST, “Guidelines for Media Sanitization,” NIST SP 800-88 Rev.1, Dec. 2014.

[6] NIST, “Secure Hash Standard (SHS),” FIPS PUB 180-4, Aug. 2015.

※ "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지원을 받아 2026년도 AI중심대학(2026-0-00032)의 결과로 수행되었음"

주파수 융합과 가우시안 적응 주의를 활용한 위장 객체 분할 연구

한인철*(발표자), 김수균O
*O제주대학교 컴퓨터공학과
e-mail:{ironin0923, nicesk}@gmail.com*O

Research on Camouflaged Object Segmentation Using Frequency Fusion and Gaussian Adaptive Attention

In-Chul Han*, Soo-Kyun KimO
*O Dept. of Computer Engineering, Jeju National University

요 약

위장 객체 분할은 객체와 배경의 색상, 질감, 구조적 단서가 유사하여 일반적인 객체 분할보다 높은 수준의 문맥 이해와 경계 복원이 요구되는 과제이다. 본 논문에서는 위장 객체 분할을 위해 Swin Transformer 기반의 다중 스케일 특징 추출 구조를 활용하고, 주파수 영역 특징 융합과 가우시안 적응 주의 기법을 결합한 분할 모델을 제안한다. 고수준 의미 특징과 저수준 공간 특징을 각각 정제한 뒤, 주파수 영역에서 두 특징의 크기 정보를 적응적으로 융합하여 위장 객체의 전역적 구조 단서를 보강한다. 또한 가우시안 기반 적응 샘플링과 주의 연산을 통해 불분명한 객체 경계와 세부 영역을 보완하도록 설계한다. 제안하는 구조는 COD10K 데이터세트에서 0.9163의 E-measure와 0.77의 WF-measure를 달성하여, 가우시안의 불확실성 샘플링이 복잡한 배경 속에서도 위장 객체의 위치와 형태를 효과적으로 포착할 수 있음을 보인다.

키워드: 위장 객체 분할, Swin Transformer, 주파수 융합, 가우시안 주의, COD10K

I. 서론

위장 객체 분할 (Camou-flaged Object Segmentation, COS) 과제는 주변 환경과 시각적으로 유사한 객체를 픽셀 단위로 식별하는 과제이며, 전경 객체와 배경이 유사한 시각적 특징을 공유하기 때문에 정확하게 구분하기 어렵다. 따라서 위장 객체를 효과적으로 분할하기 위해서는 전역적인 의미 정보와 함께 세밀한 경계 및 지역적 구조 정보를 함께 활용하는 것이 중요하다. 이러한 COS 연구에서는 고해상도 입력, 반복적 정제, 지역 영역 재처리, 복잡한 디코더 구조 등을 통해 객체와 배경 간의 미세한 차이를 강조하고자 하였다. 그러나 이러한 방법들은 다단계 추론이나 복잡한 특징 상호작용으로 인해 모델 및 추론 복잡도가 증가하는 한계가 있다. 본 연구에서는 이를 완화하기 위해 적응형 가우시안 통합 모델, GINet (Adaptive-Gaussian Integration Network)를 제안한다. GINet은 계층적 백본에서 추출한 다중 스케일 특징을 가우시안 기반 적응 통합 모듈로 결합하고, 주파수 영역 융합과 가우시안 주의 모듈인 GAT (Gaussian Attention Transformer) 기반의 디코딩을 통해 경계 보존과 특징 정보 통합을 동시에 수행한다.

본 연구는 교육부와 제주특별자치도의 지원을 받아 제주 RISE 센터를 통 해 Regional Innovation System and Education (RISE) 사업의 지원을 받아 수행된 연구임. (RISE_2-2-11)

II. 관련 연구

데이터사이언스는~~ (신명조, 9) Fan의[1] 연구는 이웃 특징 연결 디코더와 그룹 반전 주의 기법을 통해 위장 객체의 위치 탐색과 세부 영역 정제를 수행한다. Huang의[2] 연구는 특징 수축 디코더를 통해 인접 단계의 특징을 점진적으로 통합하여 위장 객체의 단서를 강화한다. Pang의[3] 연구는 사람의 확대 및 축소 관찰 방식을 착안하여 삼중 혼합 스케일 구조를 사용하여 다양한 크기의 위장 객체에 대한 판별적 표현을 학습한다.

III. 제안방법

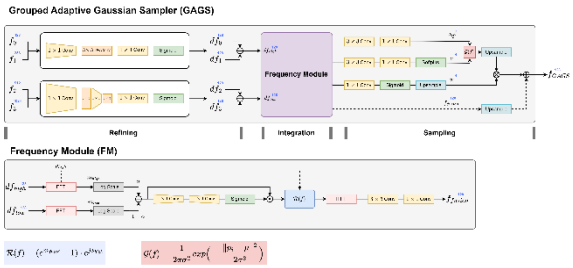


그림 1. (신명조, 9) GAGS 모듈 프레임워크

본 모델은 384 이미지 크기를 사용하는 Swin-Base를 백본으로 사용하며, 추출된 4 단계 특징으로 그림 1과 같이 구성된 GAGS(Grouped Adaptive Gaussian Sampler) 모듈을 통해 얕은 단계 특징 df_{high}과 깊은 단계 특징 df_{low}으로 통합한다. 이후 혼합된 크기 성분과 고수준 위상 정보를 이용하여 다음과 같이 계산한다.

$$\mathcal{R}(f) = (e^{m_{fused}} - 1) \cdot e^{j\phi_{high}} \quad (1)$$

복원된 특징은 3개의 GAT[4] 블록과 업샘플링을 통해 최종 분할 맵을 생성한다.

IV. 결과 및 결론

제안하는 모델은 COD10K[1] 데이터세트에서 표 1

표1. COD10K 평가 결과

모델	$F_W\uparrow$	$S\uparrow$	$E\uparrow$	$MAE\downarrow$
Ours	0.7784	0.8638	0.9163	0.0243

과 같은 결과를 달성하였다. 특히 E-measure에서 0.9163을 달성하여, SINet, FSPNet, ZoomNet과 같은 기존 모델의 성능을 뛰어 넘는 것을 확인할 수 있다. 이는 제안하는 GAGS의 주파수 융합과 GAT 디코더 구조가 위장 객체의 전역 구조와 세부 경계를 효과적으로 탐지할 수 있음을 보인다.

참고문헌

[1] D. Fan and G. Ji and M. Cheng and L. Shao, “Concealed Object Detection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.44, 10, pp. 6024-6042, 2022.

[2] Z. Huang and H. Dai and T. Xiang and S. Wang and H. Chen and J. Qin and H. Xiong, “Feature Shrinkage Pyramid for Camouflaged Object Detection with Transformers”, IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR, 2023.

[3] Y. Pang and X. Zhao and T. Xiang and L. Zhang and H. Lu, “Zoom In and Out: A Mixed-scale Triplet Network for Camouflaged Object Detection”, IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR, 2022.

[4] I. Georgios and C. Aman and E. Aaron, “Gaussian Adaptive Attention is All You Need: Robust Contextual Representations Across Multiple Modalities”, arXiv, 2401.11143, 2024.

본 연구는 교육부와 제주특별자치도의 지원을 받아 제주 RISE 센터를 통 해 Regional Innovation System and Education (RISE) 사업의 지원을 받아 수행된 연구임. (RISE_2-2-11)

NER 기반 용어 제약과 소형 언어 모델 미세 조정을 통한 한→영 도메인 번역

정성우*, 이창준O
*성균관대학교 데이터사이언스융합학과 (발표자)
O성균관대학교 글로벌융합학부 (교신저자)
e-mail:mrensol@lgensol.com*, changjunlee@skku.eduO

Korean to English Domain Translation via NER Based Terminology Constraints and Small Language Model Fine-Tuning

Sungwoo Choung*, Changjun LeeO
*Dept. of Applied Data Science, SungKyunKwan University
OSchool of Convergence, SungKyunKwan University

요 약

기계 번역은 신경망 기반 번역과 대규모 언어 모델의 발전으로 범용 도메인에서 높은 성능을 달성하였으나, 산업·기업 내부 도메인의 전문 용어를 정확하고 일관되게 번역하는 데에는 한계를 보인다. 기존 연구는 용어집 활용에 집중할 뿐 입력 문장에서 용어를 식별하는 선행 단계를 충분히 다루지 못하였으며, 정규식 기반 문자열 매칭은 동음이의어·오탈자·형태 변이에 취약하다. 본 연구는 문맥 인식형 개체명 인식(NER) 모델로 도메인 용어를 식별·교정하고, 군집 식별자 기반 용어집 조회로 용어 제약 프롬프트를 구성한 뒤, 소형 언어 모델(EXAONE 3.5 2.4B)을 미세조정하는 번역 프레임워크를 제안한다. 배터리 제조 도메인 한→영 번역 실험에서 제안 방법은 정규식 기반 방식보다 유의미하게 높은 용어 번역 정확도(97.8%)와 오탈자 강건성을 보였으며, 동시에 BLEU·COMET 기반 일반 번역 품질에서도 가장 안정적인 성능을 유지하였다. 또한 소형 미세조정 모델이 대형 모델의 ICL 기반 번역과 동등한 용어 정확도를 달성하여 온프레미스 환경의 실용적 대안이 될 수 있음을 확인하였다.

키워드: 기계 번역, 도메인 적응, 개체명 인식, 용어 제약, 소형 언어 모델

I. 서론

기계 번역은 신경망 기반 번역(Neural Machine Translation, NMT)과 대규모 언어 모델(Large

※지원기관표시(사사표기)

Language Model, LLM)의 발전으로 범용 도메인에서 큰 성과를 거두었다. 그러나 NMT는 학습 데이터 분포에 강하게 의존하여 산업·기술 도메인과 같이 충분히 관찰되지 않은 영역에서는 번역 품질이 저하되고, 전문 용어를 강제하거나 출력 형식을 제어하는 데 한계를 보인다. LLM 역시 인컨텍스트 학습(In-context Learning, ICL)을 통해 다양한 번역 과업에 적응할 수 있으나, 외부에 공개되지 않은 지엽적 도메인의 전문 용어를 일관되게 처리하지 못한다 [1].

도메인 특화 번역에서 전문 용어의 정확성은 전체 번역 품질을 좌우하는 핵심 요소이다. 이를 위해 최근 연구들은 용어집(glossary)을 프롬프트나 학습 데이터에 통합해 왔으나, 입력 문장에서 도메인 용어를 정확히 식별하는 선행 단계는 충분히 다루어지지 않았다. 실제 번역 파이프라인은 여전히 Trie 구조나 정규표현식 기반 문자열 매칭에 의존하는 경우가 많으며, 이는 동음이의어, 오탈자, 한국어 교착어 특성에 따른 형태 변이에 취약하여 과잉 매칭과 용어 누락을 유발한다. 또한 기존 용어 제약 프레임워크는 대부분 13B 이상의 대형 모델을 전제로 검증되어 [2], 민감 데이터의 외부 전송이 제한되는 기업 온프레미스 환경에 적용하기 어렵다.

본 연구는 이러한 한계를 극복하기 위해 문맥 인식형 개체명 인식(Named Entity Recognition, NER)을 활용한 용어 식별·매핑 기반의 도메인 특화 번역 프레임워크를 제안한다. 이를 검증하기 위해 두 가지 연구 질문을 설정한다. RQ1은 NER 기반 용어 식별이 규칙 기반 문자열 매칭보다 용어 정확도와 오탈자 강건성을 향상시키는지, RQ2는 NER 기반 프레임워크를 적용한 소형 미세조정 모델이 대형 모델의 ICL 기반 번역과 경쟁력 있는 성능을 달성하는지를 다룬다.

II. 관련 연구

2.1 NMT 도메인 적응

NMT 도메인 적응은 학습 데이터 구성을 조정하는 데이터 중심 접근, 모델의 학습 방식·구조를 조정하는 모델 중심 접근, 추론 단계에서 디코딩을 제어하는 디코딩 중심 접근으로 분류된다. 그러나 이러한 접근들은 주로 문장 수준의 전반적 번역 품질을 향상시키거나 사전에 정의된 용어를 출력에 포함시키는 능력에 초점을 두며, 도메인 번역에서 특히 중요한 전문 용어를 문맥적으로 식별하고 일관되게 적용하는 문제는 취약하게 남아 있다. 특히 어휘 제약 디코딩은 지정된 용어를 출력에 포함시킬 수는 있으나, 해당 용어가 현재 문맥에서 적절한 의미로 선택되었는지까지 보장하지는 못한다는 한계가 지적되어 왔다.

2.2 LLM 기반 번역과 용어 제약

LLM은 ICL과 미세조정을 통해 번역 과업을 효과적으로 수행할 수 있으며, 도메인 용어와 예제를 명시적으로 제공하고 이를 미세조정에 통합할 때 번역 성능이 향상된다 [1]. 특히 번역 태스크에서 용어집 기반 학습 데이터 재구성으로 용어 정확도와 번역 품질을 함께 개선한 연구도 보고되었다 [2]. 그러나 이들 연구에서 용어집 활용의 선행 단계인 용어 식별은 여전히 문자열 매칭에 의존하며, 성능 검증이 대부분 대형 모델을 전제로 하여 기업 온프레미스 환경에서 운용 가능한 소형 모델에서의 적용 가능성은 충분히 다루어지지 않았다. 본 연구는 이 두 공백을 직접적으로 다룬다.

III. 제안방법

제안 프레임워크는 그림 1과 같이 세 단계로 구성된다.

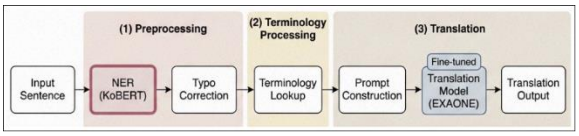


그림 1. 제안하는 번역 파이프라인의 전체 구조

3.1 NER 기반 도메인 용어 식별

약 90개의 배터리 도메인 전문 용어가 포함된 약 1만 건의 문장을 KoBERT 인코더로 임베딩하고, 용어별 대표 임베딩을 PCA로 2차원 축소한 뒤 KMeans로 군집화하여 각 용어에 군집 식별자를 부여한다. 군집 수는 4~9 범위에서 탐색하였으며, 동음이의어 충돌이 적고 군집별 학습 문장 수의 균형이 유지되면서 전체 문장 수가 급감하지 않는 값으로 7을 선택하였다. 이 군집 식별자를 클래스 레이블로 하는 BIO 태깅 체계를 적용하여 KoBERT 기반 NER 모델을 학습한다. 예컨대 '양극의 탈리로 인해 불량이 발생하였음'에서 '양극'은 한 군집(B-1, I-1)으로, 동음이의어인 '탈리'는 의미에 따라 separation 군집과 deintercalation 군집으로 분리 태깅되어 문맥에 따라 구분된다. 추론 시 모델은 입력 문장에서 도메인 용어와 그 군집을 문맥적으로 식별하며, 각 인식 결과는 원문 텍스트, 군집 식별자, 문자 위치로 구성된다.

역추출된 단어는 오탈자나 띄어쓰기 변이를 포함

※지원기관표시(사사표기)

할 수 있으므로, 한글 자모 분해 함수를 통해 단어를 자모 단위로 분해하고 Levenshtein 편집 거리를 정규화한 자모 유사도를 산출한다. 해당 군집 내 표준 용어 후보 중 유사도가 가장 높은 표준 용어로 매핑하며, 유사도가 임계값 미만이면 인식 결과에서 제외한다. 이로써 표면형이 변형된 입력도 표준 용어로 교정된다.

3.2 용어 매핑 및 번역 프롬프트 구성

번역 용어집은 군집 식별자를 상위 키, 한국어 용어를 하위 키, 영어 번역 표현을 값으로 하는 계층적 딕셔너리 구조이다. 군집 식별자를 조회 키로 사용함으로써, 문맥에 따라 의미가 달라지는 동음이의어도 군집 단위로 올바른 번역 표현이 선택된다. 조회된 용어 쌍은 '한국어 용어 : 영어 표현' 형식으로 직렬화되어 dictionary 블록을 구성하며, 최종 프롬프트는 시스템 프롬프트, dictionary 블록, 교정된 입력 문장의 세 부분으로 이루어진다.

3.3 소형 언어 모델 미세조정

추론 단계와 동일한 프롬프트 구조의 한→영 병렬 데이터를 채팅 템플릿 형식으로 변환하고, 이를 사용하여 EXAONE 3.5 2.4B [3] 모델을 전체 미세조정(SFT)한다. 어시스턴트 응답 토큰에 대해서만 손실을 계산하는 Completion-only 방식을 적용함으로써, 모델이 용어 제약을 준수하는 번역 패턴과 도메인 문체를 함께 내재화하도록 유도한다.

IV. 결론

배터리 제조 도메인 한→영 번역을 대상으로, 학습에 사용되지 않은 정상 문장 600건과 오탈자 강건성 평가용 정상·오탈자 문장 각 150건에 대해 Terminology Accuracy, Robustness, BLEU, COMET 네 가지 지표로 평가하였다.

4.1 RQ1: NER 기반 용어 식별의 효과

표 1은 EXAONE 3.5 2.4B 미세조정 모델에 세 가지 용어 식별 방식을 적용한 결과이다. NER + FT 조건은 Terminology Accuracy 97.8%로, Regex All + FT(86.2%) 및 Regex Top1 + FT(74.8%) 대비 각각 11.6%p, 23.0%p 높았으며, 용어 occurrence 단위 McNemar 검정에서도 그 차이가 통계적으로 유의하였다($p < 0.0001$). 오탈자 강건성에서도 NER + FT는 Robustness 0.167로 Regex 조건(0.733, 0.760)보다 안정적이었다.

오탈자 강건성의 차이는 표면형이 변형되면 Regex 방식이 올바른 dictionary 항목을 구성하지 못하는 반면, NER 방식은 자모 유사도 기반 표준 용어 매핑으로 이를 보정하기 때문이다.

표 1. RQ1 주요 결과 (EXAONE 3.5 2.4B FT)

식별 방식	TA(%)	Rob.	BLEU	COMET
NER+FT	97.8	0.167	53.80	.8767
RegexAll+FT	86.2	0.760	52.94	.8698
RegexTop1+FT	74.8	0.733	51.93	.8639

TA: Terminology Accuracy, Rob.: Robustness

일반 번역 품질 측면에서도 NER + FT는 BLEU 53.80, COMET 0.8767로 세 조건 중 가장 높았다. Regex 조건들은 BLEU에서는 51.93~52.94로 NER + FT와 큰 차이를 보이지 않았으나, 의미 기반 지표인 COMET에서는 각각 0.8639, 0.8698로 더 낮았다. 이는 NER 기반 용어 식별이 용어 제약 정확도를 높이면서도 전반적인 번역 품질 저하를 수반하지 않으며, 오히려 의미 기반 품질 지표에서 가장 안정적임을 보여준다.

또한 Ablation 분석 결과, 동일한 NER 식별 방식에서 미세조정(FT) 조건이 ICL 조건보다 Terminology Accuracy가 23.3%p 높았다. 이는 EXAONE 3.5 2.4B와 같은 소형 모델에서는 용어 제약을 프롬프트로만 제공하기보다 미세조정 과정에 통합하는 방식이 더 효과적임을 보여준다.

4.2 RQ2: 모델 규모에 따른 효과

표 2는 NER 기반 용어 제약을 적용한 소형 미세조정 모델과 대형 ICL 모델의 도메인 번역 성능을 비교한 결과이다. 용어 제약이 없는 Raw 조건에서는 모든 모델의 Terminology Accuracy가 40%대에 머물러, 모델 규모 확장만으로는 도메인 용어 번역 문제가 해결되지 않음을 확인하였다. 반면 NER 기반 용어 제약을 적용하면 소형 FT 모델과 대형 ICL 모델 모두에서 정확도가 크게 향상되었으며, EXAONE 3.5 2.4B FT + NER는 OSS 20B 및 OSS 120B와 동일한 97.8%의 Terminology Accuracy를 달성하였다.

일반 번역 품질 측면에서는 OSS 120B + NER가 BLEU 63.16, COMET 0.8906으로 전체 조건 중 가장 높았다. 이는 충분히 큰 모델의 일반 번역 능력이 전반적 품질에 여전히 기여함을 보여준다. 그럼에도 EXAONE 3.5 2.4B FT + NER는 BLEU 53.80, COMET 0.8767을 기록하여 OSS 20B + NER(BLEU 52.63, COMET 0.8763)와 사실상 동

※지원기관표시(사사표기)

등한 품질을 달성하였다. 즉, 소형 모델이라도 도메인 미세조정과 NER 기반 용어 제약을 결합하면 중대형 ICL 모델에 근접한 도메인 번역 품질을 확보할 수 있다.

표 2. RQ2 모델 규모별 결과 (NER 조건)

모델(NER)	TA(%)	BLEU	COMET	방식
EXAONE 2.4B	97.8	53.80	.8767	FT
EXAONE 7.8B	88.6	29.01	.8444	ICL
EXAONE 32B	96.6	34.14	.8571	ICL
OSS 20B	97.8	52.63	.8763	ICL
OSS 120B	97.8	63.16	.8906	ICL

사례 분석 결과, 세 모델이 동일하게 97.8%에 머문 잔여 오류는 번역 모델의 용어 제약 미준수가 아니라 선행 NER 단계의 동음이의어 군집 오분류에서 비롯되었다. 즉, NER는 해당 표현이 도메인 용어라는 사실은 탐지했으나 문맥상 올바른 의미 군집이 아닌 다른 군집으로 분류하였고, 그 결과 잘못된 dictionary 항목이 제공되어 세 모델 모두 동일한 오역을 생성하였다. 이는 NER 기반 프레임워크에서 최종 성능의 상한이 번역 모델의 규모보다 NER 단계의 의미 군집 분류 정확도에 의해 제한될 수 있음을 시사한다.

4.3 결론

본 연구는 NER 기반 용어 식별·매핑과 소형 언어 모델 미세조정을 결합한 도메인 특화 한→영 번역 프레임워크를 제안하였다. 실험 결과 NER 기반 용어 식별은 정규식 기반 방식보다 높은 용어 번역 정확도와 오탈자 강건성을 보였으며, 동시에 BLEU·COMET 기반 일반 번역 품질에서도 저하 없이 가장 안정적인 성능을 유지하였다. 또한 소형 미세조정 모델이 대형 ICL 모델과 동등한 용어 정확도를 달성하였다. 이는 도메인 특화 번역에서 단순한 모델 규모 확장보다 정확한 용어 식별과 표준 용어 매핑이 중요하며, 소형 모델과 NER 기반 전처리의 결합이 보안이 중요한 기업 온프레미스 환경에서 활용 가능한 실용적 방법론임을 보여준다. 향후 연구로는 다양한 산업 도메인·언어 방향으로의 확장, 위치 기반 동음이의어 제약, 오탈자 포함 데이터의 미세조정 효과 분석을 진행할 수 있다.

참고문헌

[1] J. Zheng, et al., "Fine-Tuning Large Language Models for Domain-Specific Machine Translation," arXiv:2402.15061, 2024.

[2] S. Kim, et al., "Efficient Terminology Integration for LLM-Based Translation in Specialized Domains," Proc. of the Ninth Conference on Machine Translation, pp. 1-10, 2024.

[3] S. An, et al., "EXAONE 3.5: Series of Large Language Models for Real-World Use Cases," arXiv:2412.04862, 2024.

[4] SKTBrain, "KoBERT: Korean BERT Pre-Trained Cased," GitHub repository, 2019.

※지원기관표시(사사표기)

Distributed Labor Market Intelligence: NLP and Semantic Job Recommendation at Scale

Neelam Patidar, **Jongwook Woo**
Department of Information Systems
California State University, Los Angeles, USA
{npatida, jwoo5}@calstatela.edu

ABSTRACT

The rapid growth of digital recruitment platforms has generated massive volumes of workforce data requiring scalable analytical frameworks for interpretation. This research presents a distributed analytics framework for analyzing over 1.3 million LinkedIn technology job postings using Apache Spark, natural language processing (NLP), Latent Dirichlet Allocation (LDA), and semantic recommendation modeling. The proposed pipeline integrates distributed preprocessing, semantic role normalization, topic modeling, and TF-IDF-based similarity analysis to extract meaningful labor market insights from large-scale recruitment data.

Experimental results revealed strong hiring concentration within major U.S. technology hubs and significant demand for software engineering, cloud infrastructure, data engineering, and machine learning roles. The LDA-based clustering framework identified recurring skill communities across technologies including Python, SQL, AWS, Docker, and Kubernetes, while the recommendation framework demonstrated context-aware job matching through semantic similarity computation. The proposed approach illustrates how distributed NLP and unsupervised learning techniques can support labor market analytics, intelligent recruitment systems, and career recommendation applications. The recommendation framework achieved a Precision@10 of 0.700 and NDCG@10 of 0.776, demonstrating effective semantic job matching within large-scale recruitment data.

Keywords: Labor Market Intelligence, Big Data, PySpark, NLP, LDA, TF-IDF, Semantic Recommendation Systems, Career Intelligence

I. INTRODUCTION

Digital recruitment platforms such as LinkedIn generate massive volumes of semi-structured workforce data containing valuable information related to hiring demand, technical skill evolution, and workforce specialization.

However, extracting the interpretable labor market intelligence from large-scale recruitment data remains computationally challenging due to inconsistent job titles, noisy skill descriptions, and heterogeneous workforce metadata. Recent advances in distributed computing and natural language processing (NLP) have enabled scalable analysis of unstructured employment datasets. Nevertheless, many existing workforce analytics approaches remain limited by small-scale datasets, centralized architectures, or simplistic keyword-based analytical methods that fail to capture latent semantic workforce structures. To address these limitations, this research presents a scalable workforce intelligence framework using Apache Spark, PySpark MLlib, Latent Dirichlet Allocation (LDA), and semantic recommendation modeling for analyzing over 1.3 million LinkedIn technology job postings collected during 2024.

The primary contributions of this research include:

1. Development of a scalable distributed labor market analytics framework using Apache Spark.
2. Creation of a semantic job title normalization pipeline for noisy recruitment data.
3. Application of Latent Dirichlet Allocation (LDA) for interpretable workforce skill clustering.
4. Design of a semantic recommendation system using TF-IDF and cosine similarity.
5. Large-scale workforce analysis across over 1.3 million LinkedIn technology job postings.

This research demonstrates how semantic analytics can support modern workforce intelligence systems and contribute toward the labor market analytics.

II. RELATED WORK

Recent advancements in natural language processing (NLP), machine learning, and large-scale data

analytics have significantly influenced research in recruitment intelligence and labor market analysis. Online recruitment platforms generate substantial volumes of workforce data that enable the study of skill demand evolution, occupational transitions, hiring patterns, and workforce specialization using computational approaches [5], [6].

Latent Dirichlet Allocation (LDA) remains one of the most widely adopted probabilistic topic modeling techniques for discovering latent semantic structures within unstructured textual corpora [2]. Prior studies have successfully applied topic modeling to recruitment datasets for skill extraction, occupational classification, and workforce trend analysis [7]. Similarly, information retrieval techniques such as TF-IDF and cosine similarity continue to serve as foundational approaches for semantic search and recommendation systems by transforming textual information into vector-space representations suitable for similarity computation [3].

Recent research has further explored semantic recommendation frameworks and representation learning techniques to improve job matching and career recommendation performance [5], [8]. However, many existing studies remain limited by centralized processing architectures, smaller datasets, or keyword-based matching methods that fail to capture complex relationships among skills, occupations, and workforce domains.

This research differentiates itself from prior studies through:

- Integration of distributed big data processing using Apache Spark and Hadoop.
- Large-scale analysis over 1.3 million LinkedIn technology job postings.
- Semantic role normalization across noisy recruitment data.
- Interpretable unsupervised skill clustering using LDA.
- Intelligent semantic recommendation architecture using TF-IDF and cosine similarity.
- Unified integration of NLP, unsupervised learning, and workforce intelligence.

III. DATASET AND TECHNOLOGY

3.1 Dataset Overview

The primary dataset used in this research is the publicly available "1.3M LinkedIn Jobs and Skills (2024)" dataset obtained from Kaggle. The dataset contains over 1.3 million job postings along with structured employment metadata including: *Job titles, Company names, Geographic locations, Work modes, Job levels, Employment types, Skill descriptions, Posting metadata etc.*

Table 1: Technology Stack

The proposed framework integrates multiple big data and machine learning technologies:

Technology	Purpose
Apache Spark & Hadoop	Distributed Workforce Analytics
Databricks	Distributed Experimentation/ EDA
Spark MLlib	NLP & ML workflows
Matplotlib / Seaborn	Visualization
TF-IDF	Semantic Vectorization
Cosine Similarity	Recommendation Ranking
LDA	Skill Clustering/ Topic modeling

Table 2: Hardware Specifications

Cluster Nodes	*5 (2 master nodes, 3 data nodes)*
Cluster version	*Hadoop 3.1.2*
Spark Version	*Spark 3.0.2*
CPU Speed	*1995.312 MHz*
Memory	*806.4GB*
Resource Manager	*Yarn*

IV. METHODOLOGY

The proposed methodology follows a modular pipeline architecture consisting of five major phases:

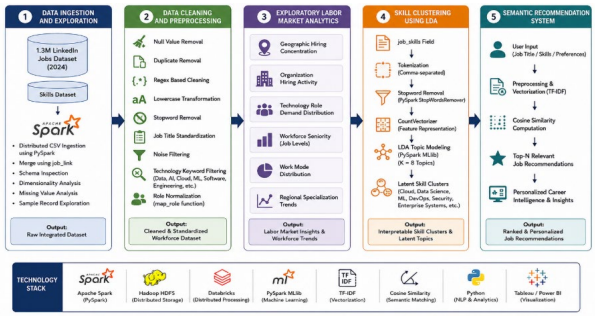


Figure 1. Distributed Workforce Analytics Pipeline

4.1 Data Ingestion and Exploration

The LinkedIn job posting and skill datasets were loaded into Spark DataFrames and merged using the job_link identifier. Initial exploratory analysis was conducted to validate schema consistency, dataset completeness, and data quality prior to preprocessing. This phase ensured that the distributed environment was correctly configured and that the dataset structure was suitable for downstream preprocessing and modeling tasks.

4.2 Tech Role Isolation

To focus the analysis specifically on technology-related positions, a predefined keyword-based filtering approach was developed. Technology-oriented job

postings were isolated using terms such as: *Data, Software, AI, Machine Learning, Cloud, Engineering, Analytics*. A dynamic OR-condition was constructed using PySpark filtering operations and lowercase pattern matching. This filtering strategy removed non-technical positions and narrowed the analysis toward workforce demand within the technology sector.

4.3 Data Engineering

Large-scale recruitment datasets often contain noisy, inconsistent, and redundant textual information. To improve analytical reliability, multiple preprocessing operations were performed.

Records missing critical attributes such as: *Job title, Company name, Job location* were removed. Duplicate postings were identified using unique job identifiers. Job titles exhibited substantial inconsistencies due to: Abbreviations, Typographical errors, Seniority prefixes, Organizational naming conventions. A custom normalization function (***map_role***) was developed to: *Lowercase titles, remove seniority terms, apply regex-based transformations, Map variants into standardized categories*. Examples included:

Table 3: Standardized Category

Raw Title	Standardized Title
Sr. Data Engineer	Data Engineer
Software Developer	Software Engineer
Lead ML Engineer	Machine Learning Engineer

Non-technical and low-confidence postings were removed including: *Legal roles, Clerical roles, Internships, Volunteer positions, rare outlier titles.* This phase produced a refined and semantically consistent technology workforce dataset.

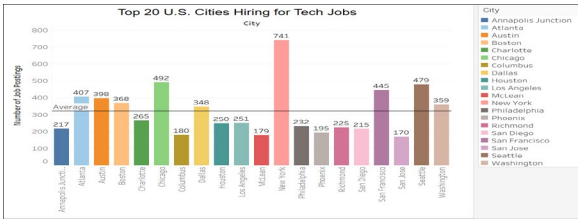


Figure 2. Top U.S. Cities Hiring for Technology Roles

V. EXPLORATORY LABOR MARKET ANALYTICS

5.1 Geographic Distribution & Top U.S. Cities for Technology Hiring

The United States accounted for approximately 87.7% of all technology-related postings, significantly exceeding contributions from the UK and Canada. The concentration of workforce demand across major U.S. metropolitan regions demonstrated the highest technology hiring activity, highlighting the dominance of major innovation and enterprise technology hubs. California-specific analysis revealed strong workforce demand across Top 5 California cities:

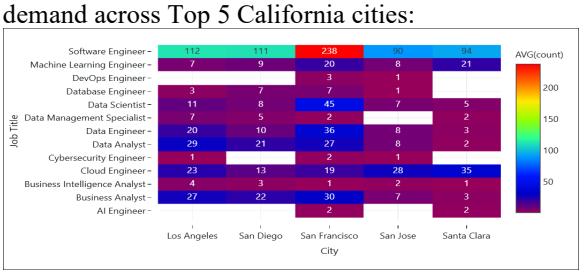


Figure 3. Technology Hiring Trends Across California Cities

The findings reaffirm California's role as a global technology workforce center.

5.2 Most In-Demand Standardized Roles

Software Engineer emerged as the dominant role with approximately 121,184 postings, nearly four times the count of the second-ranked Cloud Engineer (30,772). Data Engineer (3,419), Data Scientist (3,017), and ML Engineer (2,021) followed, together highlighting the dominance of data-oriented and cloud roles in the current technology hiring landscape.

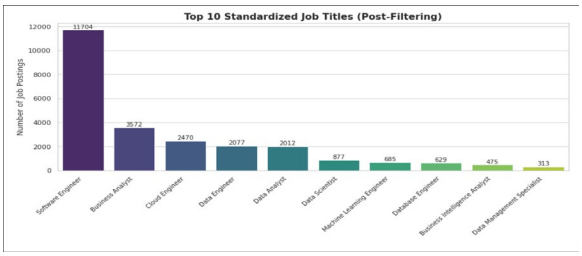


Figure 4. Most In-Demand Standardized Technology Roles

VI. SKILL CLUSTERING USING LATENT DIRICHLET ALLOCATION (LDA)

Technical skill descriptions within recruitment data are often noisy, overlapping, and semantically inconsistent. Individual skills alone may not accurately represent broader occupational specialization. To address this limitation, Latent Dirichlet Allocation (LDA) was applied to uncover latent workforce skill structures across job postings. The job_skills field was tokenized into skill arrays using comma-separated parsing. Stopword removal was applied using PySpark's StopWordsRemover to eliminate non-informative terms. The resulting *filtered_skills* column retained semantically meaningful technical competencies suitable for topic

modeling. The cleaned skill tokens were transformed into numerical vectors using CountVectorizer. Each skill was represented as a frequency-based token within a distributed feature space. This representation enabled downstream unsupervised modeling using LDA.

LDA was initially configured with a topic count of k=8 as a baseline exploration, using a PySpark implementation with token-frequency feature vectors. To identify the statistically optimal configuration, multiple topic counts (k = 4, 6, 8, 10, 12) were subsequently evaluated using Log-Likelihood and Perplexity metrics, as detailed in Section 6.2. The final model was set to k=4, which produced the strongest coherence and generalization performance across the large-scale recruitment corpus.

Table 4: Topic Category & Skills

Topic Category	Representative Skills
Cloud DevOps	AWS, Docker, Kubernetes
Data Science	Python, SQL, Analytics
Machine Learning	AI, ML, Modeling
Software Engineering	Java, Development
Enterprise Systems	SAP, Integration
Cloud Security	Security, Infrastructure

Several skills appeared across multiple clusters, indicating interdisciplinary convergence within modern technical roles.

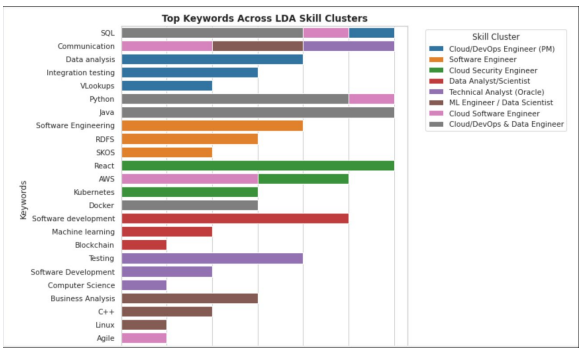


Figure 5. LDA-Based Workforce Skill Clusters

6.1 Workforce Intelligence Insights

The LDA-based clustering revealed:

- Strong dependence on Python and SQL across domains
- Convergence between cloud infrastructure and software engineering
- High overlap between analytics and development roles
- Growing integration of machine learning skills across technical categories

The resulting semantic taxonomy significantly improves interpretability of workforce demand patterns.

6.2 LDA Model Evaluation and Topic Optimization

To evaluate the effectiveness of the proposed topic modeling framework, the LDA model was assessed using probabilistic topic-modeling metrics including ***Log-Likelihood*** and ***Perplexity***. Log-Likelihood measures how well the model explains the observed corpus, while Perplexity evaluates the model's ability to generalize to unseen textual data. The initial LDA model achieved a Log-Likelihood score of **−7,279,488.48** and a Perplexity score of **55,711.83**, indicating successful extraction of meaningful latent workforce skill structures from large-scale technology recruitment data.

To determine the optimal number of latent topics, multiple topic configurations (**k = 4, 6, 8, 10, and 12**) were experimentally evaluated. The results demonstrated that the model with **4 topics** produced the highest Log-Likelihood and lowest Perplexity, indicating the strongest statistical coherence and generalization performance. As the number of topics increased, model perplexity increased substantially while Log-Likelihood deteriorated, suggesting increased topic redundancy and semantic overlap among technical skill categories. These findings indicate that broader workforce skill domains provide more stable representations of large-scale technology job postings than highly granular topic segmentation.

Table 5: LDA Topic Optimization Result

Number of Topics (k)	Log-Likelihood	Perplexity
4	-6,667,157.24	22,219.33
6	-6,940,693.89	33,501.66
8	-7,273,944.58	55,250.09
10	-7,676,769.02	101,148.42
12	-8,139,195.40	202,508.55

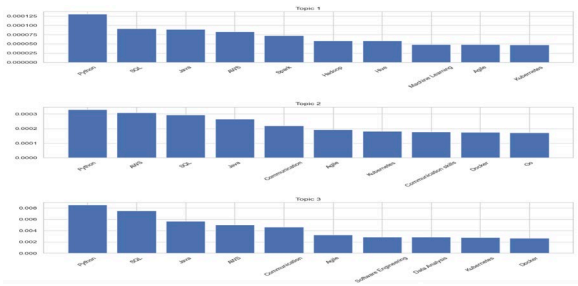


Figure 6: Topic-level dominant keywords & relative importance across skill clusters.

VII. SEMANTIC RECOMMENDATION SYSTEM

Traditional job search systems rely heavily on keyword matching and static filtering. Such approaches frequently fail to capture semantic relationships between: *Skills, Job titles, Functional domains, Technical specialization.* This research introduces a semantic recommendation framework designed to support intelligent career alignment.

Each job posting was transformed into a composite semantic representation containing: *Standardized job title, Job level, Dominant LDA skill cluster, Associated skill terms.* This aggregated representation captured both technical and functional characteristics of each role.

The composite job descriptions were transformed into TF-IDF vectors. TF-IDF weighting reduced the impact of highly common terms while emphasizing discriminative semantic features.
User input queries underwent identical preprocessing and vectorization.

Cosine similarity scoring was then applied between:

- *User semantic vector*
- *Job posting semantic vectors*

Similarity scores represented contextual alignment between user preferences and workforce opportunities.

Table 6: Recommendation Sample Output

job_title	company	job_location	job_level	job_type	similarity
Machine Learning Engineer	Tencent Games	Los Angeles, CA	Mid senior	Onsite	0.223977709229877 67
Senior Machine Learning Engineer (Computer Vision Recognition)	Verticalmove, Inc	Los Angeles, CA	Mid senior	Onsite	0.222945951426861 1
Principal Machine Learning Engineer, ML Training Platform	Snap Inc.	Los Angeles, CA	Mid senior	Onsite	0.179934086622082 2
Junior Data Engineer	GGG Demolition, Inc.	Los Angeles, CA	Mid senior	Onsite	0.177020544969525 1
Staff Machine Learning Engineer, Series A	Recruiting from Scratch	Los Angeles, CA	Mid senior	Onsite	0.162740711496259 46
Machine Learning Engineer, Generative ML , 3+ Years of Experience	Snap Inc.	Los Angeles, CA	Mid senior	Onsite	0.145840438162211 95
TS/SCI Data Scientist	Harnham	Los Angeles, CA	Mid senior	Onsite	0.136134780079056 3
Staff Machine Learning Engineer, Series A	Recruiting from Scratch	Los Angeles, CA	Mid senior	Onsite	0.130021942644882 34
(Global Oil Gas) Senior Data Scientist Expert	MatchaTalent	Los Angeles, CA	Mid senior	Onsite	0.101203590284065 48
Senior Machine Learning Engineer	Freeform	Los Angeles, CA	Mid senior	Onsite	0.099151282
Sr. Software Engineer	Eliassen Group	Los Angeles Metropolita	Mid senior	Onsite	0.097517519
Business Insight Data Analyst III	L.A. Care Health Plan	Los Angeles, CA	Mid senior	Onsite	0.094870781
Senior Data Engineer	Verticalmove, Inc	Los Angeles Metropolita	Mid senior	Onsite	0.094128059
Staff Data Scientist	Intuit	Los Angeles, CA	Mid senior	Onsite	0.092237338
Senior Machine Learning Engineer - Generative AI	Lumicity	Los Angeles Metropolita	Mid senior	Onsite	0.088979363
Staff Software Engineer	Demand.io	Los Angeles Metropolita	Mid senior	Onsite	0.086932415
Senior Data Scientist	Jobot	Los Angeles, CA	Mid senior	Onsite	0.086623338
Business Analyst II	L.A. Care Health Plan	Los Angeles, CA	Mid senior	Onsite	0.085813799
Senior Backend Software Engineer	Kindo	Los Angeles, CA	Mid senior	Onsite	0.081973271
Senior Machine Learning Engineer	StormAI	Los Angeles Metropolita	Mid senior	Onsite	0.078185211

Recommendation Workflow

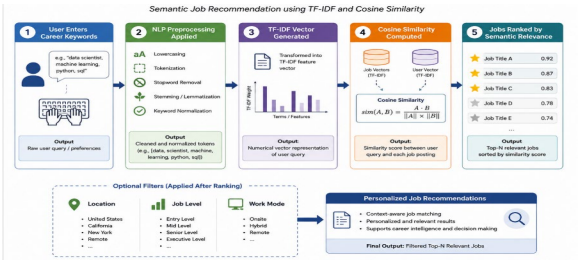


Figure 7. Semantic Recommendation Workflow

The framework demonstrates how semantic AI systems can support: *Intelligent career guidance, Workforce alignment, Personalized job discovery, Context-aware recommendation systems.* The approach moves beyond syntactic keyword matching toward semantic workforce intelligence.

7.1 Recommendation Evaluation

To quantitatively evaluate the recommendation quality, the proposed semantic recommendation framework was assessed using Precision@10 and Normalized Discounted Cumulative Gain (NDCG@10). Precision@10 measures the proportion of relevant job recommendations retrieved within the top 10 results, while NDCG@10 evaluates the ranking effectiveness by assigning greater importance to relevant recommendations appearing higher in the recommendation list.

Since explicit user interaction data and relevance labels were unavailable, relevance was approximated using query-specific technical keywords aligned with the target job profile. For a representative query containing Data Science, Python, and Machine Learning skills, the recommendation framework achieved a Precision@10 score of 0.700 and an NDCG@10 score of 0.776.

Table 7. Recommendation Evaluation Results

Metric	Score
Precision@10	0.700
NDCG@10	0.776

The Precision@10 score indicates that approximately 70% of the top-ranked job recommendations were relevant to the user query. Similarly, the NDCG@10 score of 0.776 demonstrates that highly relevant job opportunities were generally positioned near the top of the recommendation ranking. These results provide quantitative evidence that the proposed TF-IDF and cosine similarity framework can effectively support semantic job recommendation and intelligent career alignment within large-scale recruitment datasets.

VIII. DISCUSSION

The findings of this research reveal that large-scale employment data can function as a dynamic semantic representation of modern workforce evolution.
Rather than serving solely as recruitment records, technology job postings collectively encode:

- Emerging technical priorities

- Organizational transformation patterns
- Skill convergence dynamics
- Workforce specialization trends

The dominance of technical job titles reflects continued enterprise investment in scalable digital transformation ecosystems.

Additionally, the extracted latent skill clusters demonstrate that modern workforce demand increasingly favors interdisciplinary competency profiles rather than isolated technical specialization.

The research further highlights the growing relevance of interpretable AI systems for workforce intelligence applications. The semantic recommendation framework illustrates how NLP-driven architectures can support intelligent career alignment and adaptive labor market interpretation.

From an educational perspective, the findings provide actionable signals for curriculum modernization, technical upskilling strategies, and workforce preparation initiatives.

IX. CONCLUSION

This research presented a scalable data-driven labor market analytics framework for extracting interpretable workforce intelligence from over 1.3 million LinkedIn technology job postings.

By integrating Apache Spark, distributed NLP pipelines, Latent Dirichlet Allocation (LDA), and a semantic recommendation framework, the proposed approach transformed noisy unstructured recruitment data into meaningful labor market insights. The findings revealed strong geographic concentration of technology hiring, dominant demand for software engineering and cloud-oriented roles, and significant interdisciplinary skill convergence across modern workforce categories.

The effectiveness of the proposed framework was supported through both clustering and recommendation evaluation metrics. Log-Likelihood and Perplexity analysis demonstrated the ability of the LDA model to identify coherent workforce skill structures and meaningful topic representations. Furthermore, TF-IDF vectorization combined with cosine similarity successfully captured semantic relationships between user skill profiles and job postings. The recommendation framework achieved a Precision@10 score of 0.700 and an NDCG@10 score of 0.776, indicating that the majority of recommended jobs were relevant to the user query and that highly relevant opportunities were ranked near the top of the recommendation list.

Beyond descriptive workforce reporting, this research demonstrates how distributed NLP, unsupervised learning, and semantic recommendation techniques can support labor market analytics, intelligent recruitment systems, educational alignment, and adaptive career guidance. The proposed methodology further establishes a foundation for future research in workforce intelligence, recommendation systems, and computational labor market analytics.

X. FUTURE WORK

Future research directions include:

- Transformer-based semantic embeddings (BERT, RoBERTa)
- Temporal labor market forecasting
- Graph-based workforce intelligence modeling
- Hybrid recommendation architectures
- Real-time workforce monitoring systems
- AI-assisted career intelligence platforms

Future extensions may further improve semantic granularity, contextual understanding, and workforce forecasting capabilities.

REFERENCES

[1] Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. EMNLP Proceedings.

[2] Zhang, Y., Wang, H., & Li, X. (2024). Large-Scale Job Recommendation Using Semantic Representation Learning. Expert Systems with Applications, 245, 123456.

[3] Kumar, A., Singh, R., & Patel, M. (2024). Workforce Analytics and Skill Intelligence Using Natural Language Processing. IEEE Access, 12, 56789–56803.

[4] Chen, J., Liu, Z., & Zhao, Y. (2025). Topic Modeling and Skill Extraction from Online Recruitment Data Using Transformer-Based Methods. Information Processing & Management, 62(1), 104321.

[5] Park, S., Kim, J., & Lee, H. (2025). Semantic Job Recommendation Systems Using Hybrid Retrieval Models. ACM Transactions on Information Systems, 43(2), 1–25.

[6] Otani, N., Bhutani, N., & Hruschka, E. (2025). Natural Language Processing for Human Resources: A Survey. Proceedings of NAACL 2025 Industry Track.

[7] Gasco, L., Fabregat, H., García-Sardiña, L., Estrella, P., Deniz, D., Rodrigo, A., & Zbib, R. (2025). Overview of the TalentCLEF 2025: Skill and Job Title Intelligence for Human Capital Management. CEUR Workshop Proceedings, TalentCLEF 2025.

[8] Markel, K. S., Tale, M., & Belz, A. (2025). JobPulse: A Big Data Approach to Real-Time Engineering Workforce Analysis and National Industrial Policy. arXiv:2508.11014.

간소화된 설명인자 기반 혼합용매 용해도 예측 기계학습 모델 개발

김정임*, 여승연O, 탁경재**
O동국대학교 화공생물공학과 학부생
*단국대학교 화학공학과 학부생
**단국대학교 화학공학과 교수
e-mail: ysyeon@dgu.ac.krO, jiji999@dankook.ac.kr*, kjtak@dankook.ac.kr**

Development of a Machine Learning Model for Predicting Solubility in Multicomponent Solvents Based on Simplified Descriptors

SeungYeon YeoO, CheongIm Kim*, KyungJae Tak*
ODept. of Chemical and Biological Engineering, Dongguk University
*Dept. of Chemical Engineering, Dankook University

요 약

본 연구는 혼합 용매에서의 용해도 예측 정확도를 최대화하기 위한 용해도 예측 기계학습 모델을 제안한다. XGBoost 회귀 모델과 MixtureSolDB 데이터셋(약 18만 건)을 사용하여 12가지 특성을 기반으로 한 상호작용 조합을 전수 실험하였다. Phase 1에서 정량화된 분자 구조 정보를 기반한 용해도를 예측하였으며, Phase 2에서는 상호작용 설명인자를 반영한 예측 모델로 정확도 향상시켰다. 마지막으로, Phase 3는 R^2 기반으로 설명인자를 스크리닝하여 설명인자 수를 감소시키면서도 높은 예측 정확도를 유지할 수 있는 간소화된 기계학습 모델을 제안한다. 이를 통해 상호작용 특성 유효 조합 선택이 혼합 용매 용해도 예측 정확도 향상에 효과적임을 확인하였다.

키워드: 혼합 용매 용해도, XGBoost, 저차원 설명인자, 대규모 데이터

I. 서론

바이오, 제약, 화학, 소재 산업 등에서는 혼합 용매계에서의 용질 용해도 예측이 매우 중요하다. 신물질의 용해도를 예측함으로써 적절한 합성 및 분리 공정을 설계할 수 있으며, 생산 공정 설계 시 정확한 용해도 계산을 통해 제품의 수율과 품질을 향상시킬 수 있기 때문이다.

혼합 용매의 용해도는 일반적으로 순수 용매에서

측정된 용해도 실험 데이터를 기반으로 혼합 용매의 용해도를 계산하는 CNIBS/R-K, Jouyban-Acree 등의 모델을 통해 예측된다 [1-2]. 그러나 이러한 모델은 실험 용해도 데이터가 존재하지 않는 신물질에 대해서는 적용이 어렵다는 한계가 있다.

이에 기계학습 기반 용해도 예측 연구가 활발히 수행되어 왔으나, 대부분 단일 용매인 물에서의 용해도 예측에 국한되어 왔다. Cho와 Choi [3]는 graph convolutional network를 이용하여 12가지 원자 특성으로부터 도출된 49개의 설명인자를 기반으로 물에서의 용해도를 예측하였다. Ali 등 [4]은 19,219건의 데이터를 활용하여 물에서의 용해도를 예측하였으며, 다양한 기계학습 모델을 비교한 결과 298개의 설명인자를 이용한 XGBoost 모델이 가장 높은 예측 정확도를 나타내었다.

한편, 최근 혼합 용매에서의 용해도 DB 규모가 급격히 증가하고 있다. 2024년에는 28,098건의 데이터를 포함한 BaoDB [5]가 공개되었으며, 2025년에는 34,734건의 MixSolDB [6], 2026년에는 175,626건의 MixtureSolDB [7]가 연이어 공개되었다. 이러한 DB는 다양한 혼합용매-용질계의 용해도 거동을 체계적으로 분석할 수 있는 기반을 제공하며, 대규모 데이터 축적에 따라 기계학습 기법의 적용 가능성을 높여주고 있다.

본 연구에서는 대규모 용해도 데이터셋과 분자 구조 정보를 활용하여 혼합 용매계에서의 용해도를 예측하는 기계학습 모델을 제안하고자 한다. 먼저, SMILES(Simplified Molecular Input Line Entry System) 분자 표현법을 활용하여 설명인자로 사용할 분자 구조 정보를 정량화한다. 이후 기계학습 알고리즘을 이용하여 정량화된 분자 구조 정보를 기반으로 용해도를 예측한다(Phase 1). 이어서 용매와 용질 간 상호작용과 관련된 설명인자를 추가하여 예측 정확도를 향상시킨다(Phase 2). 마지막으로, 결정계수(R^2) 기반 스크리닝을 수행하여 설명인자 수를 감소시키면서도 높은 예측 정확도를 유지할 수 있는 간소화된 기계학습 모델을 제안한다(Phase 3).

II. 제안 방법

2.1 데이터셋 및 전처리

본 연구에서는 MixtureSolDB 데이터셋(175,626건) 중 SMILES 표기법이 제공되지 않은 데이터 1,450건을 제외한 데이터를 기계학습에 활용하였다. 용매1, 용매2 및 용질에 대하여 각각 RDKit으로부터 12가지 분자 특성(소수성, 분자량, 최대 부분 전하, 최소 부분 전하, 고리수, 방향족 고리수, 극성 표면적, 수소 공여체 수, 수소 수용체 수, sp3 혼성 오비탈을 가진 탄소 비율, 중원자수, 회전 가능 결합 수)을 추출하였으며, 여기에 온도와 혼합용매 조성 등 6개의 공정 변수를 추가하여 총 42개의 설명인자를 구성하였다. 이를 기반으로 혼합용매계에서의 용해도(logS)를 예측하였다. 전체 설명인자 수는 표 1에 정리하였다.

전처리가 완료된 174,176건의 데이터셋은 60:20:20의 비율로 학습(training), 검증(validation), 테스트(test) 데이터셋으로 분할하여 기계학습 모델 개발에 활용하였다. 모델의 예측 성능은 R^2와 평균제곱근오차(RMSE, Root Mean Square Error)를 기반으로 평가하였다.

표 1. 용해도 예측에 사용된 설명인자 수

	Ali et al. [4]	Phase 1	Phase 2	Phase 3
Descriptor	125	36	36	36
FP	128	0	0	0
FGs	7	0	0	0
Feature	38	6	6	6
Interaction	0	0	144	120
Total	**298**	**42**	**186**	**162**

2.2 XGBoost 모델 학습

모든 모델에는 n_estimators=500, max_depth=6,

learning_rate=0.05 등의 동일한 하이퍼파라미터를 적용하였으며, StandardScaler를 이용하여 특성을 정규화하였다. 또한 5-fold 교차 검증을 수행하여 CV R^2 및 RMSE를 평가하였으며, 최종 모델은 학습 및 검증 데이터 전체를 활용하여 구축하였다.

2.3 상호작용 특성 생성

용해도는 용매-용질, 용매-용매 및 용질-용질 간의 분자 간 상호작용에 의해 영향을 받는다. 따라서 예측 정확도를 향상시키기 위하여 용매1-용질, 용매2-용질 및 용매1-용매2 간의 상호작용 특성을 추가적인 설명인자로 도입하였다.

SMILES 기반으로 추출한 12가지 분자 특성 각각에 대하여, a) 용매1과 용질 간 특성값의 차이, 곱, 비, b) 용매2와 용질 간 특성값의 차이, 곱, 비, c) 용매1과 용매2 간 특성값의 차이, 곱, 비, d) 용매1과 용매2의 몰분율 가중 특성값, e) 몰분율 가중 특성값과 용질 특성값 간의 차이 및 곱을 계산하였다. 이를 통해 각 분자 특성마다 총 12개의 상호작용 설명인자를 추가하였으며, 최종적으로 총 186개(42 + 12 × 12)의 설명인자를 구성하였다.

2.4 R^2 기반 설명인자 간소화

설명인자의 수를 감소시키기 위하여 R^2 기반 스크리닝을 수행하였다. 각 설명인자의 유지(KEEP) 또는 제거(DROP) 여부는 용해도 예측에 대한 기여도를 기준으로 결정하였으며, 이때 임계값 θ는 0.005로 설정하였다.

Ⅲ. 실험 결과 및 분석

3.1 Phase 1: Base Case 모델

Phase 1은 상호작용 특성인자를 반영하지 않은 용해도 예측 XGBoost 모델이며, 그 결과를 표 2에 나타내었다. 혼합용매계에서의 용해도 예측임에도 불구하고, 본 모델은 R^2 0.9623 및 RMSE 0.3113을 나타내어 Ali et al. [4]이 개발한 순수용매계 용해도 예측 XGBoost 모델의 R^2 0.92 및 RMSE 0.55보다 우수한 예측 성능을 보였다. 이러한 성능 향상은 선행연구에서 사용된 19,219건 규모의 데이터셋과 비교하여, 본 연구에서는 약 9배 많은 174,176건의 데이터를 활용하였기 때문으로 판단된다. 즉, 대규모 데이터셋을 기반으로 모델이 분자 구조에 따른 용해도 거동을 보다 체계적으로 학습할 수 있었던 것으로 해석된다.

표 2. 모델별 용해도 예측 성능 평가

모델	R^2	RMSE	특성 수	ΔR^2
Phase 1	0.9623	0.3113	42	+0.0000
Phase 2	0.9762	0.2472	186	+0.0141
Phase 3	0.9766	0.2454	162	+0.0145

3.2 Phase 2: 상호작용 특성 반영 모델

Phase 2에서는 상호작용 특성을 포함한 186개의 설명인자를 이용하여 용해도 예측 XGBoost 모델을 개발하였다. 그 결과, R^2는 0.9762, RMSE는 0.2472로 나타나 상호작용 특성을 포함하지 않은 Phase 1 모델 대비 예측 정확도가 향상됨을 확인하였다. 이는 혼합용매계에서의 용해도가 용매1-용매2-용질 간의 상호작용에 크게 영향을 받음을 의미한다. 또한, XGBoost 모델이 분자 구조 정보만으로는 이러한 상호작용 효과를 충분히 학습하지 못하며, 상호작용 특성을 설명인자 형태로 명시적으로 제공할 필요가 있음을 시사한다.

3.3 Phase 3: 설명인자 간소화 모델

Phase 3 모델은 R^2 기반 스크리닝을 통해 설명인자 수를 162개로 감소시킨 XGBoost 모델이다. 설명인자 수를 12.90% 감소시켰음에도 불구하고, 모델의 예측 성능은 R^2 0.9766 및 RMSE 0.2454로 나타나 스크리닝 이전의 Phase 2 모델과 비교하여 차이가 거의 없으나 R^2 0.0004 상승하고 RMSE 0.0018 감소를 나타냈다. 또한, Ali et al. [4]이 순수용매계 용해도 예측에 298개의 설명인자를 사용한 것과 비교하였을 때, 본 연구에서는 훨

씬 적은 수의 설명인자를 이용하여 보다 복잡한 혼합용매계에서의 용해도를 더 높은 정확도로 예측할 수 있음을 확인하였다. 이는 제안된 설명인자 구성 및 스크리닝 방법이 혼합용매계 용해도 예측에 효과적으로 적용될 수 있음을 시사한다.

최종 Phase 3 모델의 parity plot은 그림 1에 나타내었으며, Phase 1-3 모델의 성능 비교 결과는 표 2에 정리하였다.

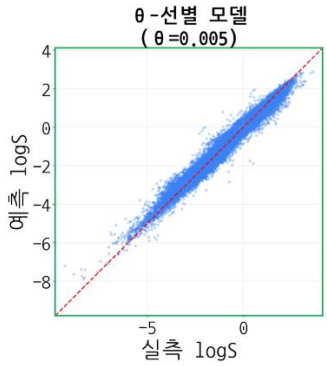


그림 1. Phase 3모델의 Paraty Plot

V. 결론

본 연구에서는 MixtureSolDB 기반의 대규모 혼합용매 용해도 데이터를 활용하여 혼합용매계에서의 용해도를 예측하는 XGBoost 기반 기계학습 모델을 개발하였다. SMILES 기반 분자 구조 정보와 공정 변수를 설명인자로 사용하였으며, 용매-용질 및 용매-용매 간 상호작용 특성을 추가적으로 반영하여 예측 정확도를 향상시키고자 하였다. Phase 1 모델은 R^2 0.9623 및 RMSE 0.3113의 우수한 예측 성능을 나타내었으며, 상호작용 특성을 추가한 Phase 2 모델에서는 R^2 0.9762 및 RMSE 0.2472로 성능이 더욱 향상되었다. 또한, R^2 기반 스크리닝을 적용한 Phase 3 모델은 설명인자 수를 감소시켰음에도 불구하고 기존과 유사하지만 미세한 정확도 향상을 보여 스크리닝으로 감소된 설명인자(고리수, 방향족 고리수)가 정확도 향상을 저해했다는 것으로, 유의미한 스크리닝이었음을 나타낸다. 이를 통해 혼합용매계 용해도 예측에서 분자 간 상호작용 특성이 중요함을 확인하였으며, 제안된 기계학습 기반 접근법이 복잡한 혼합용매계 용해도 예측에 효과적으로 활용될 수 있음을 확인하였다.

감사의 글

인공지능 공부를 지도해주신 동국대학교 이해인 교수님께 감사드립니다.

참고문헌

[1] W. E. Acree Jr., "Mathematical representation of thermodynamic properties", *Thermochim. Acta*, vol. 198, pp. 71-79, 1992.

[2] A. Jouyban and W. E. Acree Jr., "Mathematical derivation of the Jouyban-Acree model", *J. Mol. Liq.*, vol. 256, pp. 541-547, 2018.

[3] H. Cho and I. S. Choi, "Deep learning algorithm of graph convolutional network", *Bulletin of the Korean Chemical Society.*, vol. 40, pp. 485-486, 2019.

[4] M. Ali et al., "Advancing aqueous solubility prediction", *J. Chem. Inf. Model.*, vol. 65. Pp. 8426-8434, 2024.

[5] Z. Bao et al., "Towards the prediction of drug solubility in binary solvent mixtures at various temperatures using machine learning", *J. Cheminform.*, vol. 16, pp. 117, 2024.

[6] H. Jung et al., "Enhancing predictive models for solubility in multicomponent solvent systems using semi-supervised graph neural networks", *Digit. Discov.*, vol. 4, pp. 1492-1504, 2025.

[7] D. Malikov et al., "Dataset of solubility values for organic compounds in binary mixtures of solvents at various temperatures", *Sci. Data*, vol. 13, pp. 727, 2026.