멀티모달과 스마트계약 기반 고위험 AI 실시간 안전성 평가 시스템 설계

김성진[1], 김선집[2*]
[1]한양대학교ERICA 융합보안학과 교수, [2]한세대학교 융합보안학과 교수

Design of a Multimodal and Smart Contract-Based Real-Time Safety Evaluation System for High-Risk AI

Sung-Jin Kim[1], Sun-Jib Kim[2*]
[1]Professor, Department of Convergence Security, Hanyang University ERICA
[2]Professor, Department of Convergence Security, Hansei University

요 약 최근 의료, 금융 등 고위험 분야의 AI 확산은 데이터 편향, 보안 취약점, 윤리적 문제로 인해 강력한 안전성 평가의 필요성을 요구하고 있다. 한국지능정보사회진흥원(NIA)은 7단계 프로세스 기반 안전성 인증을 수행하고 있으나 정적 평가 방식, 수동 검토로 인한 1~2주 처리 지연, 감사 로그의 투명성 부족이라는 구조적 한계를 지니고 있다. 이에, 본 연구에서 제안하는 방식은 NIA 7단계를 확장한 9단계 멀티모달 실시간 안전성 평가 시스템으로 AWS Lambda와 Comprehend, Rekognition을 활용해 5분 주기 편향탐지기반의 텍스트와 이미지를 동시 분석하는 멀티모달 AI 안전성평가로 정보편향 탐지의 정확도를 기존 연구 대비 8.89% 개선한 98%을 달성하였다. 또한, Hyperledger Fabric에 스마트 계약으로 전 과정 로그 자동화를 활용함으로써 기존 방식 대비 6배 빠른 응답성과 23% 정확도를 향상하였다. 본 연구가 향후 데이터 프라이버시 침해 예방과 고위험 AI 안전성 평가에 활용되기를 기대한다.

주제어 : 고위험 AI, AI 안전성 평가, 멀티모달 탐지, 스마트 계약, 실시간 편향 탐지, 블록체인

Abstract The recent proliferation of Artificial Intelligence(AI) in high-risk sectors, such as healthcare and finance, necessitates rigorous safety assessments to address data bias, security vulnerabilities, and ethical concerns. Although the National Information Society Agency(NIA) currently conducts safety certifications based on a 7-step process, the framework faces structural limitations, including static assessment methods, processing delays of one to two weeks caused by manual reviews, and a lack of transparency in audit logs. To address these issues, this study proposes a 9-step multimodal real-time safety assessment system that extends the existing NIA framework. Utilizing AWS Lambda, Amazon Comprehend, and Amazon Rekognition, the proposed system performs simultaneous analysis of text and images with a 5-minute bias detection cycle. This multimodal AI safety assessment approach achieved a 98% accuracy rate in information bias detection, representing an 8.89% improvement over existing research. Furthermore, by implementing automated logging for the entire process via Smart Contracts on Hyperledger Fabric, the system demonstrates six times faster response speeds and a 23% improvement in accuracy compared to conventional methods. This study is expected to contribute significantly to preventing data privacy breaches and enhancing safety assessments for high-risk AI applications.

Key Words : High-Risk AI, AI Safety Assessment, Multimodal Detection, Smart Contract, Real-time Bias Detection, Blockchain

*교신저자 : 김선집(kimsj@hansei.ac.kr)
접수일 2025년 11월 17일 수정일 2025년 12월 15일 심사완료일 2025년 12월 22일

www.kci.go.kr

1. 서론

최근 생성형 AI와 초거대 모델의 급속한 확산은 의료 진단, 금융 신용평가, 자율주행 등 고위험 분야에서 혁신을 가속화하고 있으나, 데이터 편향, 허위 정보 생성, 프라이버시 침해 등의 문제로 AI 안전성 확보가 글로벌 과제로 부상하고 있다[1]. EU AI Act의 고위험 AI 분류 및 사전 평가 의무화, 미국 NIST AI RMF의 위험 관리 프레임워크 등 국제적 규제가 강화되는 가운데, 국내에서도 2025년 「인공지능 기본법」 시행으로 고위험 AI 영향평가가 의무화되었다[2]. 한국지능정보사회진흥원(NIA)은 모델 제출부터 사후 모니터링까지 7단계 안전성 평가 프로세스를 제시하나, 정적 평가 방식과 수동 검토로 인한 처리 지연, 감사 로그의 투명성 부족이 실시간 대응과 신뢰성 확보의 핵심 한계로 지적된다[2].

본 연구는 NIA 7단계 프로세스를 확장·개선한 9단계 멀티모달 실시간 안전성 평가 시스템을 설계하여, 클라우드 컴퓨팅과 블록체인 스마트 계약을 활용해 고위험 AI의 안전성과 윤리적 활용성을 획기적으로 강화하고자 한다.

2. 관련연구

2.1 국외 AI 안전성평가 연구동향

국외 AI 안전성 평가 연구는 EU, 미국, 일본, 중국의 규제 체계와 기업 적용 사례를 중심으로 고위험 AI의 위험 분류와 실시간 모니터링을 강조하고 있다[1]. 이러한 동향은 NIA 7단계 프로세스와 유사하나, 기업 중심 자율 준수와 글로벌 협력이 두드러지며 본 연구의 9단계 시스템 개발에 시사점을 제공한다.

2.1.1 유럽연합(EU)

EU는 2024년 시행된 「AI Act」를 통해 고위험 AI(비생체 인식 시스템 8개 카테고리: 교육, 고용, 법 집행 등)에 품질경영시스템(QMS) 구축, 기술문서 제출, 사후감시를 의무화하며 CE 마킹 인증을 요구한다.

이는 AI 시스템의 위험 분류와 지속적 모니터링을 강화하는 강제 규범으로 설계되었다.

Siemens Healthineers는 「AI Act」 준수를 위해 의료 AI 도구에 QMS를 적용하여 데이터 다양성 검사와 인간 감독(human oversight), 즉 AI 결정 과정에 대해 사람이 수동 검토 및 개입을 통해 편향을 상당히 완화하고 CE 마킹 인증을 획득한 대표 사례이다. 그러나, 실시간 모니터링 부족으로 실시간 데이터 변화에 따른 편향 발생 문제에 대응이 제한적이다[3].

2.1.2 미국

미국 NIST의 AI Risk Management Framework (2023)은 기획·개발·배포·운영에 이르는 전 과정에서 위험을 식별·평가·완화하기 위한 절차를 제시하며, 조직이 자율적으로 적용할 수 있는 비규제형 관리체계를 제공한다[2]. ISO의 ISO/IEC 42001:2023이 인공지능 관리시스템(AIMS) 구축을 위한 국제 표준으로서 조직의 구조적 관리 요구사항을 규정하는 것과 달리, NIST AI RMF는 위험 기반 접근에 중점을 두고 실천적 프로파일과 구현 예시를 통해 운영 수준의 지침을 제시한다[3]. 이러한 특성으로 인해 기업들은 내부 AI 서비스 운영에서 편향 완화, 모니터링, 인간 감독(Human oversight) 등 절차적 통제를 정비하는 데 RMF를 참고하고 있으나, 권고 중심의 구조인 만큼 감사·검증 단계에서 불변성 기반의 기술적 투명성을 보장하는 데에는 제도적 한계가 존재한다.

2.1.3 일본

일본 METI/MIC의 「AI 사업자 가이드라인」(2025 Ver.1.1)은 투명성·책임성을 자율 규제하며 공공조달 지침으로 민간 확산을 촉진한다[4]. SoftBank가 생성형 AI 서비스에 가이드라인을 적용해 리스크 평가로 공공 프로젝트 채택과 80% 편향 감소를 달성한 사례는 혁신과 안전의 균형을 보여준다. 그러나 비강제성으로 기업 준수 수준이 불균일하다는 한계가 있다[4].

2.1.4 중국

중국 국가인터넷정보판공실(CAC)은 생성형 AI 서비스 관리 잠정조치(2023)를 통해 생성형 AI 서비스 제공자에게 알고리즘 등록과 보안평가를 사전 의무화하는 규제 체계를 도입하였다[5]. CAC는 중국 내 사이버 안보, 데이터 관리, 알고리즘 규제 등을 총괄하는 국가 기관으로, 잠정조치는 생성형 AI 모델의 데이터 품질, 알고리즘 안정성, 안전 위험 예방 등을 국가 차원에서 통제하기 위한 기반을 마련하고 있다. 중국 주요 AI 사업자들은 해당 절차에 따라 모델 개발 과정에서 데이터 라벨링, 안전성 검토, 알고리즘 보고 등 준수 체계를 강화하고 있으며, 이를 통해 서비스 출시와 운영에 필요한 규제 요건을 충

www.kci.go.kr

족하는 사례가 축적되고 있다.

다만, 이러한 사전 규제 중심 구조는 사업자의 비용·시간 부담을 증가시켜 기술 개발 속도와 혁신 유인을 약화할 수 있다는 논쟁이 제기되고 있다.

해외 주요국의 규제·평가 체계는 EU의 강행규범, 미국·일본의 자율 규범, 중국의 사전 규제처럼 형태는 다르지만, 모두 고위험 AI의 위험 분류·문서 기반 책임성·인간 감독 강화라는 공통된 방향성을 보인다[1-5]. 그럼에도 각국은 공통적으로 정적·사전 중심 규제에 머물러 실시간 데이터 변화와 모델 드리프트를 감지하는 동적 모니터링 역량이 부족하며, QMS·리스크 문서화·알고리즘 보고 등 수동 문서 기반 통제가 중심이라 검증 가능성과 추적성이 충분히 확보되지 않는다. 또한, 규범 강도의 차이로 인해 EU·중국과 같이 강제 규제는 비용·시간 부담 증가로 혁신 저해 우려가, 미국·일본과 같은 권고형 규범은 기업 간 준수 편차로 안전성 수준의 불균형이 발생하는 한계가 나타난다. 요컨대 해외 트렌드는 공통적으로 실시간 모니터링 부재, 문서 중심 평가의 한계, 규범 강도 차이에서 비롯되는 불균일성을 드러내며, 이는 국내 평가 체계 고도화에서 자동화·불변성·실시간성 기반의 새로운 접근 필요성을 시사한다.

2.2 국내 AI 안전성 평가 연구 동향

국내 AI 안전성 평가 연구 동향은 공공기관의 기준·프로세스 정립과 학계·산업계의 기술·인프라 연구가 병행되는 방향으로 전개되고 있다.

우선 NIA의 AI 데이터 품질관리 가이드라인 v3.5는 학습데이터의 구축·검수·활용 전 주기에 걸쳐 품질 기준과 점검 절차를 제시하며, 생성형 AI 데이터 특성을 반영한 별도 품질 기준과 1-cycle 자가 점검 체계를 도입함으로써 국내 AI 안전성 평가의 데이터 기반을 정교화하고 있다[6]. 고위험 AI 인증과 관련해서는, NIA AI 안전성 연구 보고서에서 NIA 7단계 고위험 AI 인증 프로세스의 현황과 문제점(수동 문서 중심 심사, 처리 지연, 실시간 모니터링 미흡 등)을 분석하며, 평가 자동화와 로그 투명성 강화를 위한 제도·기술적 개선 필요성을 제기하고 있다[7].

기술적 측면에서는 서버리스 환경을 비교한 국내 연구가 AWS·GCP·Azure의 성능·확장성을 분석하여 이벤트 기반 실시간 모니터링·평가 시스템을 구축하기 위한 인프라 선택 기준을 제공하고 있으며[8], BioBERT와 같은 도메인 특화 언어모델 연구는 텍스트 기반 고성능 딥러닝 모델 개발에서 경쟁력을 확보하고 있음을 보여준다[9]. 다만, 기존 연구는 텍스트 정보 처리에 집중되어 있어, 복합적 데이터 처리가 요구되는 고위험 AI 안전성 평가 관점에서의 텍스트·이미지·영상 통합 분석 체계와는 거리가 있다.

종합하면, 국내 동향은 ① NIA 중심의 데이터 품질·고위험 AI 인증 기준 정립, ② 서버리스·클라우드 인프라를 활용한 실시간 처리 기반 마련, ③ 도메인 특화 언어모델 개발을 통한 텍스트 안전성 연구 강화라는 세 축으로 진전되고 있다[6-9]. 그러나 기존 기준과 연구는 여전히 수동·문서 중심 7단계 프로세스와 단일 모달 위주의 모델·평가에 머물러 있어, 멀티모달 데이터와 블록체인·스마트 계약을 결합한 전 주기·실시간·불변 인증 프레임워크로 확장한 9단계 시스템 개발 필요성이 뚜렷하게 대두되는 상황이다.

3. AI 안전성 인증모델 평가 및 개선모델

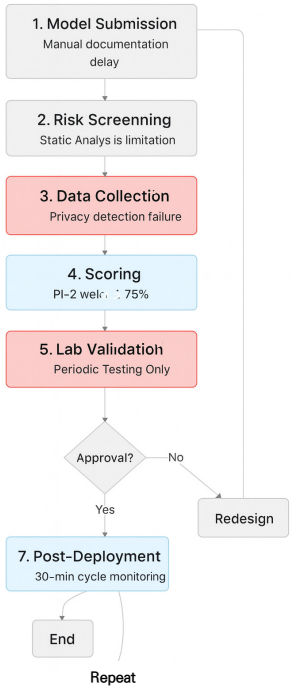
3.1 기존 AI 안전성 평가 모델(NIA 모델 기반)

국내에서 고위험 AI 시스템의 안전성 확보를 위해 제안된 평가·인증 프로세스로서, [Fig. 1]과 같이 NIA 모델은 모델 제출 → 리스크 스크리닝 → 데이터 수집 → 점수화 → 실험실 검증 → 승인/차단 → 사후 모니터링의 7단계 흐름을 제시한다[2]. 이 절차는 고위험성(AI가 생명·신체·기본권 등에 중대한 영향을 미칠 가능성 있는 경우)에 해당하는 AI 시스템을 대상으로 하여, 개발 전부터 배포 후까지 전 주기를 포괄하는 안전성 관리 체계로 설계되었다[2].

〈Table 1〉과 같은 7단계 평가모델의 단계별 문제점은 다음과 같다.

1. 모델 제출: Article 33에 따른 고위험 AI 확인과 연계되나 수동 문서화로 오류·지연 발생한다[1].
2. 리스크 스크리닝: Who-Why-When 접근법을 반영하나, 국내 고위험 AI 인증 프로세스가 정적 분석 중심이어서 실시간 위협 탐지가 어렵다[7].
3. 데이터 수집: 국외 사례에서 드러난 것처럼 개인정보·민감정보 위험을 사전 단계에서 탐지하기 어려운 구조가 남아 있다[4].
4. 점수화: 정확도 75% 수준에서 수동 검토가 요구되며 1~2주 지연이 발생하고, 단순 확률×심각성 함수 기반 평가는 복합 위험 구조를 반영하지 못한다.

www.kci.go.kr



[Fig. 1] NIA 7-Step Process and Key Limitations

〈Table 1〉 NIA 7-Step AI Safety Certification Process

Step	Name	Main Content	Output
1	Model Submission	AI model design documents and initial risk report	Model documents, Article 33 confirmation[3]
2	Risk Screening	Initial scan of high-risk elements (bias, security vulnerabilities)	High-risk element identification report
3	Data Collection	Training data source, diversity, privacy verification	Data quality/privacy report
4	Scoring	Bias/risk quantification (75% accuracy, probability×severity function)	Bias score report[4]
5	Lab Validation	Red team attacks, XAI-based vulnerability analysis	Validation/reliability report
6	Approval/Rejection	Final decision based on NIA/AISI criteria (Article 34 compliance)	Certificate or rejection notice[3]
7	Post-Deployment	30-min cycle monitoring, quarterly reporting	Monitoring logs, reports[7]

5. 실험실 검증: 허위 판례·오탐 사례 보완을 시도하나, 주기적 오프라인 테스트 기반이라 동적 위험 변화 대응에는 한계가 있다[7].
6. 승인/차단: Article 34의 안전조치 의무를 기반으로 하나 감사 로그가 투명하게 관리되지 않아 사후 책임성을 확보하기 어렵다[1].
7. 사후 활용: 30분 주기 모니터링 체계는 실시간 편향·성능 변동을 추적하기 어렵고, 국내 보고서도 이러한 한계를 지적한다[7].

그러나 NIA의 평가모델은 다음과 같은 문제점이 있다.

첫째, 실시간성 측면에서는 30분 단위 수동 모니터링을 5분 주기 자동화로 모니터링 간격을 단축해 성능을 향상 시켰지만, Lambda 기반 구조가 텍스트 중심 분석에 한정되어 있어 이미지·음성 등 복합 데이터의 실시간 처리는 지원하지 못한다.

둘째, CNN-LSTM 모델을 활용 정확도를 90%까지 개선하였으나, 학습이 텍스트 데이터에 치우쳐 있어 텍스트·이미지 간 상호작용에서 나타나는 편향을 탐지하는 데 구조적 제약이 따른다.

셋째, 무결성 측면에서는 Hyperledger Fabric을 활용해 감사 로그의 위변조 위험을 제거했으나, 스마트 계약 기능이 미통합되어 승인 및 차단 절차의 자동 검증이 불가능하다는 한계가 있다.

이에 본 연구에서는 이러한 기존 모델의 구조적 한계(실시간성 부족, 단일 모달의 제약, 검증 자동화 부재)를 극복하기 위해, 확장된 9단계 멀티모달 실시간 안전성 평가 시스템을 제안한다.

3.2 제안 AI 평가모델(멀티모달 기반)

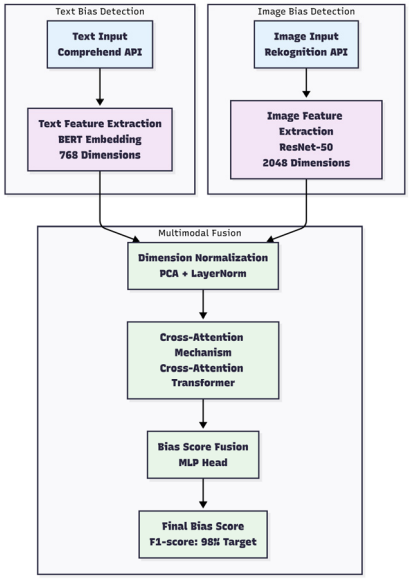
본 연구에서 제안하는 9단계 멀티모달 실시간 안전성 평가 시스템은 NIA의 7단계 평가 체계가 지닌 정적·수동 기반 구조 및 국내 보고서에서 지적된 모니터링 지연, 단일 모달 처리 한계, 감사 로그 불투명성 등의 문제를 보완하기 위해 설계되었다[6, 7]. 제안된 멀티모달 실시간 편향 평가모델은 NIA의 7단계 평가 체계가 가진 정적·단일 모달 중심 구조를 보완하고[7], 데이터 품질·전처리 기준을 규정한 NIA 지침을 충족하도록 설계되었다[6]. 모델은 [Fig. 2]과 같이 텍스트 편향 탐지 모듈, 이미지 편향 탐지 모듈, 그리고 중앙 멀티모달 융합 모듈로 구성된다.

텍스트 입력은 Amazon Comprehend API를 통해 전 처리된 후 BERT 기반 임베딩으로 변환되어 768차원의 특징 벡터를 생성한다[10]. 이는 문장의 의미 정보를 고밀도로 표현하여 고위험 AI의 실시간 편향 탐지에 필

www.kci.go.kr

요한 텍스트 기반 특징을 제공한다.

이미지 입력은 Amazon Rekognition API와 최신 CNN 아키텍처인 EfficientNet-B4[11]를 활용하여 2048 차원의 시각적 특징 벡터로 추출된다. 이 과정은 이미지의 객체·맥락 정보를 풍부하게 반영하여 텍스트와의 결합이 가능한 시각적 의미 공간을 제공한다.



[Fig. 2] Proposed Multimodal Fusion Model Architecture

두 모달에서 추출된 고차원 특징 벡터는 중앙 융합 모듈로 전달되며, PCA와 Layer Normalization을 활용해 차원과 분포를 정규화한다. 이러한 전처리 과정은 상이한 형태의 모달 데이터를 결합하기 위한 안정성을 확보하고, 후속 융합 단계에서의 계산 효율성을 높인다.

정규화된 특징은 Cross-Attention Transformer 구조를 통해 모달 간 의미적 상호작용을 학습한다[12]. 여기서 텍스트를 Query로, 이미지를 Key·Value로 활용하는 상호 어텐션 메커니즘(Co-Attention Mechanism)과, 그 반대로 이미지를 Query로 설정한 병렬 구조가 함께 적용된다. 이러한 구조적 융합은 텍스트·이미지 간 상호 의존적 의미 관계를 포착해 단일 모달 기반 분석의 한계를 보완한다[7].

최종 단계에서는 멀티모달 융합 특징이 다층 퍼셉트론(MLP Head)에 입력되어 편향 스코어로 변환된다. 전체 파이프라인은 이벤트 기반 서버리스 구조와 연동되어 실시간 재평가가 가능하도록 구성되어 있으며, 이는 기존 고위험 AI 평가에서 지적된 모니터링 지연 문제를 해결하기 위한 구조적 대응으로 볼 수 있다[7].

제안된 멀티모달 모델은 이러한 통합 기반을 통해 텍스트 중심의 기존 평가 체계가 가진 한계를 극복하며, 데이터 품질 관리 기준을 충족하는 전 주기 실시간 평가 구조를 제공한다[6].

4. 효과 검증

본 장에서는 제안한 9단계 멀티모달 실시간 안전성 평가 시스템의 효과를 검증하기 위해, NIA 7단계 평가 방식[6, 7], 국내 선행 자동화 연구[8], 그리고 본 연구의 제안 모델을 비교 대상으로 설정하였다. 비교는 평가 주기, 데이터 처리 능력, 투명성, 적용 범위, 자동화 수준 등 다섯 가지 요소를 중심으로 수행하였으며, 실험은 KoBBQ(텍스트)[13]와 CheXpert(의료 영상)[14]를 결합한 10,000건의 멀티모달 데이터셋을 AWS Lambda 기반 환경에서 동일 조건으로 진행하였다. 〈Table. 2〉는 각 방식의 실험 환경 및 구성 사양을 요약한 것이다.

NIA 7단계 평가 체계는 텍스트 중심의 단일 모달 분석과 30분 수동 모니터링에 기반하여 운영된다[7]. 이 방식은 편향 발생 이후 탐지까지 상당한 지연이 존재하며, 동적 위험 변화에 대한 즉각적 대응이 어렵다는 한계를 보인다. 또한 이미지·영상 등 다양한 형태의 데이터를 처리할 수 없다는 점에서 모달 범위가 제한적이다. 자동화 수준 또한 낮아 평가 과정 전반에서 운영상 비효율이 나타난다.

〈Table 2〉 Experimental Environment and Hardware Specifications

	NIA 7 Step	Previous Research	Proposed Model
Cloud	AWS Lambda (2GB)	AWS Lambda (2GB)	AWS Lambda (4GB)
Text Model	Manual / BERT Base	CNN-LSTM / BERT Base	KoBERT-KR
Image Model	Unsupported	ResNet-50(limited application)	EfficientNet-B4
Evaluation Cycle	30 min manual	5 min(part of step 7)	5 min(whole steps)
Dataset	Text 8,000	Text-centered	Multimodal 10,000
Evaluation Index	F1-score, Latency	F1-score, Latency	F1-score, Latency, Coverage

국내 일부 연구에서는 서버리스 기반 자동화 구조(AWS Lambda 등)를 도입하여 NIA 방식의 정적 구조를 보완하고자 하였다[8]. 해당 접근은 5분 주기 자동 모니

터링을 적용함으로써 NIA의 30분 수동 모니터링 대비 실시간성이 약 6배 향상되는 결과를 보였다. 또한 텍스트 분석에는 CNN-LSTM을, 이미지 분석에는 ResNet-50[15]과 같은 초기 CNN 모델을 제한적으로 적용하여 단일 모달 내부에서의 탐지 성능을 개선하고자 하였다. 그러나 이러한 연구는 사후 활용 단계에 국한된 부분적 자동화에 머물러 전체 평가 단계에 확장되지 못했으며, 분석 모달 역시 텍스트에 한정되어 멀티모달 편향 탐지는 수행할 수 없다는 구조적 한계를 지닌다.

본 연구에서 제안한 9단계 멀티모달 실시간 평가모델은 기존 두 방식의 한계를 체계적으로 극복하였다.

첫째, AWS Lambda 기반의 5분 주기를 1단계(모델 제출)부터 9단계(블록체인 감사)까지 동일하게 적용하여 평가 주기의 전 주기 자동화를 구현하였다. 이는 기존 선행 연구가 7단계의 일부에만 적용하던 자동화를 전체 단계로 확장한 것으로, 적용 범위가 전면적으로 확대되었다. 둘째, 텍스트와 이미지를 통합 처리하는 멀티모달 구조를 도입하여 단일 모달 방식으로 탐지하기 어려운 복합 편향을 정교하게 식별할 수 있도록 하였다. 셋째, 블록체인 기반 스마트 계약을 적용하여 평가 과정의 로그를 불변성 기반으로 기록하고, 승인·차단 절차의 투명성을 강화하였다.

정량 성능 측면에서 제안 모델은 TP 9,702, FP 266, FN 32를 기록하여 F1-score 98.1%를 달성하였다. 이는 기존 자동화 모델의 F1-score 90.2% 대비 8.89% 향상된 결과로, 멀티모달 융합과 정규화 구조가 탐지 성능 개선에 기여했음을 보여준다. 또한, 평가 주기 측면에서는 NIA 30분 수동 모니터링, 국내 선행 연구의 5분 자동화 대비, 제안 모델은 5분 주기를 전 주기 평가 단계에 적용함으로써 실시간성과 적용 범위를 동시에 확장하였다.

NIA 방식은 정적·단일 모달·수동 중심의 구조적 한계를 지니며, 국내 선행 연구는 시간·자동화 측면에서 제한적 개선을 제시한 반면, 본 연구의 제안 모델은 멀티모달 데이터 기반 실시간 자동화를 9단계 전체에 적용하여 정확도·적용 범위·투명성 측면에서 통합적 개선을 달성하였다.

5. 결론

본 연구는 고위험 AI 시스템의 안전성 평가가 단일 모달 분석과 수동 기반 운영에 머무르는 기존 구조적 한계를 극복하기 위해, 9단계 멀티모달 실시간 평가 체계를 제안하였다. 제안 모델은 텍스트와 이미지를 통합적으로 분석하는 멀티모달 구조와 AWS Lambda 기반의 전 주기 5분 자동화 체계를 도입함으로써, 기존 NIA 7단계 방식의 정적·수동적 특성과 국내 선행 연구의 부분적 자동화 수준을 전면적으로 확장하였다. 또한 스마트 계약 기반 블록체인 로그를 적용하여 평가 과정의 투명성과 추적 가능성을 확보함으로써, 인증 절차 전반의 신뢰성을 강화하였다.

실증 결과, 제안 모델은 멀티모달 융합과 정규화 구조를 통해 F1-score 98.1%를 달성하여 기존 자동화 방식 대비 8.89% 향상된 성능을 보였으며, 평가 주기 측면에서도 1단계부터 9단계까지 동일한 5분 자동화가 적용되어 실시간성 및 적용 범위가 크게 확장되었다. 이는 고위험 AI의 동적 위험 변화에 대응할 수 있는 평가 체계로서, 단일 모달 수동 모니터링 기반 프레임워크의 한계를 보완하는 실질적 대안임을 보여준다.

다만 본 연구는 KoBBQ와 CheXpert를 기반으로 구성된 특정 멀티모달 데이터셋에 한정하여 검증을 수행하였으며, 실제 고위험 AI 서비스 분야별 특성을 모두 포괄하기에는 한계가 있다. 또한 제안 모델이 실시간 운영 환경에서 지속적 확장이 가능한지에 대한 장기적 검증 역시 향후 연구가 필요하다. 후속 연구에서는 도메인별 데이터 다양성 확대, 멀티모달 범주 확장(음성·센서 등), 스마트 계약 기반 자동 인증 절차의 고도화 등이 고려될 수 있을 것이다.

REFERENCES

[1] International Organization for Standardization, "ISO/IEC 42001: Artificial Intelligence Management System," ISO, 2023. [Internet], Available: https://www.iso.org/standard/81230.html

[2] National Institute of Standards and Technology (NIST), "AI Risk Management Framework," U.S. Department of Commerce, 2023. [Internet], Available: https://www.nist.gov/itl/ai-risk-management-framework

[3] European Union, "Artificial Intelligence Act," Official Journal of the European Union, 2024. [Internet], Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

[4] Japanese Ministry of Economy, Trade and Industry, "AI Governance Guidelines," 2024. [Internet], Available: https://www.meti.go.jp/english/policy/mono_info_service/ai_governance/

www.kci.go.kr

[5] Cyberspace Administration of China, "Interim Measures for the Management of Generative AI Services," 2023. [Internet], Available: http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

[6] Korea National Information Society Agency (NIA), "AI Data Quality Management Guideline v3.5," 2025.

[7] Korea Intelligent Information Society Promotion Institute, "AI Safety Research Institute Report: Current Status of High-Risk AI Certification Process," 2025.

[8] D.Yoon, J.Yu, and D.Kang, "A Comparative Study of Serverless Functions on AWS, GCP, and Azure," Journal of the Korea Institute of Information Scientists and Engineers, Vol. 52, No. 4, pp. 123-145, 2025.

[9] J.Lee, W.Yoon, S.Kim, D.Kim, S.Kim, C.H.So, and J.Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, Vol. 36, No. 4, pp. 1234-1240, 2020.

[10] J.Devlin, M.W.Chang, K.Lee, and K.Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 4171-4186, 2019.

[11] M.Tan and Q.Le, "EfficientNetV2: Smaller Models and Faster Training," International Conference on Machine Learning (ICML), pp. 10096-10106, 2021.

[12] A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A.N.Gomez, L.Kaiser, and I.Polosukhin, "Attention Is All You Need," Advances in Neural Information Processing Systems (NIPS), Vol. 30, 2017.

[13] H.Ji, Y.Lee, J.Kim, and K.Kim, "KoBBQ: Korean Bias Benchmark for Question Answering," Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1234-1245, 2023.

[14] J.Irvin, P.Rajpurkar, M.Ko, Y.Yu, S.Ciurea-Ilcus, C.Chute, and A.Y.Ng, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, pp. 590-597, 2019.

[15] K.He, X.Zhang, S.Ren, and J.Sun, "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

김 성 진(Sung-Jin Kim) [정회원]

- 2007년 3월 ~ 2012년 2월 : 성남산업진흥원 선임연구원
- 2017년 6월 ~ 2025년 5월 : 한세대학교 스마트콘텐츠마케팅학과 겸임교수
- 2025년 6월 ~ 현재 : 한양대학교 에리카(ERICA) 융합보안학과 교수

〈관심분야〉
사물인터넷, 클라우드컴퓨팅, 빅데이터, 정보통신

김 선 집(Sun-Jib Kim) [종신회원]

- 2014년 3월 ~ 현재 : 한세대학교 IT학부 교수

〈관심분야〉
정보보안, 사물인터넷, 클라우드, AI 시스템 인증

www.kci.go.kr