Check for updates

Journal of KIIT. Vol. 24, No. 10, pp. 289-302, Oct. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.10.289

생성형 AI 기반 스마트시티를 위한 정량적 AI 보안 위험평가 및 통합 보안 프레임워크

백승수*, 윤승정**

A Quantitative AI Security Risk Assessment and Integrated Security Framework for Generative AI-Enabled Smart Cities

SeungSoo Baek*, SeongJeong Yoon**

요 약

최근 생성형 AI의 확산으로 스마트시티는 도시 운영의 효율성과 서비스 품질이 크게 향상되고 있다. 그러나 Prompt Injection, Data Poisoning 및 Model Drift와 같은 AI 특화 보안위협은 기존 보안체계만으로는 효과적으로 대응하기 어렵다. 본 연구에서는 한국, 미국, 영국, 일본 및 캐나다의 스마트시티 보안모델을 비교·분석하고, OWASP LLM Top 10, NIST AI RMF 및 Microsoft AI Risk Assessment를 통합하여 생성형 AI 기반 스마트시티의 보안 요구사항을 도출하였다. 또한 Prompt Injection Risk, Data Poisoning Risk 및 Model Drift Severity를 활용한 정량적 AI 보안 위험평가 모델과 통합 AI 보안 프레임워크를 제안하였다. 제안한 프레임워크는 생성형 AI 기반 스마트시티의 AI 보안위험을 정량적으로 평가하고 정책 및 운영 의사결정을 지원하기 위한 실질적인 기준으로 활용될 수 있다.

Abstract

The rapid adoption of generative AI has significantly improved the operational efficiency and service quality of smart cities. However, AI introduces new security threats, including prompt injection, data poisoning, and model drift, which cannot be effectively addressed by conventional security approaches. This study compares smart city security models in South Korea, the United States, the United Kingdom, Japan, and Canada, and derives AI security requirements by integrating the OWASP LLM Top 10, NIST AI Risk Management Framework (AI RMF), and Microsoft AI Risk Assessment. A quantitative AI security risk assessment model based on Prompt Injection Risk, Data Poisoning Risk, and Model Drift Severity is proposed, together with an integrated AI security framework. The proposed framework provides a practical foundation for quantitative AI risk assessment and risk-informed decision-making in generative AI-enabled smart cities.

Keywords
generative AI, smart city security, AI risk assessment, AI security framework

* ICT폴리텍대학 정보보안학과
- ORCID: https://orcid.org/0000-0001-5531-7468
** Swiss School of Management, DBA(교신저자)
- ORCID: https://orcid.org/0000-0001-5188-801X

· Received: Jul. 10, 2026, Revised: Jul. 30, 2026, Accepted: Aug. 02, 2026
· Corresponding Author: SeongJeong Yoon
Dept. of AI Bigdaata, Swiss School of Management, DBA, USA, 9535 Forest Lane, Suite 125D Dallas, Texas 75243
Tel.: +82-31-760-3397, Email: sj9416@naver.com

Ⅰ. 서 론

최근 스마트시티는 사물인터넷(IoT, Internet of Things), 빅데이터(Big data), 클라우드 컴퓨팅 및 생성형 인공지능(Generative AI) 등 첨단 디지털 기술의 발전과 함께 도시 운영의 효율성과 시민 삶의 질을 향상시키는 핵심 도시 모델로 빠르게 발전하고 있다. 특히 생성형 AI는 대규모 도시 데이터를 실시간으로 분석·예측하여 교통, 에너지, 공공안전, 재난 대응, 의료 및 행정서비스 등 다양한 분야에서 의사결정을 지원하는 핵심 기술로 활용되고 있다. 최근 연구에서는 생성형 AI와 AI 기반 보안기술이 스마트시티의 운영 효율성과 도시 서비스의 지능화를 가속화하는 핵심 기술로 평가되고 있다[1][2]. 이처럼 생성형 AI는 스마트시티의 지능화와 자율화를 가속화하며 도시 운영의 생산성과 서비스 품질을 향상시키는 핵심 기반기술로 자리매김하고 있다. 그러나 생성형 AI의 확산은 기존 정보보안 체계에서 충분히 고려하지 못한 새로운 보안위협을 발생시키고 있다. 프롬프트 인젝션(Prompt injection), 데이터 포이즈닝(Data poisoning), 모델 탈취(Model Theft), 환각현상(Halluciantion) 및 모델 드리프트(Model drift)와 같은 AI 특화 공격은 기존의 네트워크 및 시스템 중심 보안기술만으로는 효과적으로 대응하기 어렵다[3][4]. 특히 학습데이터 변조나 악성 프롬프트 입력은 AI 모델의 판단 결과를 왜곡하여 잘못된 의사결정을 유도할 수 있으며, 스마트시티와 같이 전력망, 교통신호, 상수도 및 공공안전 등 국가 핵심 기반시설을 AI가 자동으로 제어하는 환경에서는 단일 AI 시스템의 오류가 연쇄적인 장애(Cascading failure)로 확산되어 사회·경제적 피해를 초래할 가능성이 높다[5][6]. 또한, 생성형 AI는 시민의 이동정보, 에너지 사용량, 의료정보 등 대규모 개인정보를 지속적으로 수집·활용하기 때문에 개인정보 보호, 데이터 거버넌스, AI 의사결정의 투명성(Transparency) 및 설명 가능성(Explainability) 확보가 새로운 보안 요구사항으로 부각되고 있다[5][6]. 기존 스마트시티 보안 연구는 네트워크 보안, 접근통제, 암호화 및 개인정보 보호를 중심으로 발전해 왔으며, 정보시스템을 보호하기 위한 다양한 보안기술과 관리체계를 제안하였다. 그러나 이러한 연구는 생성형 AI 모델의 학습·추론 과정에서 발생하는 프롬프트 인젝션, 데이터 포이즈닝, Model Drift와 같은 AI 특화 보안위협을 충분히 반영하지 못하고 있으며[7], 대부분 정성적인 보안 요구사항을 제시하는 수준에 머물러 있다. 또한, 최근 AI 위험관리 프레임워크는 AI 거버넌스와 위험관리 원칙을 제시하고 있으나, 스마트시티 환경에서 AI 보안위험을 정량적으로 평가하고 운영 의사결정을 지원할 수 있는 위험평가 모델은 아직 부족한 실정이다[8]. 이에 본 연구에서는 생성형 AI 기반 스마트시티 환경에서 요구되는 AI 보안관리 체계를 구축하기 위하여 한국, 미국, 영국, 일본 및 캐나다의 스마트시티 보안모델을 비교·분석하였다. 또한, OWASP LLM Top 10, NIST AI RMF(Risk Management Framework) 및 Microsoft AI Risk Assessment Framework를 통합 분석하여 생성형 AI 기반 스마트시티에 적합한 보안 요구사항을 도출하였다. 이를 바탕으로 프롬프트 인젝션 Risk, 데이터 포이즈닝 Risk 및 모델 드리프트 Severity를 활용한 정량적 AI 보안 위험평가 모델을 제안하고, AI 거버넌스, 위험관리, AI 생애주기 관리, 개인정보 보호, 지속적 모니터링 및 교육·인식 제고를 포함하는 통합 AI 보안 프레임워크를 설계하였다.

본 연구의 차별성은 기존 연구가 생성형 AI 보안 요구사항을 정성적으로 제시하는 수준에 머물렀던 한계를 극복하고, AI 보안위험을 정량적으로 평가하여 정책 및 운영 의사결정을 지원할 수 있는 통합 보안관리 체계를 제안하였다는 점에 있다. 제안한 프레임워크는 생성형 AI 기반 스마트시티뿐만 아니라 자율주행, 스마트팩토리, 디지털 헬스케어 등 다양한 AI 기반 서비스의 보안관리 모델로 확장 활용될 수 있으며, 안전하고 신뢰할 수 있는 AI 중심 사회를 구현하기 위한 실질적인 기준을 제공할 것으로 기대된다. 본 논문의 구성은 다음과 같다. 제Ⅱ장에서는 한국, 미국, 영국, 일본 및 캐나다의 스마트시티 보안모델과 주요 생성형 AI 보안 프레임워크를 비교·분석하고, 기존 연구의 한계와 정량적 AI 보안 위험평가의 필요성을 고찰한다. 제Ⅲ장에서는 국제 AI 보안 프레임워크를 기반으로 프롬프트 인젝션, 데이터 포이즈닝, 모델 드리프트 및 알고리즘 편향에 대한 핵심 위험지표를 도출하고, 이를 통합한 정량적 AI 보안 위험평가 모델과 통합

위험지수 산정 방법을 제시한다. 제Ⅳ장에서는 생성형 AI 기반 스마트시티 교통신호 제어 시스템을 대상으로 가상 데이터 기반 시뮬레이션을 수행하여 제안한 위험평가 모델의 적용 가능성을 분석하고, 위험등급별 대응방안과 실무적 활용 방법을 제시한다. 마지막으로 제Ⅴ장에서는 연구결과를 종합하여 이론적·실무적 시사점과 연구의 한계를 논의하고, 향후 연구 방향을 제시한다.

II. 관련 연구

2.1 국가별 스마트시티 보안모델 비교

스마트시티는 교통, 에너지, 공공안전, 재난관리 등 다양한 도시 서비스를 정보통신기술(ICT, Information and Communication Technology 인공지능(AI)을 기반으로 통합 운영하는 도시 모델로 발전하고 있다. 이에 따라 각국은 도시 인프라의 안정성과 시민의 안전을 확보하기 위하여 국가별 정책과 환경에 적합한 스마트시티 보안모델을 구축하고 있다[9][10].

우리나라는 국토교통부와 과학기술정보통신부를 중심으로 스마트시티 통합플랫폼을 구축하여 교통, 재난, 방범 및 에너지 서비스를 연계하고 있으며, 개인정보 보호와 통합관제 중심의 보안체계를 운영하고 있다. 또한, 국내에서는 스마트시티 환경에 적합한 사이버보안 체계와 정보보호 산업 기반의 보안모델에 관한 연구가 지속적으로 수행되고 있다[9]-11]. 미국은 NIST CSF(ybersecurity Framework)와 AI RMF를 기반으로 AI 거버넌스와 위험관리를 통합적으로 관리하고 있으며, 캐나다는 데이터 거버넌스와 국가안보를 동시에 고려하는 스마트시티 정책을 추진하고 있다. 일본은 Smart City Security Guidelines를 통해 IoT 기반 도시 서비스의 위험관리를 체계화하였고[4], 영국은 NCSC(ational Cyber Security Centre)를 중심으로 Security by Design과 Zero Trust 기반의 스마트시티 보안체계를 운영하고 있다. 그러나 대부분의 국가들은 네트워크, IoT 및 도시 인프라 보호를 중심으로 보안정책을 수립하고 있으며, 생성형 AI 환경에서 발생하는 프롬프트 인젝션, 데이터 포이즈닝, 모델 드리프트와 같은 AI 특화 보안위협을 체계적으로 관리하거나 정량적으로 평가할 수 있는 체계는 아직 부족한 실정이다[12][13]. 표 1은 한국, 미국, 영국, 일본 및 캐나다의 스마트시티 보안모델을 비교한 결과를 나타낸다. 국가별 보안정책의 중점 분야와 AI 보안 성숙도를 비교함으로써 생성형 AI 환경에서의 한계를 분석하였다.

2.2 생성형 AI 보안 프레임워크 비교

생성형 AI는 기존 정보시스템과 달리 학습, 추론 및 지속적인 모델 업데이트 과정을 수행하므로 새로운 형태의 보안관리 체계가 요구된다. 최근 AI 보안 분야에서는 NIST AI RMF, OWASP Top 10 for LLM Applications, Microsoft AI Risk Assessment가 대표적인 국제 프레임워크로 활용되고 있다. NIST AI RMF는 AI 거버넌스와 위험관리 중심의 정책 프레임워크로 AI 생애주기 전반에 걸쳐 Govern, Map, Measure, Manage의 네 가지 핵심 기능을 제시한다[2]. OWASP LLM Top 10은 프롬프트 인젝션, 데이터 포이즈닝, 모델 탈취, 민감정보누설 등 생성형 AI 환경에서 발생하는 대표적인 기술적 공격을 정의하고 있으며, Microsoft AI Risk Assessment는 STRIDE 기반 위협모델을 AI 환경으로 확장하여 데이터, 모델 및 추론과정을 포함하는 위험평가 방법을 제안하였다 [14]-[16]. 표 2는 대표적인 생성형 AI 보안 프레임워크를 비교한 결과이다. 각 프레임워크의 주요 특징과 장단점을 분석하여 본 연구에서 통합 프레임워크를 설계하기 위한 기반으로 활용하였다.

표 2. 생성형 AI 보안 프레임워크 비교
Table 2. Comparison of generative AI security frameworks

Framework	Key caracteristics	Strength	Limitations
NIST AI RMF	Governance and risk management	AI policy framework	Insufficient quantitative assessment
OWASP LLM tp 10	Response to AI-specific attacks	Technical threat analysis	Limited policy support
Microsoft AI rsk asessment	AI threat analysis	Development-stage risk assessment	Limited applicability at the city scale

표 1. 국가별 스마트시티 보안모델 비교
Table 1. Comparison of national smart city security models

Country	Security fcus	AI scurity mturity	Key Imitations
South Korea	Integrated Smart City Platform	Low	Lack of AI risk assessment framework
United States	NIST-based AI Risk Management	Medium	Limited quantitative AI risk metrics
Canada	Data Governance and Privacy	Medium	Insufficient operational evaluation indicators
Japan	IoT Security and Risk Management	Low	Model drift not considered
United kingdom	Security by Design and Zero Trust	Medium	Limited protection against generative AI threats

이들 프레임워크는 생성형 AI 보안 요구사항을 체계적으로 제시하고 있다는 장점이 있으나, 스마트시티와 같이 다양한 AI 서비스가 동시에 운영되는 환경에서 AI 보안위험을 정량적으로 평가하고 운영 우선순위를 결정할 수 있는 기준은 제시하지 못하고 있다.

2.3 AI 보안 정량 위험평가의 필요성

생성형 AI 기반 스마트시티에서는 AI가 교통신호, 에너지 관리, 공공안전 및 재난 대응과 같은 핵심 서비스를 실시간으로 제어하므로 AI 모델 자체의 신뢰성과 안전성을 지속적으로 평가할 수 있는 정량적 위험평가 체계가 필수적이다. 최근에는 AI 미들웨어 기반의 실시간 영상분석과 분산 AI 처리 기술이 스마트시티 핵심 인프라로 활용되면서 AI 모델에 대한 지속적인 보안평가의 중요성이 더욱 강조되고 있다[17][18]. 기존 스마트시티 보안모델은 네트워크 및 시스템 보호에 중점을 두고 있어 AI 모델의 학습과 추론 과정에서 발생하는 위험을 객관적으로 평가하기 어렵다. 또한, 생성형 AI 환경에서는 프롬프트 인젝션, 데이터 포이즈닝, 모델 드리프트, 할루시네이션 등 다양한 위험이 동시에 발생하며, 운영자는 제한된 시간 안에 어떤 위험을 우선적으로 대응해야 하는지를 판단해야 한다[19][20]. 그러나 기존 연구는 대부분 정성적 보안 요구사항을 제시하는 수준에 머물러 있어 AI 위험의 상대적 중요도를 비교하거나 운영 의사결정을 지원하기에는 한계가 있다[21][22]. 이에 본 연구에서는 생성형 AI 기반 스마트시티 환경에서 발생 빈도와 영향도가 높은 프롬프트 인젝션 Risk, 데이터 포이즈닝 Risk 및 모델 드리프트 Severity를 핵심 위험지표로 선정하고, 이를 통합하여 AI 시스템의 전체 위험수준을 평가할 수 있는 Integrated AI Risk Score를 제안하였다.

2.4 기존 연구의 한계 및 연구 방향

기존 스마트시티 보안 연구는 네트워크 보안과 개인정보 보호를 중심으로 발전하였으며, 최근에는 AI 거버넌스와 AI 보안 프레임워크 연구가 활발히 진행되고 있다[9]-[16]. 그러나 대부분의 연구는 ① 생성형 AI 특화 보안위협을 충분히 반영하지 못하고, ② AI 보안위험을 정량적으로 평가할 수 있는 기준을 제시하지 못하며, ③ 정책 및 운영 의사결정을 지원하는 통합 관리체계를 제공하지 못하는 한계를 가진다. 따라서 본 연구에서는 국가별 스마트시티 보안모델과 국제 AI 보안 프레임워크를 통합적으로 분석하고, AI 보안위험을 정량적으로 평가할 수 있는 위험평가 모델과 통합 AI 보안 프레임워크를 제안함으로써 기존 연구의 한계를 보완하고자 한다. 표 3은 기존 연구와 본 연구를 비교한 것이다. 기존 연구가 정성적 보안관리 중심인 반면, 본 연구는 정량적 AI 위험평가와 운영 의사결정 지원을 포함하는 통합 프레임워크를 제안한다.

표 3. 생성형 AI 보안 프레임워크 비교
Table 3. Comparison between previous studies and the proposed study

Category	Previous studies	Proposed study
Research scope	Network and system security	Generative AI-enabled smart city security
Target threats	Conventional cyber threats	Prompt injection, data poisoning, and model drift
Risk assessment	Qualitative security assessment	Quantitative AI security risk assessment
Decision support	Security guidelines	Risk-informed policy and operational decision-making
Research outcome	Individual security frameworks	Integrated AI security framework

III. AI 보안 정량 위험평가 모델도출 방법론

3.1 AI 보안 위험지표 선정 과정

생성형 AI 기반 스마트시티에서는 기존의 네트워크 및 시스템 중심 보안만으로는 AI 모델의 학습, 추론 및 운영 과정에서 발생하는 새로운 보안위협을 효과적으로 관리하기 어렵다. 특히 스마트시티는 교통, 에너지, 공공안전, 의료 및 행정서비스 등 다양한 AI 서비스가 실시간으로 운영되므로, AI 모델의 신뢰성과 안전성을 지속적으로 평가할 수 있는 정량적 위험평가 체계가 요구된다[23]. 본 연구에서는 생성형 AI 환경에서 발생 가능한 보안위협을 체계적으로 도출하기 위하여 OWASP LLM Top 10, NIST AI RMF 및 Microsoft AI Risk Assessment Framework를 종합적으로 분석하였다. OWASP LLM Top 10은 생성형 AI에서 발생하는 대표적인 기술적 공격 유형을 제시하고 있으며, NIST AI RMF는 AI 거버넌스와 위험관리 체계를 제공한다. 또한, Microsoft AI Risk Assessment는 AI 모델의 개발, 운영 및 배포 과정에서 발생할 수 있는 위험요인을 체계적으로 관리하기 위한 평가 기준을 제시하고 있다.

그러나 세 가지 프레임워크 모두 생성형 AI의 보안 요구사항을 제시하는 데 중점을 두고 있을 뿐, 스마트시티 환경에서 다양한 AI 보안위험의 우선순위를 정량적으로 평가하거나 운영 의사결정을 지원할 수 있는 위험지표는 제시하지 못하고 있다[24][25]. 이에 본 연구에서는 다음의 세 가지 기준을 적용하여 AI 보안 위험지표를 선정하였다. ① 국제 AI 보안 프레임워크에서 공통적으로 제시되는 위험요인일 것, ② 스마트시티 운영의 안정성과 시민 안전에 직접적인 영향을 미칠 것, ③ 로그 및 운영데이터를 이용하여 정량적으로 측정이 가능할 것. 이러한 기준을 적용한 결과 프롬프트 인젝션, 데이터 포이즈닝, 모델 드리프트 및 알고리즘 편향(Algorithmic bias)를 생성형 AI 기반 스마트시티의 핵심 AI 보안 위험지표로 선정하였다. 프롬프트 인젝션은 생성형 AI 서비스에서 가장 빈번하게 발생하는 입력 기반 공격이며[26][27], Data Poisoning은 AI 모델의 학습데이터를 변조하여 잘못된 의사결정을 유도하는 대표적인 위협이다. 또한, 모델 드리프트는 AI 모델의 성능 저하를 지속적으로 평가하기 위한 핵심 지표이며, 알고리즘 편향은 특정 지역이나 집단에 대한 편향된 의사결정을 측정하여 AI의 공정성과 신뢰성을 확보하기 위한 지표이다.

표 4. AI 보안위험지표 선정 과정
Table 4. Selection process of AI security risk indicators

AI security threat	Quantitative assessment	International frameworks	OWASP LLM top 10	NIST AI RMF	Microsoft AI risk assessment	Smart city impact	Selected
Prompt injection	Yes	High	●	○	●	High	✓
Data poisoning	Yes	High	●	●	●	High	✓
Model theft	No	Medium	●	○	●	Medium	-
Hallucination	No	Medium	●	○	○	Medium	-
Model drift	Yes	High	○	●	●	High	✓
Algorithmic bias	Yes	High	○	●	●	High	✓
Sensitive information disclosure	Yes	High	●	○	●	Medium	-

Note: ● Applicable, ○ Not applicable

본 연구에서는 이들 개별 위험지표를 통합하여 AI 서비스의 전체 보안수준을 평가할 수 있는 통합 AI 위험지수를 제안하였으며, 이를 통해 스마트시티 운영자가 AI 보안위험을 객관적으로 평가하고 대응 우선순위를 결정할 수 있도록 하였다. 표 4는 생성형 AI 기반 스마트시티에 적용하기 위한 AI 보안 위험지표의 선정 과정이다.

본 연구에서는 OWASP LLM Top 10, NIST AI RMF 및 Microsoft AI Risk Assessment에서 제시하는 주요 AI 보안위협을 비교·분석하고, ① 국제 프레임워크에서 공통적으로 중요하게 다루는 위험인지, ② 스마트시티 운영에 미치는 영향이 큰지, ③ 운영 로그와 데이터를 이용하여 정량적으로 측정 가능한지를 기준으로 핵심 위험지표를 선정하였다. 분석결과 프롬프트 인젝션, 포이즈닝, 모델 드리프트 및 알고리즘 편향은 세 가지 기준을 모두 충족하는 대표적인 AI 보안위협으로 나타났다. 반면 Model Theft, Hallucination 및 Sensitive Information Disclosure는 중요한 보안위협이지만, 스마트시티 운영환경에서 실시간으로 정량화하기 어렵거나 운영 영향도를 직접 평가하기 어려워 본 연구의 핵심 위험지표에서는 제외하였다. 최종적으로 위험평가 모델은 국제 AI 보안 프레임워크와 스마트시티 운영 특성을 동시에 반영한 정량적 AI 보안 위험평가 모델로 구성되었다.

3.2 AI 보안 위험지표

본 연구에서 제안하는 AI 위험평가 모델은 생성형 AI에서 가장 빈번하게 발생하는 네 가지 보안위협을 중심으로 구성하였다.

3.2.1 프롬프트 인젝션 Risk

프롬프트 인젝션 위험(Prompt injection risk)은 악의적인 입력이 AI 모델의 정상적인 동작을 우회하거나 보안정책을 위반하도록 유도할 가능성을 정량적으로 평가하는 지표이다. 본 연구에서는 전체 AI 요청 대비 정책을 위반한 응답의 비율을 이용하여 위험도를 계산하였다.

$$R_{\Pi} = N_v / N_t \tag{1}$$

여기서,
N_v : 정책 위반 응답 수
N_t : 전체 AI 요청 수

3.2.2 데이터 포이즈닝 Risk

데이터 포이즈닝 위험(Data poisoning risk)은 정상 학습데이터와 재학습 데이터 간의 분포 차이를 이용하여 데이터 변조 가능성을 정량적으로 평가하는 지표이다. 본 연구에서는 데이터 분포 간의 차이를 측정하기 위해 KL(Kullback-Leible Divergence를 이용하여 데이터 무결성 수준을 평가하였다.

$$R_{DP} = \sum_{i=1}^{n} P_i \log\left(\frac{P_i}{Q_i}\right) \tag{2}$$

R_{DP} : 데이터 포이즈닝 Risk
P_i : 정상 학습데이터(Baseline dataset)의 i번째 데이터 분포 비율
Q_i : 재학습데이터(Retraining dataset)의 i번째 데이터 분포 비율
n: 데이터 분포를 구성하는 전체 범주의 개수

R_{DP}값이 증가할수록 재학습 데이터의 분포가 정상 학습데이터와 크게 달라졌음을 의미하며, 데이터 포이즈닝에 의해 AI 모델의 학습데이터 무결성이 저하될 가능성이 높은 것으로 판단한다. 따라서 일정 임계 값을 초과하는 경우 데이터 무결성 검증, 이상 데이터 탐지 및 재학습 데이터에 대한 보안점검이 수행되어야 한다.

3.2.3 모델 드리프트 Severity

모델 드리프트 심각도(Model drift severity)는 시간의 경과에 따라 AI 모델의 예측 성능과 설명가능성이 저하되는 정도를 정량적으로 평가하는 지표이다. 본 연구에서는 기준 모델(Baseline model)과 현재 운영 모델(Current model)의 Accuracy, F1-score 및

Explainability 변화량을 종합하여 모델 성능 저하 수준을 산출하였다.

$$R_{MD} = \frac{1}{3}(\Delta Accuracy + \Delta F1 + \Delta Explainability) \quad (3)$$

R_{MD}: 모델 드리프트 심각도(Model drift severity)
$\Delta Accuracy$: 기준 모델과 현재 모델 간 정확도(Accuracy)의 변화량
$\Delta F1$: 기준모델과 현재 모델 간 F1-score 변화량
$\Delta Explainability$: 기준 모델과 현재 모델 간 설명 가능성(Explainability)의 변화량

3.2.4 알고리즘 편향 Risk

알고리즘 편향 위험(Algorithmic bias risk)은 AI가 특정 지역이나 집단에 대해 편향된 의사결정을 수행하는 정도를 정량적으로 평가하는 지표이다. 본 연구에서는 각 집단의 긍정적 의사결정 비율과 전체 평균 의사결정 비율의 차이를 이용하여 알고리즘의 공정성을 평가하였다.

$$R_{AB} = \frac{1}{n}\sum_{i=1}^{n} abs(P_i - \overline{P}) \quad (4)$$

R_{AB}: 알고리즘 편향 위험(Algorithmic bias risk)
P_i: i번째 지역 또는 집단의 긍정적 의사결정 비율
$\overline{P}$: 전체 평균 긍정적 의사결정 비율
n : 비교 대상 지역 또는 집단의 개수

R_{AB} 값이 증가할수록 특정 지역이나 집단에 대한 AI 의사결정의 편향성이 커짐을 의미한다. 이는 AI 모델의 공정성과 신뢰성이 저하될 가능성을 나타내며, 데이터 재균형(Data rebalancing), 목적함수 개선(Objective function optimization) 및 공정성(Fairness) 검증을 통해 편향을 완화할 필요가 있다.

3.2.5 통합 인공지능 위험 점수

통합 AI 위험지수(Integrated AI risk score)는 개별 AI 보안 위험지표를 하나의 종합지표로 통합하여 스마트시티 AI 서비스의 전체 보안위험 수준을 정량적으로 평가하기 위한 지표이다. 본 연구에서는 개별 위험지표의 상대적 중요도를 반영하기 위하여 예시적 가중치(Illustrative weights)를 적용하여 통합 위험지수를 산출하였다.

$$R_{AI} = \sum_{i=1}^{n} w_i R_i \quad (5)$$

R_{AI}: 통합 AI 위험지수(Integrated AI risk score)
w_i : i번째 위험지표의 가중치(Weight)
R_i : 정규화된 i번째 위험지표 값
n : 통합에 사용된 위험지표의 개수

통합 AI 위험지수는 프롬프트 인젝션 위험(R_{PI}), 데이터 포이즈닝 위험(R_{DP}), 모델 드리프트 심각도(R_{MD}) 및 알고리즘 편향 위험(R_{AB})을 하나의 지표로 통합하여 AI 시스템의 종합적인 보안위험 수준을 나타낸다.

R_{AI}값이 증가할수록 AI 서비스의 전체 보안위험이 높음을 의미하며, 운영자는 이를 기반으로 위험 수준을 신속하게 판단하고 데이터검증, 모델 재학습, 정책 수정 및 보안점검 등의 대응 우선순위를 결정할 수 있다.

Ⅳ. 분석 결과

4.1 AI 보안 위험지표 시나리오 검증

본 연구에서 설정한 시뮬레이션 시나리오는 표 5와 같이 생성형 AI 기반 스마트시티 교통신호 제어 시스템을 대상으로 구성하였다. 본 연구에서는 정상 운영 상태를 기준으로 생성형 AI 환경에서 발생 가능한 대표적인 보안위협을 단계적으로 적용하였다. 먼저 AI가 실시간 교통량을 분석하여 교통신호를 자동제어하는 환경을 설정한 후, Prompt Injection 공격, 학습데이터 변조(Data poisoning), AI 모델 성능 저하(Model drift) 및 알고리즘 편향(Algorithmic bias)을 순차적으로 발생시키는 상황을 가정하였다.

표 5. 생성형 AI 기반 스마트시티 AI 보안 위험평가 시나리오
Table 5. Simulation scenarios for AI security risk assessment in a generative AI-enabled smart city

Step	Simulation scenario	AI security impact	Evaluation item
Scenario 1	Generative AI automatically controls traffic signals during peak-hour traffic conditions.	AI-enabled real-time operational decision-making	Normal operation
Scenario 2	A malicious user performs a prompt injection attack to manipulate traffic signal operations.	Increased risk of AI policy violations	Prompt injection risk
Scenario 3	Erroneous traffic sensor data are continuously collected and incorporated into the AI retraining dataset.	Degradation of data integrity	Data poisoning risk
Scenario 4	Long-term operation causes a decline in the AI model's prediction accuracy and explainability.	Degradation of AI model performance	Model drift severity
Scenario 5	Traffic signals are repeatedly prioritized for a specific district, resulting in citizen complaints.	Biased AI decision-making	Algorithmic bias risk
Scenario 6	Individual AI security risk indicators are aggregated to assess the overall AI security risk level.	Integrated AI security risk assessment	Integrated AI risk score

마지막으로 각 위험지표를 통합하여 Integrated AI Risk Score를 산출함으로써 생성형 AI 기반 스마트시티의 전체 보안위험 수준을 평가하였다.

4.2 시뮬레이션 입력 데이터 셋 구성

본 연구에서는 제안한 AI 보안 위험평가 모델의 적용 가능성을 검증하기 위하여 생성형 AI 기반 스마트시티 교통신호 제어 시스템을 가상의 운영환경으로 설정하였다. 본 시뮬레이션은 실제 운영데이터가 아닌 개념 검증(Concept validation)을 위한 가상 데이터 셋(Synthetic dataset)을 이용하였다.

입력 데이터는 생성형 AI 서비스 운영 중 일반적으로 수집 가능한 AI 요청 로그, 교통센서 데이터, AI 모델 성능지표 및 지역별 의사결정 결과를 기반으로 구성하였다.

표 6은 제안한 AI 보안 위험평가 모델의 검증을 위해 구성한 가상 입력 데이터 셋(Synthetic dataset)을 나타낸다. 본 데이터 셋은 프롬프트 인젝션 Risk, 데이터 포이즈닝 Risk, 모델 드리프트 Severity 및 알고리즘 편향 Risk를 산출하기 위한 입력 변수와 예시 값을 포함한다. 입력 데이터는 AI 요청 로그, 교통센서 데이터, AI 모델 성능지표 및 지역별 의사결정 결과를 기반으로 구성되었으며, 실제 운영환경이 아닌 개념 검증(Concept validation)을 위한 시뮬레이션 데이터를 사용하였다. 또한, 통합 인공지능 위험 점수 산출을 위한 예시 가중치를 함께 제시하여 이후 위험지표 계산에 활용하였다.

표 7은 표 6에서 제시한 입력 데이터를 이용하여 각 AI 보안 위험지표를 계산한 결과를 나타낸다. 본 시뮬레이션에서는 생성형 AI 기반 스마트시티 교통신호 제어 환경을 가정하여 프롬프트 인젝션 데이터 포이즈닝, 모델 드리프트 및 알고리즘 편향시나리오를 순차적으로 적용하였다. 각 단계에서 제안한 정량적 위험평가 수식을 이용하여 위험도를 산출하고, 위험 수준에 따른 대응방안을 함께 제시하였다. 마지막으로 개별 위험지표를 통합하여 Integrated AI Risk Score를 계산함으로써 전체 AI 보안위험 수준을 종합적으로 평가하였다. 표 7의 계산 결과를 살펴보면, Prompt Injection Risk는 0.080으로 나타나 입력 검증 강화가 필요한 수준으로 평가되었다.

표 6. 생성형 AI 기반 스마트시티 AI 보안 위험평가 시나리오
Table 6. Synthetic input dataset for AI security risk assessment

AI security risk indicator	Input variable	Synthetic input data	Data source	Used in equation
Prompt injection risk	Total AI requests	100 requests	AI Request logs	Eq. (1)
	Policy-violating responses	8 responses		
Data poisoning risk	Baseline dataset distribution	Free Flow: 3,300 Moderate: 3,300 Congested: 3,400	Traffic sensor data	Eq. (2)
	Retraining dataset distribution	Free Flow: 2,000 Moderate: 3,000 Congested: 5,000		
Model drift severity	Model dccuracy	95% (Baseline) → 91% (Current)	AI model performance metrics	Eq. (3)
	F1-score	94% (Baseline) → 88% (Current)		
	Explainability	92% (Baseline) → 84% (Current)		
Algorithmic bias risk	AI decision ratio by region	Region A: 75% Region B: 58% Region C: 45%	Regional AI decision outcomes	Eq. (4)
	Overall average decision ratio	60%		
Integrated AI risk score	Illustrative weights	Prompt injection = 0.30 Data poisoning = 0.25 Model drift = 0.20 Algorithmic bias = 0.25	Risk indicator weights	Eq. (5)

표 7. AI 보안 위험평가를 위한 시뮬레이션 입력 데이터 및 계산 결과
Table 7. Simulation input data and calculation results for a security risk assessment

Step	Simulation scenario	Risk indicator	Simulation input data	Data source	Risk acore	Recommended action
①	A malicious user repeatedly submits prompts to manipulate traffic signal operations.	Prompt injection risk	Total AI requests: 100Policy-violating responses: 8	AI request logs	0.080 식(1)	Strengthen input validation and AI policy filtering
②	Abnormal traffic sensor data are continuously collected due to sensor malfunction or a poisoning attack.	Data poisoning risk	Baseline: free flow 3,300, Moderate 3,300, Congested 3,400Poisoned: Free flow 2,000, Moderate 3,000, Congested 5,000	Traffic sensor data	0.064 식(2)	Verify data integrity and suspend AI model retraining
③	Long-term operation causes a decline in AI model performance.	Model drift severity	Accuracy 95% → 91%F1-score 94% → 88%Explainability 92% → 84%	AI model performance metrics	0.054 식(3)	Retrain the AI model and improve model performance
④	Traffic signals are repeatedly prioritized for a specific district, resulting in unfair service allocation.	Algorithmic bias risk	Region A 75% Region B 58% Region C 45% Overall average 60%	Regional AI decision outcomes	0.150 식(4)	Review fairness and rebalance training data
⑤	Individual AI security risk indicators are integrated to estimate the overall AI security risk level.	Integrated AI risk score (Illustrative)	Illustrative weights PI 0.30, DP 0.25, MD 0.20, Bias 0.25	Risk indicator weights	0.088 식(5)	Perform integrated security inspection and support operational decision-making

Data Poisoning Risk는 0.064로 산출되어 데이터 무결성 검증과 재학습 데이터 확인이 요구되는 것으로 분석되었다. 또한, Model Drift Severity는 0.054로 계산되어 장기간 운영에 따른 모델 성능 저하 가능성을 확인하였으며, Algorithmic Bias Risk는 0.150으로 나타나 지역 간 의사결정의 공정성 개선이 필요한 것으로 평가되었다. 마지막으로 개별 위험지표를 식 (5)에 따라 통합한 결과 Integrated AI Risk Score는 0.088로 산출되었으며, 이는 제안한 모델이 다양한 AI 보안위험을 하나의 종합지표로 통합하여 운영자의 보안 대응 우선순위 결정과 정책적 의사결정을 지원할 수 있음을 보여준다.

4.3 AI 보안 위험지표 실무 활용 방법

표 8은 생성형 AI 기반 스마트시티에서 발생할 수 있는 주요 보안 위험을 정량적으로 평가하기 위하여 선정한 핵심 위험지표와 위험 수준의 기준값을 나타낸 것이다. 기존 스마트시티 보안 모델은 네트워크 및 시스템 중심의 보안관리에 초점을 두고 있으나, 생성형 AI 환경에서는 AI 모델 자체의 신뢰성과 안전성을 평가할 수 있는 새로운 위험지표가 요구된다. 본 연구에서는 생성형 AI의 대표적인 보안 위험을 반영하기 위하여 프롬프트 인젝션 Risk, 데이터 포이즈닝 Risk, 모델 드리프트 Severity 및 알고리즘 편향 Risk를 핵심 위험지표로 선정하였다. 또한, 개별 위험지표를 하나의 통합지수(Integrated AI risk score)로 산출하여 AI 시스템의 전체 위험 수준을 직관적으로 파악할 수 있도록 하였다.

제안한 위험지표는 AI 서비스 운영 과정에서 지속적으로 수집되는 로그와 운영데이터를 기반으로 실시간 위험을 평가하며, 위험 수준이 임계 값을 초과할 경우 데이터 검증, 모델 재학습, 정책 수정 및 운영자 개입 등의 대응 의사결정을 지원한다. 표 8과 9는 본 연구에서 제안하는 생성형 AI 보안 위험지표의 측정 목적과 운영 활용 방안을 나타낸 것이다.

표 8. 생성형 AI 보안 위험지표와 정의 및 활용
Table 8. Definition, risk thresholds, and applications of generative AI security risk indicators

Risk Indicator	Measurement objective	Illustrative risk threshold	Operational response
Prompt injection risk	Measures the proportion of AI responses affected by malicious prompt injection using AI input/output logs.	Low: <0.05 Medium: 0.05 – 0.10 High: >0.10	Detect policy violations, strengthen prompt validation, and perform administrator review.
Data poisoning risk	Measures abnormal changes in training data distribution using KL Divergence.	Low: <0.05 Medium: 0.05 – 0.10 High: >0.10	Verify data integrity and suspend AI retraining when necessary.
Model drift severity	Measures AI model degradation based on Accuracy, F1-score, and Explainability changes.	Low: <0.03 Medium: 0.03 – 0.07 High: >0.07	Determine the timing for model retraining and performance optimization.
Algorithmic bias risk	Measures fairness by evaluating deviations among different regions or user groups.	Low: <0.10 Medium: 0.10 – 0.20 High: >0.20	Review fairness and rebalance training data.
Integrated AI risk	Evaluates the overall AI security risk by integrating individual risk indicators.	Low: <0.05 Medium: 0.05 – 0.10 High: >0.10	Prioritize AI security responses and support operational decision-making.

표 9. AI 보안위험등급 기준 및 대응방안
Table 9. AI Security risk classification and response guidelines

Risk score	Level	Meaning	Action
0.00 – 0.05	Low	Normal operation	Continuous monitoring
0.05 – 0.10	Moderate	Increased risk	Administrator review
0.10 – 0.20	High	Immediate action required	Model retraining / Data verification
>0.20	Critical	AI service at risk	Suspend AI service & incident response

또한, 운영자가 AI 보안위험 수준을 직관적으로 판단할 수 있도록 예시적인 위험등급(Illustrative risk threshold)을 함께 제시하였다. 본 연구에서 제안한 위험등급은 실제 스마트시티 운영환경의 절대적인 기준값이 아니라, 제안한 위험평가 모델의 활용 가능성을 설명하기 위한 개념적 기준(Conceptual threshold)이다.

표 8은 본 연구에서 제안한 AI 보안 위험지표의 측정 목적과 활용 방안을 제시한다. 각 위험지표는 측정 대상, 예시 위험등급 및 운영 대응방안을 함께 제시하여 AI 운영자가 위험 수준을 직관적으로 판단할 수 있도록 구성하였다.

표 9는 통합 AI 위험지수에 따른 위험등급과 대응 절차를 나타낸다. 운영자는 위험 수준에 따라 지속 모니터링, 관리자 검토, 모델 재학습 및 서비스 중단 등의 대응 전략을 선택할 수 있다. 향후, 실제 운영데이터와 전문가 평가를 이용하여 위험등급을 재설정하고, 분야별 특성을 반영한 동적 위험 기준(Dynamic risk threshold)으로 확장할 수 있다.

Ⅴ. 시사점 및 결론

생성형 AI는 스마트시티의 운영 방식을 근본적으로 변화시키고 있으며, 교통, 에너지, 재난안전, 공공행정, 의료 등 다양한 도시 서비스에서 실시간 의사결정을 수행하는 핵심 기술로 자리잡고 있다. 그러나 생성형 AI는 기존 정보시스템과 달리 스스로 학습하고 추론하는 특성을 가지므로 Prompt Injection, Data Poisoning, Model Drift, Algorithmic Bias와 같은 새로운 보안위험이 지속적으로 발생한다. 이러한 위협은 기존의 네트워크 및 시스템 중심 보안체계만으로는 효과적으로 대응하기 어려우며, AI 모델 자체의 신뢰성과 안전성을 확보할 수 있는 새로운 보안관리 체계가 요구된다.

본 연구에서는 이러한 한계를 해결하기 위하여 한국, 미국, 영국, 일본 및 캐나다의 스마트시티 보안 모델을 비교·분석하고, OWASP LLM Top 10, NIST AI RMF(Risk Management Framework), Microsoft AI Risk Assessment를 통합하여 생성형 AI 기반 스마트시티에 적합한 보안 요구사항을 체계적으로 도출하였다. 또한, Prompt Injection Risk, Data Poisoning Risk, Model Drift Severity 및 Integrated AI Risk Score를 활용한 정량적 위험평가 모델을 제안하고, AI 거버넌스, 위험관리, AI 생애주기 관리, 개인정보 보호, 지속적 모니터링 및 교육체계를 포함하는 통합 AI 보안 프레임워크를 제시하였다.

본 연구의 가장 큰 의의는 기존 연구가 보안 요구사항을 정성적으로 제시하는 수준에 머물렀던 한계를 극복하고, AI 보안위험을 객관적으로 측정할 수 있는 정량적 평가 기준과 운영 의사결정을 지원하는 통합 관리체계를 제안했다는 점이다. 특히 개별 위험지표를 하나의 통합 위험지수(Integrated AI risk score)로 산출함으로써 AI 시스템의 보안 상태를 직관적으로 파악할 수 있으며, 위험 수준에 따라 데이터검증, 모델 재학습, 정책 수정 및 운영자 개입과 같은 대응 전략을 우선적으로 결정할 수 있도록 하였다. 이는 AI 보안관리를 기술적 대응 수준에서 정책 및 운영 의사결정 수준으로 확장하였다는 점에서 중요한 의미를 가진다. 또한, 본 연구에서 제안한 프레임워크는 특정 스마트시티 서비스에 한정되지 않고 자율주행, 스마트 팩토리, 디지털 헬스케어, 스마트 국방 및 지능형 공공서비스 등 다양한 생성형 AI 기반 시스템에서도 적용 가능하다. 최근에는 디지털트윈과 AI 기반 보안 서비스를 결합한 스마트시티 연구도 활발히 수행되고 있으며[27], 본 연구에서 제안한 정량적 AI 위험평가 프레임워크는 이러한 지능형 도시 서비스의 AI 보안관리 모델로도 확장 적용될 수 있을 것으로 기대된다. 따라서 AI 기술이 사회 전반으로 확산되는 환경에서 안전성과 신뢰성을 확보하기 위한 범용적인 AI 보안관리 모델로 활용될 수 있으며, 향후 국가 AI 정책 수립과 산업현장의 AI 보안 가이드라인 마련에도 실질적인 기준을 제공할 것으로 기대된다.

다만 본 연구에서 제안한 위험평가 모델과 통합 보안 프레임워크는 개념적 설계(Conceptual

framework)를 기반으로 제안되었으며, 실제 스마트시티 운영환경에서의 실증 검증은 수행하지 못하였다.

향후 연구에서는 교통관제, 스마트에너지, 재난관리 등 실제 스마트시티 서비스를 대상으로 제안한 위험지표의 신뢰성과 적용성을 검증하고, AI 운영 로그와 실시간 데이터를 활용한 자동 위험예측 및 대응기술로 확장할 필요가 있다. 또한, AHP, 머신러닝, 강화학습 등을 활용한 동적 위험가중치 산정기법을 적용한다면 보다 지능적이고 자율적인 AI 보안관리 체계를 구축할 수 있을 것으로 기대된다.

결국, 생성형 AI 시대의 스마트시티 경쟁력은 AI를 얼마나 많이 활용하는가가 아니라, AI를 얼마나 안전하고 신뢰성 있게 운영할 수 있는가에 의해 결정될 것이다. 본 연구는 생성형 AI 기반 스마트시티에서 요구되는 새로운 보안 패러다임과 정량적 위험평가 기준을 제시함으로써, 안전하고 신뢰할 수 있는 AI 중심 스마트시티를 구현하기 위한 실질적인 이론적·실무적 기반을 마련하였다는 점에서 중요한 의의를 가진다.

References

[1] Z. Chen, L. Zhao, and G. Yin, "A predictive deep learning model for cyber threat detection in a smart city environment", Security and Privacy, Vol. 9, No. 3, Art. no. e70218, May-Jun. 2026. https://doi.org/10.1002/spy2.70218.

[2] S. Ashraf, C. Choi, and A. Bilal, "CTDMF-SC: Cybersecurity threat detection and mitigation framework for smart cities", International Journal of Information Security, Vol. 25, pp. 1-29, Feb. 2026. https://doi.org/10.1007/s10207-025-01195-w.

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

[4] R. Pedro, M. E. Coimbra, D. Castro, P. Carreira, and N. Santos, "Words become SQL: Securing AI assistants that talk to databases", IEEE Security & Privacy, Vol. 24, No. 3, pp. 8-14, May 2026. https://doi.org/10.1109/MSEC.2025.3650332.

[5] A. Milani, V. Franzoni, and E. Florindi, "Indirect prompt injection in large language models", Neural Computing and Applications, Vol. 38, Art. no. 530, Jun. 2026. https://doi.org/10.1007/s00521-026-12266-x.

[6] Y. Gan, et al., "Navigating the risks: A survey of security and privacy threats in LLM-based agents", ACM Transactions on Software Engineering and Methodology, Art. no. 3807666, Mar. 2026. https://doi.org/10.1145/3807666.

[7] X. Li, et al., "Prompt injection attacks on large language models: A survey of attack methods, root causes, and defense strategies", Computers, Materials & Continua, Vol. 87, No. 1, pp. 1-39, Feb. 2026. https://doi.org/10.32604/cmc.2025.074081.

[8] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning", ACM Computing Surveys, Vol. 54, No. 6, Art. no. 115, pp. 1-35, Jul. 2021. https://doi.org/10.1145/3457607.

[9] S. Choenni, M. S. Bargh, T. Busker, and N. Netten, "Data governance in smart cities: Challenges and solution directions", Journal of Smart Cities and Society, Vol. 1, No. 1, pp. 31-51, Feb. 2022. https://doi.org/10.3233/SCS-210119.

[10] Z. He, J. Khazaei, F. Moazeni, and J. D. Freihaut, "Detection of false data injection attacks leading to line congestions using neural networks", Sustainable Cities and Society, Vol. 82, Art. no. 103861, Jul. 2022. https://doi.org/10.1016/j.scs.2022.103861.

[11] S. M. Kim, H. S. Jung, and Y. W. Lee, "Smart City Cyber Security Based on Information Security Industry", Journal of KIIT, Vol. 18, No. 4, pp. 129-136, Apr. 2020. https://doi.org/10.14801/jkiit.2020.18.4.129.

[12] Ł. Korycki and B. Krawczyk, "Adversarial Concept Drift Detection under Poisoning Attacks for Robust Data Stream Mining", Machine Learning, Vol. 112, No. 10, pp. 4013-4048, Oct. 2023. https://doi.org/10.1007/s10994-022-06177-w.

[13] F. Yan, S. Wen, S. Nepal, C. Paris, and Y. Xiang, "Explainable Machine Learning in Cybersecurity: A Survey", International Journal of Intelligent Systems, Vol. 37, No. 12, pp. 12305-12334, Dec. 2022. https://doi.org/10.1002/int.23088.

[14] S. N. M. Moghaddam and H. Cao, "The metrics trap: How technical sophistication masks social harm in urban AI systems", npj Urban Sustainability, Art. no. 94, Apr. 2026. https://doi.org/10.1038/s42949-026-00394-1.

[15] S. Gupta, J. Singh, R. Agrawal, and U. Batra, "Explainable AI and machine learning for robust cybersecurity in smart cities", Cyber Security and Applications, Art. no. 100104, Dec. 2025. https://doi.org/10.1016/j.csa.2025.100104.

[16] C. Wang, Y. Wang, and Y. Sun, "Artificial intelligence in sustainable governance of smart cities: A review of data and algorithmic governance challenges", Buildings, Vol. 16, No. 12, Art. no. 2363, Jun. 2026. https://doi.org/10.3390/buildings16122363.

[17] S. Y. Yoon and Y. W. Lee, "Cloud based Distributed Parallel Processing for Deep Learning based Real-time Person Detection in an AI Middleware for Smart City", Journal of KIIT, Vol. 20, No. 5, pp. 9-19, May 2022. https://doi.org/10.14801/jkiit.2022.20.5.9.

[18] M. Ziosi, et al., "Smart cities: Reviewing the debate about their ethical implications", AI & Society, Vol. 39, no. 3, pp. 1185-1200, Jun. 2024. https://doi.org/10.1007/s00146-022-01558-0.

[19] I. Alsmadi, et al., "Securing tool-using AI agents against injection and authority misuse", Computation, Vol. 14, No. 5, Art. no. 98, Apr. 2026. https://doi.org/10.3390/computation14050098.

[20] N. K. Corrêa, et al., "Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance", Patterns, Vol. 4, No. 10, Art. no. 100857, Oct. 2023. https://doi.org/10.1016/j.patter.2023.100857.

[21] Z. Wang, J. Ma, X. Wang, J. Hu, Z. Qin, and K. Ren, "Threats to training: A survey of poisoning attacks and defenses on machine learning systems", ACM Computing Surveys, Vol. 55, No. 7, Art. no. 134, Dec. 2022. https://doi.org/10.1145/3538707

[22] A. Poudel, et al., "Governance risks of AI reasoning in urban infrastructure through Delphi audit of human and large language model judgment", Discover Cities, Vol. 3, Art. no. 81, May 2026. https://doi.org/10.1007/s44327-026-00268-2.

[23] F. Abtahi, F. Seoane, I. Pau, and M. Vega-Barbas, "Data poisoning vulnerabilities across health care artificial intelligence architectures: Analytical security framework and defense strategies", Journal of Medical Internet Research, Vol. 28, Art. no. e87969, Jan. 2026. https://doi.org/10.2196/87969.

[24] A. Oprea, A. Singhal, and A. Vassilev, "Poisoning attacks against machine learning: Can machine learning be trustworthy?", Computer, Vol. 55, No. 11, pp. 94-99, Nov. 2022. https://doi.org/10.1109/MC.2022.3190787.

[25] Y. Zhu, H. Wen, J. Wu, and R. Zhao, "Online data poisoning attack against edge AI paradigm for IoT-enabled smart city", Mathematical Biosciences and Engineering, Vol. 20, No. 10, pp. 17726-17746, Sep. 2023. https://doi.org/10.3934/mbe.2023788.

[26] E. Ismagilova, L. Hughes, N. P. Rana, and Y. K. Dwivedi, "Smart Cities: Advances in Research—An Information Systems Perspective", International Journal of Information Management, Vol. 47, pp. 88-100, Aug. 2019. https://doi.org/10.1016/j.ijinfomgt.2019.01.004.

[27] D. S. Lee, C. U. Jeon, J. H. Park, and S. H. Ok, "Context-Aware Security Robot System for Dangerous Person Discrimination with Digital Twin Integration", Journal of KIIT, Vol. 24, No. 3, pp. 69-79, Mar. 2026. https://doi.org/10.14801/jkiit.2026.24.3.69.

저자소개

백 승 수 (SeungSoo Baek)



2002년 2월 : 육군사관학교 전산학(학사)
2007년 9월 : Naval Postgraduate School 전산학(석사)
2018년 8월 : 고려대학교 정보보호(박사)
2025년 3월 ~ 현재 : ICT폴리텍대학 정보보안학과 조교수
관심분야 : AI보안, 네트워크보안, 스마트시티, 보안SW개발

윤 승 정 (SeongJeong Yoon)



2000년 8월 : 건국대학교 컴퓨터공학과(공학사)
2010년 2월 : 건국대학교 정보통신(공학석사)
2014년 2월 : 경희대학교 MIS(경영학박사)
2026년 8월 : Swiss School of Management, AI&Bingdata, DBA
2000년 7월 ~ 2003년 8월 : 중소벤처진흥공단 전산팀 과장대리
2005년 9월 ~ 2009년 7월 : 프리랜서 소프트웨어 개발자
2017년 5월 ~ 2018년 10월 : 피앤티컨설팅 이사
2018년 7월 ~ 2022년 1월 : 글로벌오픈파트너스, 수석컨설턴트
2022년 6월 ~ 2024년 10월 : 슈타겐 수석컨설턴트
2025년 3월 ~ 현재 : ICT폴리텍대학 AI소프트웨어학과 교수
2025년 9월 ~ 현재 : Swiss School of Managment DBA
관심분야 : 기술수용의도, 스마트시티, 소프트웨어 개발, AI 알고리즘 개발, 빅데이터 분석, 사물인터넷 계측