

인공지능을 이용한 산업제어시스템 운영 데이터 기반 이상탐지 방안 연구

21.10.16.(토)

전남대학교
정보보안협동과정
김가영
rkdud8727@gmail.com

 전남대학교

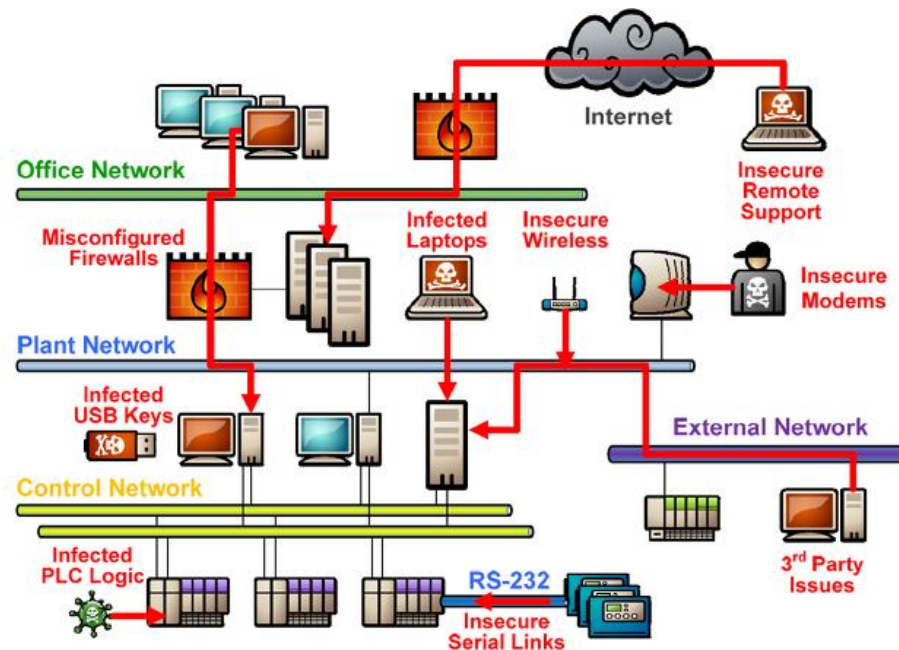
 **System Security
Research Center**

목 차

- I. 연구 목적 및 필요성
- II. 기존 연구 배경
- III. 기존 연구 분석
- IV. 연구 내용
- V. 결론 및 향후 일정

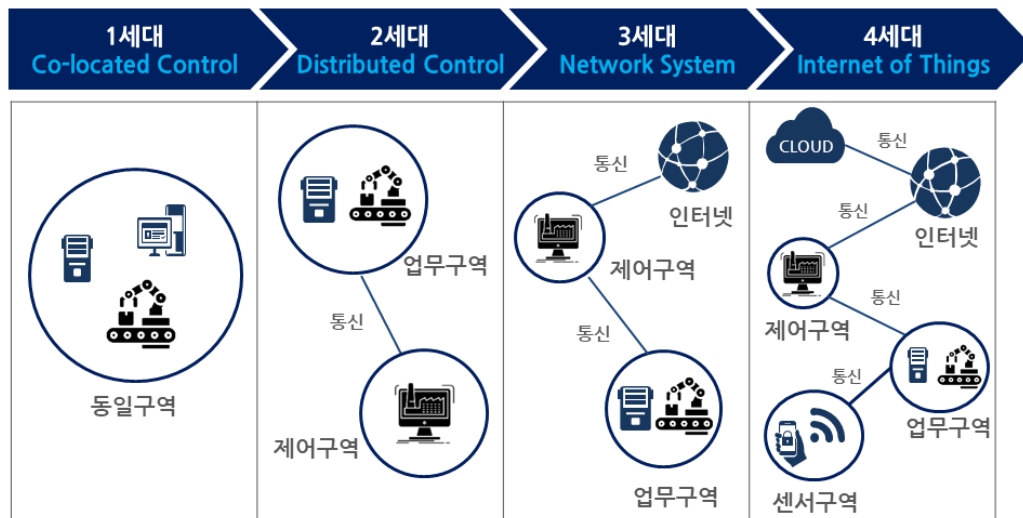
연구 목적 및 필요성

- 산업제어시스템 환경의 변화: 폐쇄망 → 외부 네트워크 연결
- 산업용 IoT(IIoT)의 발전
- 사이버 공격 사례가 급증하고 그에 따라 보안성이 중요해짐
- 따라서, 머신 러닝 기술을 이용한 이상 탐지 방법이 등장

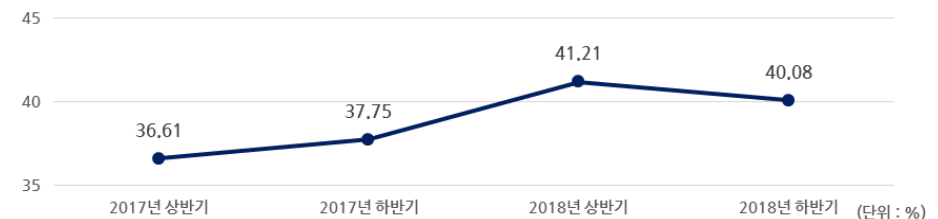


기존 연구 배경 (1/4)

- 산업제어시스템 대상 공격 증가
 - 기존의 OT와 IT의 결합으로 인해 새로운 보안위협이 발생
 - IT망을 통해 내부망을 침입하는 고도화된 공격 사례가 증가
 - 따라서, 산업제어시스템에 인공지능 보안을 적용하는 사례가 늘어나고 있음



[산업제어시스템의 변화과정]



[ICS에서 발생한 사이버 공격 발생 추이 (출처: Kaspersky)]

기존 연구 배경 (2/4)

- 산업제어시스템 데이터
 - 네트워크 트래픽을 수집한 데이터, 테스트 베드에서 수집한 운영 데이터로 구분
 - 실제 산업제어시스템 환경을 구성하기 어렵기 때문에 산업제어시스템 데이터 자체가 부족
 - ICS에서의 이상 탐지를 위해서는 물리적 프로세스 측정 값이 포함되어 있는 데이터가 적합
- 산업제어시스템 네트워크 트래픽 데이터
 - KDD Cup 1999 Data
 - ISCX(Intrusion Detection Evaluation Dataset)
- 산업제어시스템 운영 데이터
 - ICS Cyber Attack Datasets
 - BATADAL(The BATtle of Attack Detection Algorithms)
 - WADI(Water Distribution Testbed)
 - SWaT(Secure Water Treatment) Datasets
 - HAI(HIL-based Augmented ICS Security) Datasets

기존 연구 배경 (3/4)

- SWaT
 - 수처리 시스템에 대한 테스트베드를 구성하고 다양한 시나리오 기반의 공격을 수행
 - 일반 도시의 수처리 시스템의 각 단계를 묘사하고 있으며, 41가지의 공격을 포함
 - 현재까지 7차례의 데이터 셋 리뉴얼을 거쳐 8가지의 버전이 존재
 - 파일의 형태: csv, pcap

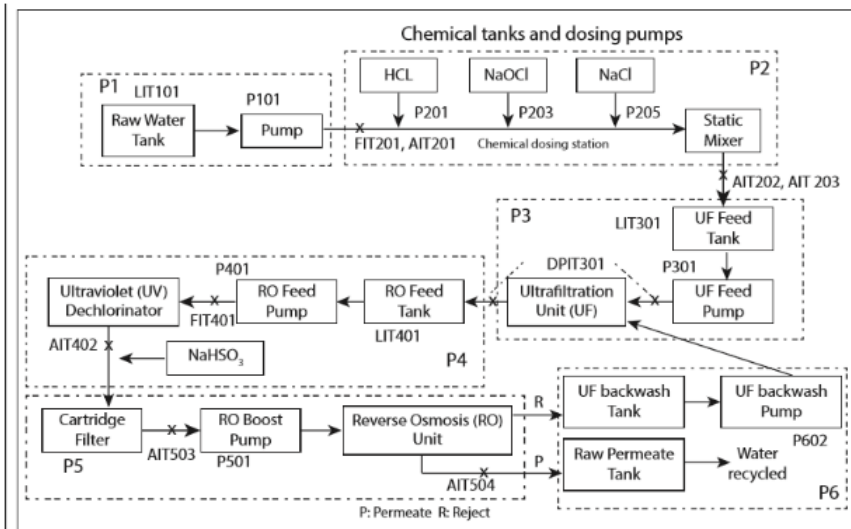


Fig. 2: SWaT testbed processes overview

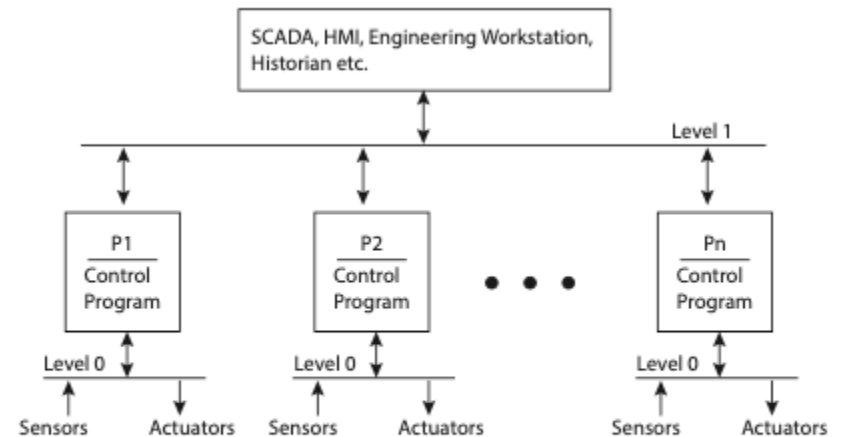


Fig. 3: SWaT testbed processes overview

기존 연구 배경 (4/4)

- HAI
 - dSPACE HIL(hardware-in-the-loop) 시뮬레이터 + 세 가지의 물리적 제어시스템 (GE 터빈, Emerson 보일러, FESTO 수처리 시스템)
 - 제어시스템에 의해 측정되는 변수와 제어되는 변수를 나타내는 78개의 지점에서 매 초마다 샘플링
 - 주요 버전으로 초기 버전인 HAI 20.07과 개선된 HAI 21.03 버전이 존재 (형태: csv)
 - 21.03 버전은 데이터 포인트가 59개에서 78개로 증가, 11가지의 공격 시나리오 추가

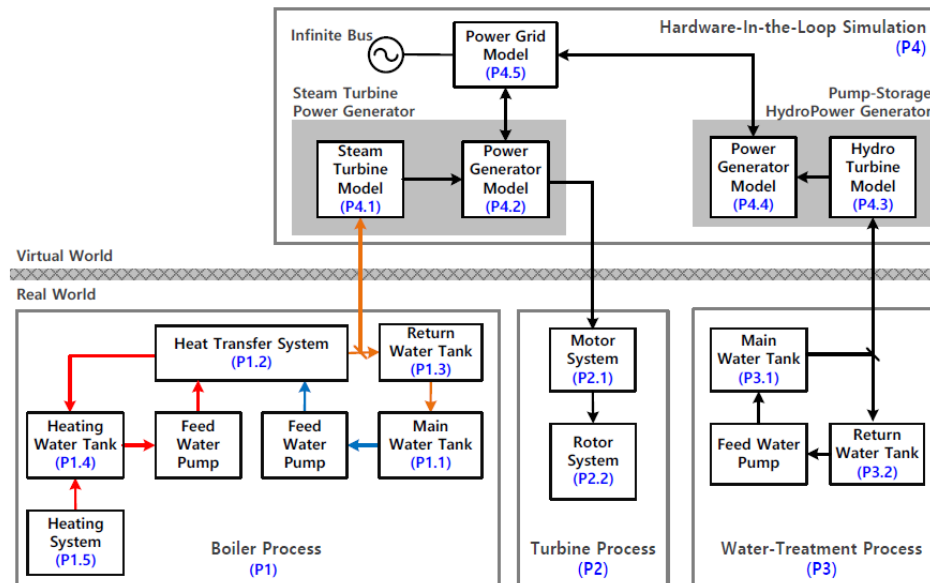


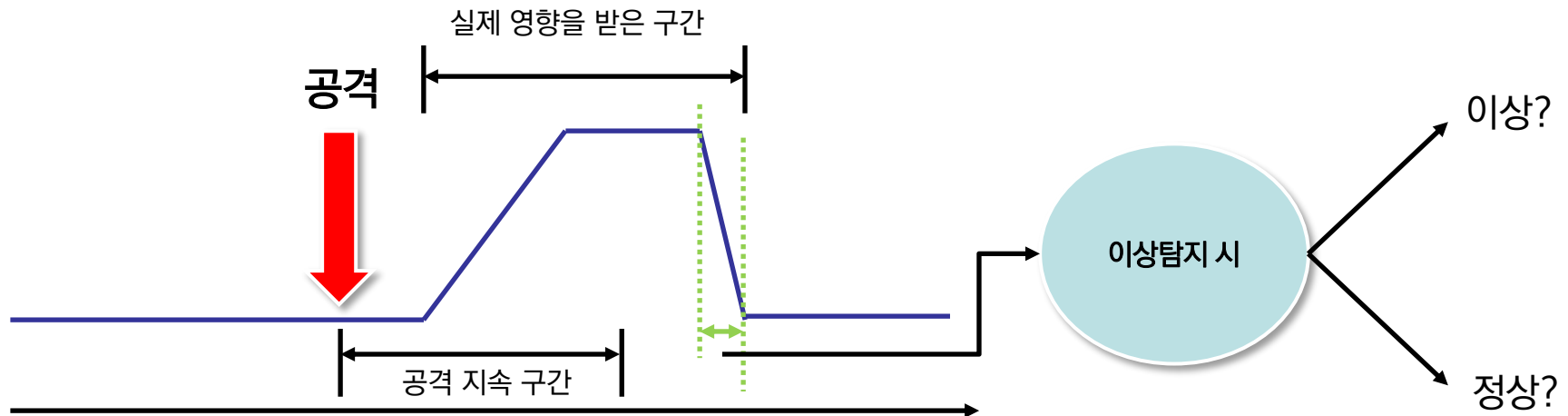
FIGURE 1. PROCESS FLOW DIAGRAM.

기존 연구 분석 (1/4)

No.	연도	사용 모델	데이터 셋	평가 지표
1	2018	LSTM/GRU	SWaT	NAB(the Numenta Anomaly Benchmark)
2	2019	Seq-to-Seq	SWaT	False Negative, False Positive 분석
3	2019	SVM, RF, KNN	SWaT	Accuracy, F1-score
4	2021	KNN, RF, DT	HAI	Precision, Recall, Accuracy, AUC
5	2021	SAE, SVDD	HAI	Accuracy

기존 연구 분석 (2/4)

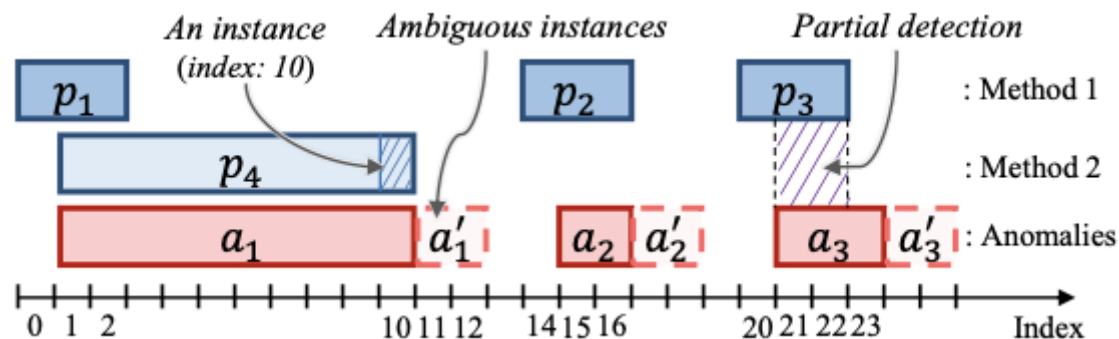
- ‘시계열’ 데이터를 고려하지 않는 평가 지표
 - 5가지의 관련 연구는 SWaT과 HAI 데이터 셋을 사용
 - SWaT과 HAI 데이터 셋은 초 단위로 구성된 데이터로 ‘시계열’ 데이터 특성을 가지고 있음
 - 시계열 데이터의 이상 탐지에 대한 평가 지표로 이진 분류(이상, 정상)을 나타내는 것은 한계점이 존재
 - 또한, 사전 징후나 공격으로 인해 제어시스템을 영향을 받은 후 정상으로 돌아가는 과정을 이진 분류로 나타내는 것은 힘들



기존 연구 분석 (3/4)

- TaPR(Time-series aware Precision and Recall)
 - Won-Seok Hwang, et al. “Time-Series Aware Precision and Recall for Anomaly Detection”, 2019
- 기존 평가 방법의 문제점
 - 탐지한 이상 항목 수와 정확하게 탐지되었는지를 고려하지 않음
 - 긴 이상만 탐지하는 방법에 높은 점수를 부여
 - 모호한 인스턴스(ambiguous instances)의 부재
 - 모호한 인스턴스란 기계의 오작동/사이버 공격 등에 의한 이상으로 영향을 받았지만 ‘정상’으로 표시된 인스턴스
- 따라서, 개선된 지표를 제시하여 탐지된 이상 항목의 다양성이 중요하다고 강조
 - 탐지 점수 (탐지된 이상 항목 수)
 - 부분 점수 (각 이상 항목이 얼마나 정확하게 탐지되는지)

기존 연구 분석 (4/4)



- 공격은 a_1, a_2, a_3 으로 이루어지며, a'_1, a'_2, a'_3 은 호모한 인스턴스
- 방법1과 방법2는 이상 탐지 결과를 보여줌
 - 방법 1은 a_1, a_2, a_3 일부를 모두 다 탐지
 - 방법 2는 a_1 만 정확하게 탐지
- 기존 평가 방법(정밀도, 재현율) 점수는 방법 2가 방법 1보다 높은 점수를 받음
 - 공격에 의한 이상 발생이라고 한다면 방법 2가 놓친 두 가지 사이버 공격(a_2, a_3)에 의해 시스템이 심각하게 손상될 수 있음

연구 내용 (1/4)

- 이전 연구 내용

- 시계열 데이터의 이상 탐지 평가 방법 연구

- 기존의 시계열 데이터 이상 탐지 평가 방법에는 Range-based Precision & Recall, TaPR이 존재
 - 두 가지의 평가 방법은 기존의 정확도 평가에 대한 문제점을 지적하며 ‘범위 기반’ 평가를 연구
 - Range-based Precision & Recall은 Recall 점수는 시계열 데이터에 적용하는게 효과가 없다는 것을 입증하며, 각 이상에 대한 중요도를 나타내는 함수를 도입하여 보상 점수 조정

- TaPR의 모호한 인스턴스의 범위 확장

- TaPR은 RR과 RP에 모호한 인스턴스를 더한 평가 방법
 - TaPR에서 말하는 모호한 인스턴스 추가 보상 기법은 이상 발생 후에만 해당
 - ‘사전 징후’를 탐지하는 모호한 인스턴스에도 보상을 적용하여 평가 방법을 발전

- 한계점

- 모호한 인스턴스의 범위를 늘리면 추가 점수에 대한 범위가 일정하지 않음
 - 공격이 연속으로 발생하는 경우에는 사전 징후와 이상탐지가 겹쳐 중복되는 경우가 발생
 - 실제 알고리즘 수정 후 적용했을 경우 점수가 적용되지 않음

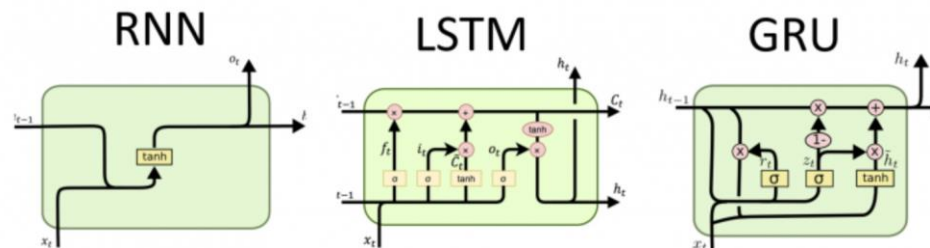
연구 내용 (2/4)

- 산업제어시스템 데이터 선택
 - 네트워크 트래픽 데이터 VS. 물리적인 운영 데이터
 - ICS에 특화된 이상 탐지를 실행하기 위해서는 데이터 셋이 OT 환경의 다양한 장치 간의 통신을 나타내는 물리적 프로세스 측정 값이 필요
 - 따라서, 물리적인 운영 데이터를 선택
 - SWaT VS. HAI
 - 산업제어시스템 이상 탐지에 대한 연구에 사용되는 두 가지 물리적인 운영 데이터
 - SWaT과 HAI를 비교 시, 모든 구분 항목에서 HAI가 더 좋은 데이터로 판단

구분	SWaT	HAI
공통점	수처리 공정 시스템	
데이터 포인트	51	78
공격 포인트	41	50
이상 탐지 분류	이상 / 정상	이상(P1, P2, P3) / 정상

연구 내용 (3/4)

- 이상 탐지에 사용할 딥러닝 모델 선택
 - RNN
 - 시간순으로 여러 개의 입력 값을 현재의 입력으로 받아서 계산
 - 하지만, RNN의 구조상 학습 데이터의 길이가 길면 먼 과거의 정보를 현재에 전달하기 힘들기 소실 문제가 발생 (Long-Term Dependency)
 - LSTM
 - RNN의 기울기 소실 문제를 해결하는 모델
 - LSTM은 신경망의 중간 계층에서의 각 유닛을 LSTM 블록이라는 메모리 유닛으로 치환한 구조
 - GRU
 - LSTM을 구성하는 셀을 간소화한 버전
 - 구조가 LSTM에 비해 단순하고 학습시킬 파라미터가 적어 연산 속도가 빠름



연구 내용 (4/4)

사용한 모델	경계값	성능 지표		
		F1-score	TaP	TaR
RNN	0.43	78.02	97.20	64.40
LSTM	0.43	81.60	97.40	67.66
GRU	0.43	83.10	97.50	71.50

- 결론 도출

- 문제점

- 시계열 데이터를 대상으로 한 이상 탐지 연구에서 대부분 표준 분류 메트릭을 사용하여 점수 도출
 - 표준 분류 메트릭, TaPR에 대한

- 평가 지표 비교를 통한 분석

- 전체 데이터에서 이상이 발생한 구간을 설정 (사례 구분)
 - 이상 탐지 후 평가 지표를 통한 점수 도출 (표준 메트릭, TaPR)
 - 설정한 구간을 중심으로 점수 비교를 통해 분석

결론 및 향후 일정

- 결론

- 기존 이상 탐지 방법의 평가 지표는 ‘시계열’ 특성을 고려하지 않음
 - 정확도, 정밀도, 재현율, ROC curve, AUC를 사용
 - 위의 평가 지표는 단순히 긴 이상을 탐지하면 높은 점수를 부여
 - 시계열 특성을 고려한 평가 지표 TaPR가 연구됨
 - 산업제어시스템 운영 데이터에서의 이상탐지 후, 시계열 데이터의 특성을 고려한 평가 지표를 이용

- 향후 일정

- 최적의 딥러닝 모델을 선정 후 이상 탐지 + 여러 평가 지표를 통해 비교
 - 시계열 데이터 이상을 감지하는 모델 LSTM, RNN, GRU 선정
 - HAI 21.03 데이터를 이용하여 학습
 - 이상 탐지 판단을 위한 ‘경계값’ 설정에서 평가 지표를 활용
 - 경계값 설정 후 테스트 데이터 셋을 대상으로 이상 탐지 실행
 - 여러 평가지표를 통해 비교 후 결론 도출

연구 실적

No.	컨퍼런스/ 저널	학회명	주관	제목	저자
1	컨퍼런스	한국정보보호학회 호남지부 추계학술대회	한국정보보호학회	인공지능을 악용한 적대적 공격의 분석과 대응방안	주저자
2	컨퍼런스	한국통신학회 통신망 운용관리 학술대회	한국통신학회	자연어 처리를 이용한 네트워크 트래픽 이상 탐지 기법	주저자
3	컨퍼런스	한국융합보안학회 하계학술대회	한국융합보안학회	개인정보 생애주기를 고려한 인공지능시스템 프라이버시 위험 대응 프레임워크	주저자
4	저널 (교내학술지)	시스템보안 연구지	전남대학교 시스템보안연구센터	인공지능을 악용한 적대적 공격의 분석과 대응방안	주저자
5	저널 (KCI)	정보화연구	한국EA학회	산업제어시스템에 대한 고도화된 Kill Chain 공격 기법 연구	공저자
6	저널 (KCI)	정보화연구	한국EA학회	개인정보 생애주기를 고려한 인공지능시스템 프라이버시 노출 대응 프레임워크	주저자

감사합니다

