Journal of KIIT. Vol. 23, No. 4, pp. 179-190, Apr. 30, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.4.179

Asherah Simulator를 활용한 원자력 시설 사이버 공격 시나리오 시뮬레이션 및 이상 탐지 방법

김성아*[1], 한준서*[2], 최선오**

Simulation of Cyberattack Scenarios on Nuclear Facilities using the Asherah Simulator and A Method for Anomaly Detection

Seongah Kim*[1], Junseo Han*[2], and Sunoh Choi**

요 약

원자력발전소는 물리적 공정과 산업제어시스템이 결합된 고위험 인프라로서, 사이버 공격에 취약하며 이에 대한 실시간 이상 탐지가 중요하다. 본 연구에서는 Asherah Simulator와 OPC-UA Simulation Server 및 Wireshark를 활용하여 원자력발전소의 주요 공정을 모사하고, FW Pump와 SG Level을 대상으로 사이버 공격 시나리오를 설계하였다. Normal Data와 Attack Data를 전처리하여 Dataset을 구축하였고, AI-Based Anomaly Detection Model(TranAD, Bi-LSTM, One-class SVM)에 Normal Data를 학습시키고 Attack Data로 이상 탐지 테스트를 수행하였다. 세 가지 모델의 실험 결과로 성능 Matrix를 평가하였고, 그 중 TranAD 모델이 FW Pump SpeedCmd Attack에서 F1-Score(0.99), FPR(0.005)를, SG Level Setpoint Attack에서 F1-Score(0.99), FPR(0.000)로 가장 우수한 탐지 성능을 보였다.

Abstract

A nuclear power plant, as a high-risk infrastructure combining physical processes and industrial control systems, is vulnerable to Cyberattacks, making real-time anomaly detection crucial. This study simulates key processes using Asherah Simulator, OPC-UA Simulation Server, and Wireshark, designing attack scenarios for the FW Pump and SG Level. After preprocessing Normal and Attack Data, an AI-Based Anomaly Detection Model (TranAD, Bi-LSTM, One-class SVM) was trained and tested. Among the models, TranAD showed the best performance, achieving an F1-Score of 0.99 and FPR of 0.005 in the FW Pump SpeedCmd Attack, and an F1-Score of 0.99 and FPR of 0.000 in the SG Level Setpoint Attack.

Keywords

asherah simulator, OPC-UA simulation server, nuclear simulation, anomaly detection

* 전북대학교 소프트웨어공학과 석사과정
- ORCID[1]: https://orcid.org/0009-0000-1733-5315
- ORCID[2]: https://orcid.org/0009-0007-7529-812X
** 전북대학교 소프트웨어공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-0654-7109

· Received: Feb. 04, 2025, Revised: Feb. 27, 2025, Accepted: Mar. 02, 2025

· Corresponding Author: Sunoh Choi
Dept. of Software Engineering Jeonbuk National University Korea
Tel.: +82-63-270-4784, Email: suno7@jbnu.ac.kr

Ⅰ. 서 론

원자력 발전을 기반으로 하는 원자력 시설(NPP, Nuclear Power Plant)은 이산화탄소 배출을 감소하며, 값싼 전기를 생성할 수 있어서 전기 생산의 주요 원천으로 자리 잡고 있다. 이러한 원자력 시설은 물리적 공정과 이를 제어 및 모니터링 하는 산업제어시스템(ICS, Industrial Control System)이 연결되어 있다. ICS의 내부 구성요소는 크게 센서 데이터 수집(PLC, Programmable Logic Controller), 데이터 처리(SCADA, Supervisory Control and 공격 Acquisition / DCS, Distributed Control System), 표준화된 데이터 교환(OPC-UA, Open Platform Communications Unified Architecture)[1], 시각적 표시 역할(HMI, Human-Machine Interface)등으로 이루어져 있다[2]. 이러한 ICS는 가압기(Pressurizer), 주급수 펌프(Feedwater pump), 증기 발생기(Steam generator) 등의 공정들의 시스템 간 상호운용성을 향상시키며 공정을 제어하는 역할을 한다.

ICS의 데이터 교환 과정은 물리적 공정 → Sensor → PLC → SCADA/DCS → OPC-UA → HMI로 전달되며, 필요시 제어 명령이 역방향으로 전달된다. 데이터 교환 과정에서 데이터를 표준화된 방식으로 처리하는 것이 OPC-UA이며, 네트워크를 통해 실시간 데이터 교환이 가능하다. 데이터에 대한 외부 접근이 가능하므로, 사이버 공격 위협에 노출될 가능성이 높다. 대표적인 사건으로 ICS를 표적으로 삼는 웜 바이러스 스턱스넷(Stuxnet)[3]이 있다. 스턱스넷은 지멘스의 PLC를 감염시켜 이란 나탄즈 원자력 시설의 원심분리기 회전 속도를 비정상적으로 변화시키고, 장비를 물리적으로 손상시키도록 설계된 바이러스이다.

2010년 발생한 스턱스넷 사건은 ICS의 사이버 공격 취약성을 입증한 대표적인 사례로, 국가 간 사이버 전쟁의 가능성과 고위험 인프라의 중요 공정 보호 필요성을 부각시켰다. 원자력 시설은 물리적 공정과 ICS가 연결된 고위험 인프라이므로, 시스템 오류나 악의적 사이버 공격 발생 시 방사능 누출 등 치명적 결과를 초래할 수 있다. 특히, 주요 공정에 대한 사이버 공격은 물리적 오작동과 시설 파괴를 유발할 가능성이 높아, 실시간 이상 탐지를 통한 조기 대응 및 능동적 보안 체계 구축이 필수적이다. 그러나 실제 원자력 시설은 공정의 복잡성과 안정이 직결되기 때문에 직접적인 공격 실험을 실행하기 어려워서, 잠재적 공격 경로와 이상을 탐지하고 검증하는데 한계가 존재한다. 이러한 한계를 극복하고자 원자력 시뮬레이션을 환경을 활용하여 직접적인 공격 실험을 수행하여 원자로 핵심 공정 변수 등 잠재적 공격 경로를 파악하고, 이상 탐지 방법을 체계적으로 연구할 필요가 있다.

본 연구에서는 원자력 시뮬레이션 환경으로 Asherah Simulator[4]를 활용하였으며, Asherah Simulator는 IAEA CRP J02008 프로젝트에서 개발된 Matlab/Simulink 기반의 가압수형 원자로(PWR) 시뮬레이터이다. 해당 시뮬레이터는 원자로의 1차 계통과 2차 계통을 포함하며, 실제 원자력발전소 환경을 모사하여 OPC-UA Simulation Server를 통해 네트워크 연결이 가능하다. 이를 활용하여 원자로 핵심 공정 변수의 변화를 감지하고, Attack Scenario 적용 시 나타나는 시스템 반응을 분석할 수 있다.

본 연구는 Asherah Simulator와 OPC-UA Simulation Server, Wireshark를 활용하여 핵심 공정 변수를 기반으로 Attack Scenario를 설계한 후 AI-Based Anomaly Detection Model로 이상 탐지를 하여 실험을 설계하였다. Asherah Simulator와 OPC-UA Simulation Server가 데이터를 교환 할 때, Wireshark를 통해 핵심 공정 변수의 패킷을 캡처하였다. 캡처 된 패킷을 기반으로 Attack Scenario(FW Pump SpeedCmd Attack, SG Level Setpoint Attack)를 설계하고 설계된 Attack Scenario를 기반으로 공격하였다. 그 후 AI-Based Anomaly Detection Model인 O-C SVM(One-Class Support Vector Machine)[5], Bi-LSTM(Bidirectional Long Short-Term Memory)[6], TranAD(Transformer-based Anomaly Detection)[7]를 통해 정상 데이터셋을 학습시키고 공격 데이터셋으로 이상 탐지를 수행하였다. 세 가지 모델 동일한 방식으로 학습한 후 이상 탐지 테스트를 수행하였고, 실험 결과로 성능 Matrix를 평가하였다.

세 가지 모델 중 TranAD가 FW Pump SpeedCmd Attack에서 F1-Score(0.99), FPR(0.005)를, SG Level Setpoint Attack에서 F1-Score(0.99), FPR(0.000)를 기록하여 가장 높은 성능으로 평가되었다.

이러한 실험 결과로 인해 TranAD가 가장 적합한 모델임을 확인하였고, 고위험 인프라의 사이버 보안을 강화하기 위한 실증적인 접근법을 제시하였다. 이 논문은 우리가 아는 한 원자력발전 시뮬레이터에서 이상 탐지를 수행한 첫 번째 연구이다.

II. 배경지식

2.1 원자력 시설

원자력 시설(NPP)은 원자력 발전으로 전기 에너지를 생성한다. 원자력 발전은 원자로의 핵 분열 반응에서 방출되는 열 에너지를 활용해 고온·고압의 증기를 생성하고, 이 증기가 터빈을 회전시켜 기계적 에너지를 전기에너지로 변환하는 방식으로 작동한다. 생성된 증기는 열 교환 및 냉각 시스템을 통해 다시 물로 환원되어 순환 과정을 반복하며, 이러한 열역학적 사이클을 통해 지속적으로 전력을 생산하는 방식으로 운영된다.

이러한 원자력 시설은 공정의 복잡성과 안정성이 직결되기 때문에, 운영 중인 원자력 시설에 직접적인 테스트를 수행하는 것은 큰 제약이 따른다. 이를 보완하기 위해 NPP의 운영 환경을 모사한 Testbed가 활용된다. 원자력 시설의 Testbed는 실제 발전소에서 사용되는 시스템의 동작을 재현하며, 새로운 기술 도입 전 정비 및 보수 훈련, 데이터 수집 및 분석, 그리고 사이버 보안 위협 시뮬레이션 등을 수행하기 위한 환경을 제공한다. 이러한 Testbed에서 검증된 결과를 바탕으로 실제 원자력 발전소 운영에 반영함으로써 시스템의 안정성과 신뢰성을 높일 수 있다.

2.2 Asherah NPP Simulator

2.2.1 Asherah NPP Simulator 소개

Asherah Simulator는 IAEA CRP J02008 프로젝트에서 USP 연구팀에서 개발한 NPP 시뮬레이터이다. 이 NPP Simulator Model은 Matlab/Simulink[8]를 사용하여 구현된 가압수형 원자로(PWR)에 기반을 두고 있다. 이 시뮬레이터는 2,772 MWt PWR Babcock & Wilcox (B&W) 원자로를 중심으로 설계되어 있으며, OPC-UA Simulation Server로 네트워크 연결이 가능하다.

2.2.2 Asherah NPP Simulator 구성

Asherah NPP Simulator는 그림 1과 같이 LOCAL HMI와 ASHERAH NPP로 구성되어 있다. LOCAL HMI는 Reactor Protect System의 제어, Reactor 조정, 그리고 시스템 주요 변수 값을 실시간으로 모니터링할 수 있는 인터페이스를 제공한다.

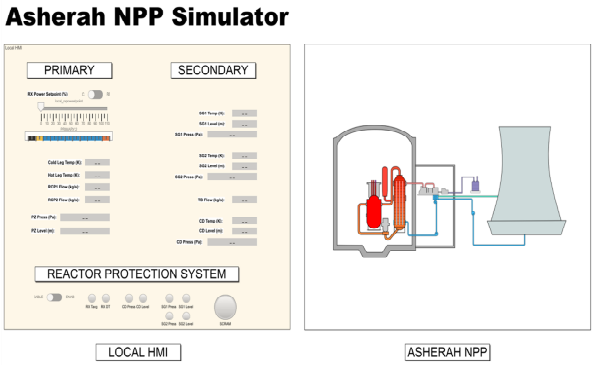

그림 1. 아슈라 원자력 시설 시뮬레이터
Fig. 1. Asherah NPP simulator

ASHERAH NPP는 Primary Loop(1차 계통)과 Secondary Loop(2차 계통)으로 구성되며, 그림 2와 같이 1차 계통은 원자로 내부에서 열 에너지를 생성하여 2차 계통으로 전달하는 역할을 수행한다.

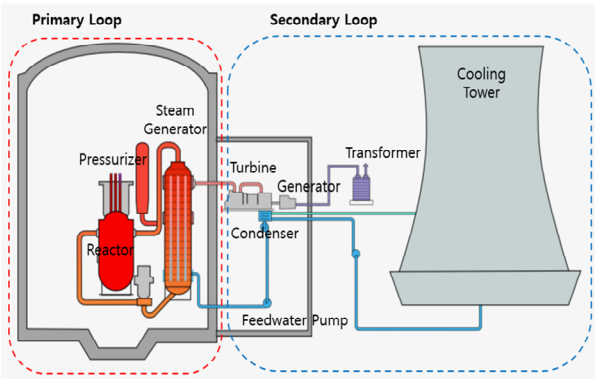

그림 2. 아슈라 시뮬레이터 1·2차 계통
Fig. 2. Asherah simulator primary and secondary loops

1차 계통의 주요 구성 요소에는 열 발생을 담당하는 Reactor, 냉각재 순환 및 압력을 조절하는 Pressurizer, 열을 2차 계통으로 전달하는 Steam Generator가 포함된다. 2차 계통은 1차 계통에서 전달받은 열에너지를 활용하여 증기를 생성하고, 이를 통해 전기 에너지를 생산한다. 2차 계통의 주요 구성 요소에는 증기를 사용하는 Turbine, 전력을 생성하는 Generator, 증기를 응축하는 Condenser, 그리고 응축수를 순환시키는 Feedwater Pump가 포함된다.

2.3 OPC-UA

OPC UA(Open Platform Communications Unified Architecture)는 산업 자동화 분야에서 기기, 시스템, 소프트웨어 간의 안전하고 표준화된 데이터 교환을 가능하게 하는 통신 프로토콜이다. Asherah NPP는 OPC UA를 통해 특정 값을 주고받을 수 있는 기능을 탑재하고 있다. 그림 3과 같이 작동하며, TCP/IP 통신을 이용해서 시뮬레이터와 값을 주고받는다.

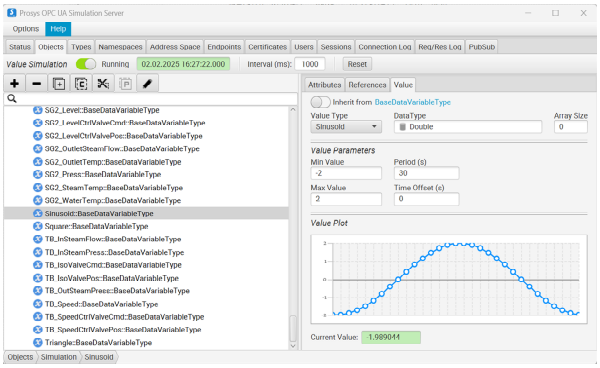

그림 3. OPC-UA 사용 예시
Fig. 3. Examples of OPC-UA usage

III. 관련 연구

3.1 설명 가능한 AI 기반 이상 탐지 연구(CPS 환경)

K. K. Kim et al.[9]은 사이버-물리 시스템(CPS) 환경에서 설명 가능한 AI 기반 이상 탐지 시스템을 개발하는 연구를 수행하였다. 원자력 발전소와 같은 산업 제어 시스템(ICS)에서 AI 기반 자동화 시스템이 확대되면서, 동시에 사이버 위협도 증가하고 있다. 특히, 기존의 AI 기반 이상 탐지 시스템은 블랙박스 모델이 많아 결과를 해석하기 어렵고, Attacker가 이를 우회하거나 조작할 가능성이 높다.

이 연구에서는 EKF(Extended Kalman Filter)를 적용하여 CPS에서 발생하는 노이즈를 보정하고, LIME과 SHAP 기법을 활용하여 탐지 결과의 해석 가능성을 높이는 방법을 제안하였다. 연구의 실효성을 검증하기 위해 HIL 기반 산업 제어 시스템 데이터셋(HAI 23.05)을 활용하여 실험을 수행했으며, EKF 기반 이상 탐지 시스템이 기존 방법 대비 사이버 공격 탐지 성능을 향상시키고 AI 모델의 설명 가능성을 높일 수 있음을 입증하였다.

3.2 PCTRAN과 AI 기반 보안 시스템을 활용한 악성코드 분석연구(NPP 사이버 보안 강화)

K. Marsh et al.[10]은 AI 기반의 호스트 침입 탐지 시스템(HIDS)과 보안 컨테이너 시스템을 PCTRAN 원자력 시뮬레이터에 통합하여 사이버 보안을 강화하는 연구를 수행하였다. PCTRAN은 원자력 발전소(NPP)의 사고 및 이상 상태를 시뮬레이션하는 프로그램으로, 실제 발전소 운영 환경을 재현할 수 있다. 하지만, 이 소프트웨어는 Windows XP 기반에서 실행되며, 최신 보안 패치를 적용하는 데 한계가 있었다.

이를 해결하기 위해 연구진은 AI 기반 이상 탐지 모델(One-Class SVM, Isolation Forest)을 활용한 HIDS를 개발하여 PCTRAN 시스템의 이상 행위를 감지하고, Docker 기반 보안 컨테이너 시스템을 구축하여 악성코드 분석과 격리를 수행하는 방안을 제안하였다. 연구 결과, HIDS가 실시간으로 이상 행위를 탐지할 수 있으며, 보안 컨테이너 시스템이 악성코드 확산을 방지하고 핵심 시스템을 보호하는 데 효과적임을 입증하였다.

3.3 NPP 시뮬레이터를 통한 사이버 보안 연구 및 교육 플랫폼 개발

M. T. Rowland et al.[11]은 국제 원자력 발전소(NPP)의 사이버 보안 연구 및 교육을 위해 ANSP(Asherah NPP Simulator Platform)을 개발하였다.

원자력 산업에서 디지털 기술이 점점 더 많이 사용되면서, 사이버 공격에 대한 대비가 필수적이 되었지만, 국제적으로 활용할 수 있는 저비용의 원자력 발전소 시뮬레이션 플랫폼이 부족한 상황이었다.

이에 대응하여 국제 원자력 기구(IAEA)와 협력하여 Asherah 가압수형 원자로 시뮬레이터를 개발하였으며, 해당 시뮬레이터는 실제 산업 장비와 인터페이스할 수 있는 하드웨어 인 더 루프(HIL) 기능을 포함하고 있다. 이 플랫폼은 원자력 사이버 보안 연구뿐만 아니라, 공학 및 보안 전문가를 위한 교육 목적으로 활용 가능하며, 실제 사이버 Attack Scenario를 테스트하고 보안 대응책을 검증하는 데 사용될 수 있다.

3.4 디지털 자산 특성 기반 보안 통제 프레임워크 연구

S. M. Lim[12]은 원자력 시설과 디지털 자산의 특성에 따라 차별화된 사이버 보안 통제 방안을 연구하였다. 기존의 보안 규정은 산업별 특성을 충분히 반영하지 못하고 일괄적으로 적용되어, 자원의 비효율적 배분과 보안 취약점을 초래할 가능성이 있었다. 이를 개선하기 위해 연구진은 원자력 시설 내 필수 디지털 자산(CDA)을 분류하고, 보안 조치를 차등 적용하는 방법론을 제안하였다. 특히, 미국 원자력 산업 협회(NEI)에서 개발한 NEI 13-10 분류법을 기반으로 CDA를 A1~B3 등급으로 구분하고, 각 등급에 맞는 보안 대책을 설계하였다. 또한, 미국 원자력 규제위원회(NRC)에서 제시한 DG-5062 방법론을 활용하여 디지털 자산의 위험 수준을 평가하고 보안 통제 방안을 최적화하는 방안을 제시하였다.

이 연구는 원자력 산업뿐만 아니라, 다른 산업에서도 보안 자원을 보다 효과적으로 배분하고 최적화된 보안 정책을 수립하는 데 기여할 수 있음을 시사하였다.

Ⅳ. 이상탐지방법

본 연구는 NPP의 ICS 사이버 보안 강화를 위해 Asherah Simulator, OPC-UA Simulation Server 및 Wireshark를 이용한 새로운 접근 방법을 제안한다. 제안된 방법은 그림 4와 같이 Attack Scenario와 Anomaly Detection의 두 단계로 구성되며, Attack Scenario는 Step1: Data Flow Definition과 Step2: Attack Scenario Simulation로 구성되고, Anomaly Detection은 Step3: AI-Based Anomaly Detection 으로 구성된다.

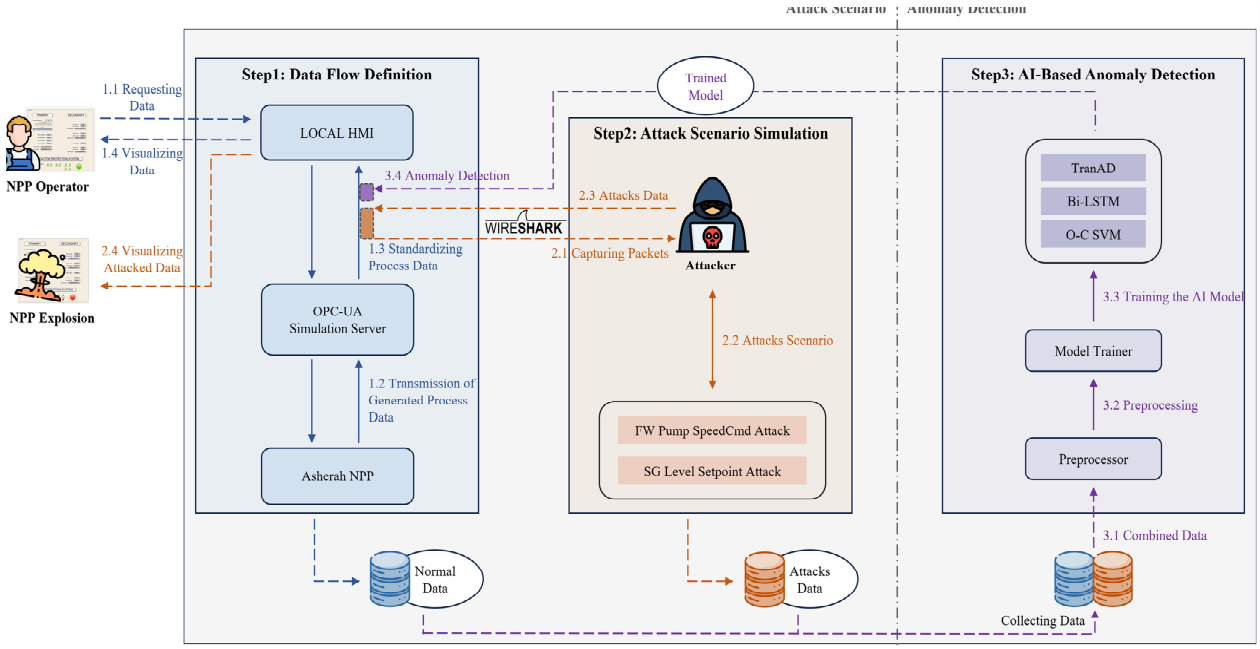

그림 4. 공격 시나리오 시뮬레이션 & 이상 탐지
Fig. 4. Attack scenario & anomaly detection

4.1 Attack Scenario

이 단계에서는 정상 데이터와 공격 데이터를 수집하고, 데이터셋을 구축하는 과정을 다룬다. Step 1에서는 정상 데이터를 수집하여 데이터셋을 구축하는 과정을, Step 2에서는 공격 데이터를 생성하고 수집하여 데이터셋을 구축하는 과정 다룬다. step1과 step2의 과정은 다음과 같다:

4.1.1 Step1: Data Flow Definition

Step1은 NPP에서 정상 데이터를 요청하고 수집한 후 데이터셋 구축 과정을 다룬다. NPP Operator가 LOCAL HMI를 통해 데이터를 요청하면, LOCAL HMI는 OPC-UA Simulation Server를 통해 Asherah NPP에 데이터를 요청한다. Asherah NPP는 요청된 데이터를 생성하고 수집하는데, 생성된 데이터는 OPC-UA Simulation Server에서 표준화 과정을 거친 후 LOCAL HMI에 제공된다. 이렇게 수집된 표준화된 정상 데이터로 데이터셋을 구축하고, 구축된 데이터셋은 AI-Based Anomaly Detection Model의 학습 데이터로 활용된다.

4.1.2 Step2: Attack Scenario Simulation

Step 2는 Wireshark를 이용하여 핵심 공정 변수의 패킷을 캡처 한 후, 캡처된 패킷을 분석하여 Attack Scenario를 설계하고 공격하는 과정을 다룬다. Attacker가 OPC-UA Simulation Server와 LOCAL HMI 간에 주고받는 핵심 공정 변수의 패킷을 캡처한다. 캡처된 패킷의 핵심 공정 변수를 분석하여 Attack Scenario를 설계한다. 설계된 Attack Scenario는 LOCAL HMI와 OPC-UA Simulation Server 간의 데이터 전송 과정에 개입하여 공격 데이터를 삽입한다. 삽입한 Attack Scenario로 인해 핵심 공정 변수의 변수 값이 변경되며, LOCAL HMI를 통해 공격된 데이터를 확인할 수 있다. 이렇게 수집된 공격 데이터로 데이터셋을 구축 하고, 구축된 데이터셋은 AI-Based Anomaly Detection Model의 이상 탐지를 수행하는데 활용된다.

4.2 Anomaly Detection

이 단계에서는 Step1, Step2에서 수집된 정상 데이터 데이터셋과 공격 데이터 데이터셋을 기반으로 AI-Based Anomaly Detection Model의 학습 및 이상 탐지를 수행하는 과정을 다룬다. step3의 과정은 다음과 같다:

Step 3는 수집된 정상 데이터 데이터셋과 공격 데이터 데이터셋을 전처리기를 통해 전처리하여 AI-Based Anomaly Detection Model의 학습 및 평가에 활용할 수 있는 형태로 변환한다. 수집된 정상 데이터 데이터셋은 AI-Based Model(TranAD, Bi-LSTM, O-C SVM)을 학습하는데 사용되고, 공격 데이터 데이터셋은 이상 탐지를 하는데 사용된다. 이상 탐지 수행 결과 각 모델의 성능 Matrix를 통해 최적의 Anomaly Detection Model을 도출한다.

V. 실험 설정

5.1 Normal Data

정상 데이터 수집은 5분간 원자력발전소 시뮬레이터(Asherah NPP)를 실행하여 0.1초 단위로 17개종류의 센서 데이터를 수집하였다. 비지도 학습을 위해 공격이나 이상 상태가 전혀 없는 정상 데이터만을 수집하였다.

5.2 Attack Scenario Data

Attack Scenario 데이터는 네트워크 트래픽 분석 도구인 Wireshark를 사용하여 17가지 변수 중 원자력 시설의 두 가지 주요 핵심 변수(FW Pump SpeedCmd, SG1 Level Seetpoint)를 캡처하였다. 캡처한 변수로 Attack Scenario를 설계하고 OPC-UA서버의 해당 변수의 값을 수정하는 것을 통하여 공격을 수행하였다.

첫 번째 Scenario는 약 60초에 FW Pump Speed를 비정상적으로 낮춤으로써 냉각수 공급이 부족한 상황을 만드는 Scenario이다.

기존 펌프의 속도를 비정상적으로 느리게 만들어, FW Pump 냉각수의 공급이 줄어들게함으로써 SCRAM 신호를 발생하게 설계하였다.

두 번째 Scenario는 약 60초에 SG Level을 정상 수위보다 낮춤으로써 냉각 효율을 저하시키는 상황을 만드는 Scenario이다. SG Level Setpoint의 변수 값을 변경하여 공격을 수행하였다. SG 정상 수위보다 변수 값을 낮춤으로써 SCRAM 신호를 발생하게 설계하였다.

5.3 AI 모델 구현

AI-Based Anomaly Detection Model인 TranAD, Bi-LSTM, O-C SVM의 세 가지 모델을 구현하였다. 모든 데이터는 전처리 과정을 거쳐 각 모델에 입력되었으며, 이를 통해 정상 상태와 비정상(Anomaly) 상태를 분류하였다. 각각의 모델은 다음과 같다:

TranAD는 Transformer 기반의 구조를 이용해 장기 의존성을 학습할 수 있게 하였다[7]. 인코더-디코더 구조를 통해 데이터를 복원하는 과정에서 생긴 오류를 이상(anomaly)으로 판단하고 적대적 학습을 활용하여 이상 탐지 성능을 높였다. 또한 이중 재구성 기법을 이용해 하나는 표준 학습(Standard Training), 다른 하나는 적대적 학습(Adversarial Training)으로 진행하였다.

Bi-LSTM은 시계열 데이터의 시간적 의존성을 학습하여 과거와 미래의 정보를 반영함으로써 이상(anomaly)을 예측 기반으로 탐지하였다[6]. 또한 LSTM 레이어를 3개로 쌓아 더 복잡한 패턴을 학습할 수 있게 구성하였다. 직전 시점들의 정보를 바탕으로, '다음 시점'에 해당하는 다변량 센서값을 예측하도록 설계하였다.

O-C SVM은 정상 데이터를 학습해 정상 패턴을 모델링하고, 그 경계를 벗어나는 데이터를 '이상(Anomaly)'이라 판단한다[5]. 정상 데이터가 존재하는 영역을 최대한 감싸는(Decision boundary) 초평면 혹은 영역을 찾고 그 영역 밖에 떨어지는 데이터 포인트들은 '정상 범위 밖'이라고 보고 이상으로 분류한다.

우리는 각 모델을 사용하여 정상데이터를 학습하였고, 학습한 이후에 이상데이터를 이용하여 예측값을 구하고 실제값과 예측값의 차이를 이상점수(Anomaly score)로 계산한다. 그 다음, 이상점수가 임계치(Threshold)보다 클 경우 이상로 판단한다.

5.4 성능 Matrix

AI 기반 이상 탐지 모델의 성능 평가는 Confusion Matrix를 기반으로 한 성능 지표를 활용하여 이루어졌다. 주요 성능 지표로는 TP(True Positive), FN(False Negative), FP(False Positive), TN(True Negative)가 사용되었으며, 이를 바탕으로 F1-Score, Precision, Recall, FPR(False Positive Rate)을 계산하였다. F1-Score는 Precision과 Recall의 균형을 평가하는 지표로 사용되었으며, Precision은 탐지된 비정상 데이터 중 실제 비정상인 데이터의 비율을 나타낸다. Recall은 실제 비정상 데이터 중 탐지된 데이터의 비율을 나타내며, FPR은 정상 데이터를 비정상으로 잘못 탐지한 비율을 의미한다. 이러한 성능 지표를 통해 각 모델의 탐지 정확도와 신뢰도를 정량적으로 평가하였다.

VI. 실험 결과

6.1 Attack Scenario Simulation 결과

Asherah Simulator와 OPC-UA Simulation Server 및 Wireshark를 활용하여 FW Pump SpeedCmd와 SG Level Setpoint에 대한 Attack Scenario를 실행한 결과는 다음과 같다:

6.1.1 FW Pump SpeedCmd Attack

Wireshark로 핵심 공정 변수를 캡처한 후 만든 Attack Scenario로 NPP가 정상 상태로 운영된 후, 60초 지점에서 공격을 수행하였다. FW Pump SpeedCmd Attack Scenario 수행 결과 그림 5와 같이 Speed가 100에서 0으로 떨어지는 진 후 0을 유지하는 것을 확인 할 수 있었다. 또한, 원자로 시스템이 정상 상태로 운영된 60초간 그림 7과 같이 안정적으로 유지되던 공정 시스템이 FW Pump Speed

의 공격으로 인해 냉각수 공급 부족 현상이 발생하였다. 그로 인해 그림 8과 같이 SCRAM 신호가 활성화되었다.

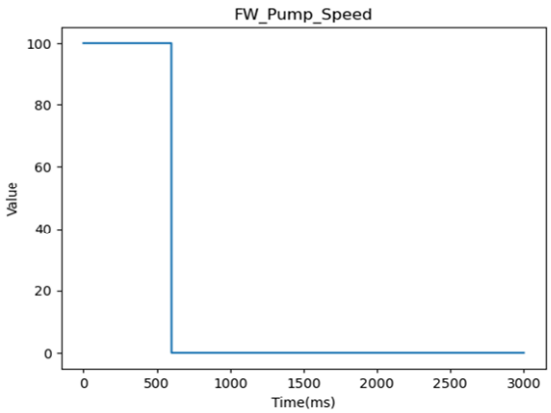

그림 5. FW pump speed 시간별 값 그래프
Fig. 5. FW pump speed value over time graph

6.1.2 SG Level Setpoint Attack

Wireshark로 핵심 공정 변수를 캡처한 후 만든 Attack Scenario로 NPP가 정상 상태로 운영된 후, 60초 지점에서 공격을 수행하였다. SG Level Setpoint Attack Scenario 수행 결과 그림 6과 같이 SG Level 수위가 15에서 0으로 떨어진 후 0으로 유지되는 것을 확인할 수 있었다. 또한, 원자로 시스템이 정상 상태로 운영된 60초간 그림 7과 같이 안정적으로 유지되던 공정 시스템이 SG Level Setpoint 공격으로 인해 냉각수가 급격히 감소하는 현상이 발생하였다. 그로 인해 그림 8과 같이 SCRAM 신호가 활성화되었다.

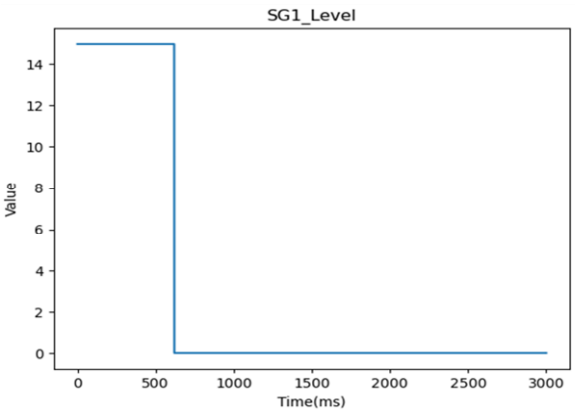

그림 6. SG1 level 시간별 값 그래프
Fig. 6. SG1 level value over time graph

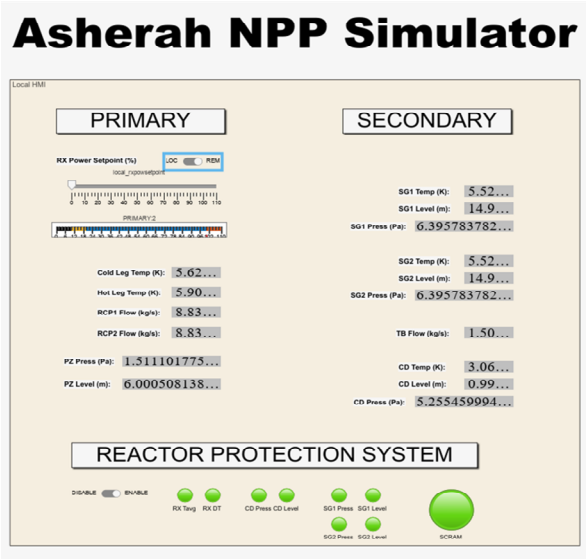

그림 7. Asherah NPP simulator 정상 운전
Fig. 7. Normal operation of asherah NPP simulator

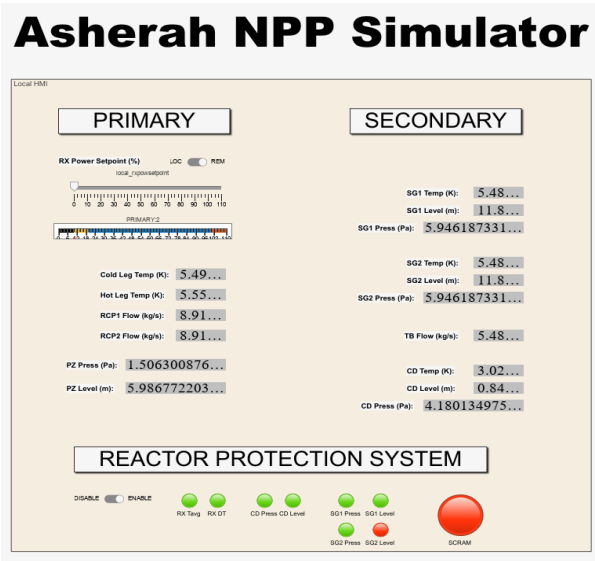

그림 8. Asherah NPP simulator 공격
Fig. 8. Attack on asherah NPP simulator

6.2 이상 탐지 결과

AI-Based Anomaly Detection Model(TranAD, Bi-LSTM, O-C SVM)을 구현하여 정상 데이터 데이터셋을 전처리 하여 학습시켰다. 공격 데이터 데이터셋을 전처리 하여 이상 탐지를 수행하였다. FW Pump SpeedCmd Attack과 SG Level Setpoint Attack에 대한 이상 탐지 결과는 다음과 같다:

6.2.1 FW Pump SpeedCmd Attack 이상 탐지

AI-Based Anomaly Detection Model(TranAD, Bi-LSTM, O-C SVM)을 구현하여 정상 데이터 데이터셋로 학습시키고, FW Pump SpeedCmd Attack 데이터셋으로 이상 탐지를 수행하였다. O-C SVM은 그림 9와 같이 공격 발생 이전에 이상 신호가 탐지되었다. Bi-LSTM은 그림 10과 같이 Attack Scenario 발생 시 급격히 상승하는 이상 신호를 보였지만, 공격이 유지되었을 때 정확한 이상 신호를 보이지 않았다.

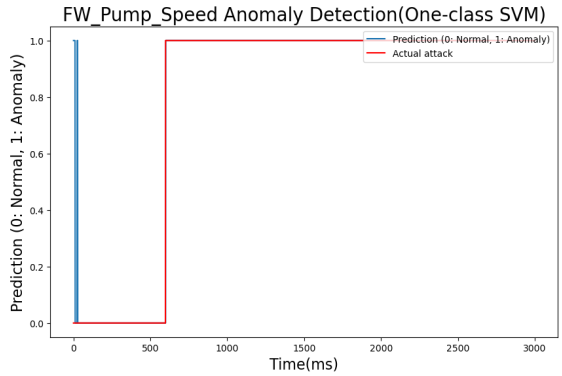

그림 9. 주급수 펌프 이상 탐지(O-C SVM)
Fig. 9. FW pump anomaly detection(O-C SVM)

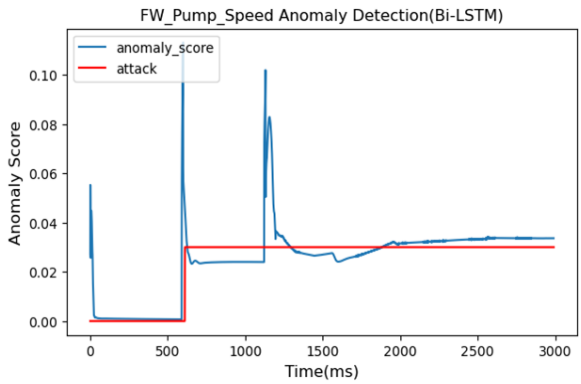

그림 10. 주급수 펌프 이상 탐지(Bi-LSTM)
Fig. 10. FW pump anomaly detection(Bi-LSTM)

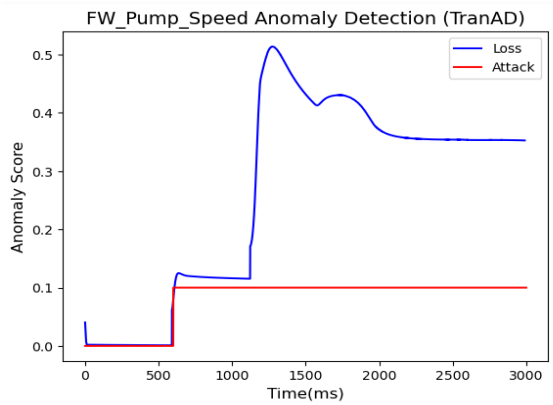

그림 11. 주급수 펌프 이상 탐지(TranAD)
Fig. 11. FW pump anomaly detection(TranAD)

TranAD는 그림 11과 같이 Attack Scenario 발생 시 이상 신호가 급격히 상승하였으며, 공격이 유지되었을 때 바로 하강하지 않고 점진적으로 이상 신호가 상승하였다.

6.2.2 SG Level Setpoint Attack 이상 탐지

AI-Based Anomaly Detection Model(TranAD, Bi-LSTM, O-C SVM)을 구현하여 정상 데이터 데이터셋로 학습시키고, SG Level Setpoint Attack 데이터셋으로 이상 탐지를 수행하였다. O-C SVM은 그림 12와 같이 Attack Scenario 발생 전 이상 신호가 탐지되었다. Bi-LSTM은 그림 13과 같이 Attack Scenario 발생시 그림 13과 같이 급격히 상승하는 이상 신호를 보였지만, 공격이 유지되었을 때, 이상 신호가 하강 후 상승을 반복하였다. TranAD는 그림 13과 같이 Attack Scenario 발생 시 이상 신호가 상승하였고, 공격이 유지되었을 때 이상 신호가 하강하지 않고 점진적으로 상승하였다.

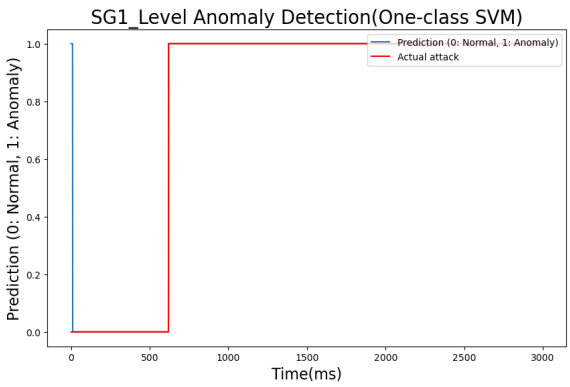

그림 12. 증기발생기 레벨 이상 탐지(Bi-LSTM)
Fig. 12. SG level anomaly detection(Bi-LSTM)

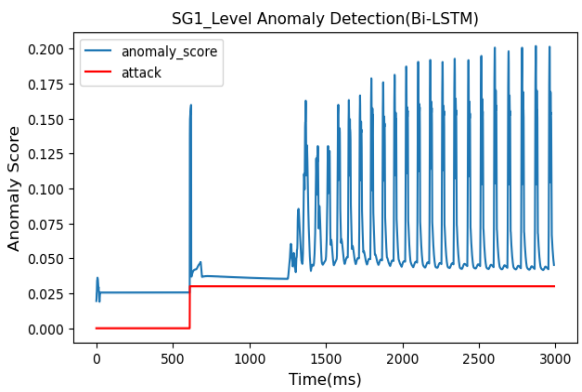

그림 13. 증기발생기 레벨 이상 탐지(Bi-LSTM)
Fig. 13. SG level anomaly detection(Bi-LSTM)

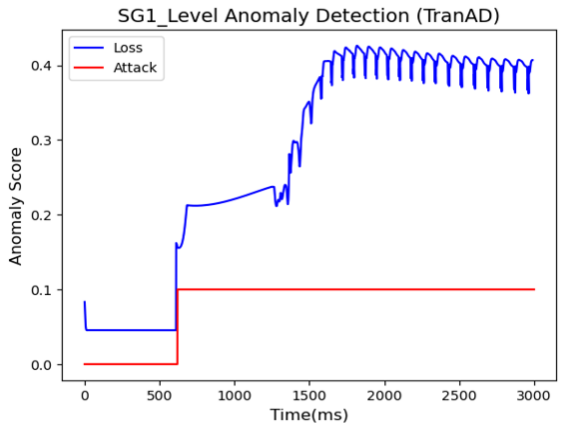

그림 14. 증기발생기 레벨 이상 탐지(TranAD)
Fig. 14. SG level anomaly detection(TranAD)

6.3 이상 탐지 모델 성능 비교

FW Pump SpeedCmd Attack 데이터셋과 SG Level Setpoint Attack 데이터셋을 기반으로 한 AI-Based Anomaly Detection Model(TranAD, Bi-LSTM, O-C SVM)의 이상 탐지 수행 결과로 성능을 평가하였다. 성능 지표는 Confusion Matrix기반으로 F1-Score, Precision, Recall, FPR로 측정되었다. 성능 평가는 다음과 같다:

6.3.1 FW Pump 성능 비교

표 1과 같이 O-C SVM은 F1-Score(0.99), Rrecision(0.99), Recall(0.999), FPR(0.025)으로 F1-Score와 Rrecision, Recall에 높은 성능을 보였다. Bi-LSTM은 F1-Score(0.99), Recall(0.99), Precision(0.98), FPR(0.043)으로 F1-Score와 Recall에 높은 성능을 보였다. TranAD는 F1-Score(0.99), Rrecision(0.99), Recall(0.999), FPR(0.005)로 F1-Score와 Rrecision, Recall에 높은 성능을 보였다. 세 모델 중 TranAD가 F1-Score, Rrecision, Recal를 0.99로, FPR는 제일 낮은 0.005를 기록하였다. 세 모델 중 TranAD가 가장 신뢰도 높은 모델임을 확인하였다.

6.3.2 FW Pump SpeedCmd Attack 이상 탐지 성능 비교

표 2와 같이 O-C SVM은 F1-Score(0.99), Rrecision(0.99), Recall(0.99), FPR(0.16)으로 Precision과 Recall에 높은 성능을 보였다. Bi-LSTM은 F1-Score(0.85), Rrecision(0.98), Recall(0.75), FPR(0.038)으로 Rrecision에 높은 성능을 보였다. TranAD는 F1-Score(0.99), Rrecision(0.99), Recall(0.99), FPR(0.00)으로 모든 지표에서 가장 우수한 성능을 보였다. 세 모델 중 TranAD가 F1-Score, Rrecision, Recal를 0.99로, FPR가 0으로 기록되었다. 세 모델 중 TranAD가 가장 신뢰도 높은 모델임을 확인하였다.

표 1. 주급수 펌프 이상 탐지 성능 지표
Table 1. Performance Metrics of FW Pump Anomaly Detection

Model	TP	FN	FP	TN	F1	Precis-ion	Recall	FPR	thresh old
O-C SVM	2401	1	15	584	0.99	0.99	0.99	0.025	X
Bi-L STM	2400	2	26	573	0.99	0.98	0.99	0.043	0.019
Tran AD	2402	0	3	596	**0.99**	**0.99**	**0.99**	**0.005**	0.06

표 2. 증기발생기 레벨 이상 탐지 성능 지표
Table 2. Performance Metrics of SG Level Anomaly Detection

Model	TP	FN	FP	TN	F1	Precis-ion	Recall	FPR	thresh old
O-C SVM	2381	20	10	590	0.99	0.99	0.99	0.016	X
Bi-L STM	1937	464	21	579	0.85	0.98	0.75	0.038	0.027
Tran AD	2401	0	0	600	**0.99**	**0.99**	**0.99**	**0.00**	0.32

Ⅶ. 결론 및 향후 과제

본 연구에서는 Asherah Simulator와 OPC-UA Simulation Server 및 Wireshark를 활용하여 원자력 시설 사이버 Attack Scenario 시뮬레이션 및 이상 탐지 방법을 제안하였다. Asherah Simulator를 활용하여 Wireshark를 통해, 원자력 시설의 핵심 공정 변수를 캡처하고, 캡처한 핵심 변수를 이용하여 Attack Scenario(FW Pump SpeedCmd Attack, SG Level Setpoint Attack)를 설계한 후 공격을 수행하였다.

AI-Based Anomaly Detection Model(TranAD, Bi-LSTM, O-C SVM)을 구현하여 정상 데이터 데이터셋으로 모델을 학습한 후, 공격 데이터 데이터셋으로 이상 탐지를 수행하였다.

이상 탐지 결과로 세 모델의 성능 지표를 평가해 본 결과 AI-Based Anomaly Detection Model 중 TranAD Model이 가장 높은 이상 탐지 성능을 보였다. 원자력 발전소 사이버 보안 강화에 TranAD Model이 높은 기여도를 가질 가능성을 확인하였다.

향후 연구에서는 원자력 발전소의 여러 종류의 핵심 공정 변수를 캡처하여, 여러 종류의 Attack Scenario를 통해 정상 데이터와 공격 데이터를 구축하여 AI-Based Anomaly Detection Model을 구축할 계획이다. 더 많은 데이터를 활용하여 연구의 실효성을 높이고, 실시간 이사 탐지 기능을 강화할 계획이다.

References

[1] Prosys OPC, "OPC-UA Simulation Server", https://prosysopc.com/products/opc-ua-simulation-server/. [accessed: Jan. 05, 2023].

[2] S. A. Kim, J. Han, A. Chaudhary, and S. Choi, "Methods for Data Collection and Analysis in Nuclear Power Plant Testbeds", Proc. Korean Institute of Information Security and Cryptology (KIISC) Summer Conf. (CISC-S'24), Sejong, Korea, pp. 101, Jun. 2024.

[3] J. Y. Lim, "A Study on Cyber Warfare: Focusing on the Stuxnet Incident (2007-2012)", M.S. thesis, Dept. of Political Science and International Relations, Ewha Womans Univ., Seoul, Korea, 2023.

[4] R. A. Busquim e Silva, K. Shirvan, J. R. C. Piqueira, and R. P. Marques, "Development of the Asherah Nuclear Power Plant Simulator for Cyber Security Assessment", Proc. IAEA Int. Conf. on Nuclear Security, Vienna, Austria, pp. 1-10, Feb. 2020.

[5] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Support Vector Method for Novelty Detection", Advances in Neural Information Processing Systems, Denver CO, USA, Vol. 12, pp. 582-588, Dec. 1999.

[6] A. Chaudhary, J. Han, S. Kim, A. Kim, and S. Choi, "Anomaly Detection and Analysis in Nuclear Power Plants", Electronics, Vol. 13, No. 22, pp. 4428, Nov. 2024. https://doi.org/10.3390/electronics13224428.

[7] S. Tuli, G. Casale, and N. R. Jennings, "TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data", arXiv preprint arXiv:2201.07284, May 2022. https://doi.org/10.48550/arXiv.2201.07284.

[8] MathWorks, "MATLAB", https://kr.mathworks.com/products/matlab.html. [accessed: Jan. 05, 2023].

[9] K. K. Kim and I. C. Euom, "Research on Implementation of Explainable Anomaly Detection Using Artificial Intelligence on Cyber-Physical System", Proc. Transactions of the Korean Nuclear Society Autumn Meeting, Changwon, Korea, Oct. 2024.

[10] K. Marsh, P. Sabharwall, and P. K. Bhowmik, "Integrating PCTRAN with AI-Driven Host-Intrusion Detection and Secured Container Systems for Advanced Malware Analysis", OSTI Tech. Rep., OSTI ID: 2448404, Sep. 2024. https://doi.org/10.2172/2448404.

[11] M. T. Rowland, R. P. Marques, N. White, A. Hahn, and C. Spirito, "NPP Simulator Platform for Cyber Security Research and Training", Sandia National Laboratories, Tech. Rep. SAND2022-7654C, 2022.

[12] S. M. Lim, "Trends on Graded Application of Cyber Security Control Based on Characteristic of Facilities and Digital Assets", Proc. Joint Summer Conf. Korea Institute of Information Technology and Korea Digital Contents Society, pp. 117-119, Jun. 2020.

저자소개

김 성 아 (Seongah Kim)

2021년 8월 ~ 2021년 8월 : 군산대학교 법학과(법학사)
2023년 9월 ~ 현재 : 전북대학교 소프트웨어공학과 석사과정
2023년 9월 ~ 현재 : 전북대학교 지능형보안연구실 연구원
관심분야 : 지능형보안, 원자력보안, 인공지능, 적대적공격, 법학

한 준 서 (Junseo Han)

2020년 3월 ~ 현재 : 전북대학교 소프트웨어공학과 학사과정
2023년 1월 ~ 현재 : 전북대학교 지능형보안연구실 연구원
관심분야 : 지능형보안, 원자력보안, 인공지능, 데이터분석

최 선 오 (Sunoh Choi)

2005년 2월 : 고려대학교 컴퓨터학과(이학사)
2014년 5월 : Purdue Univ. 컴퓨터공학부(공학박사)
2014년 8월 ~ 2019년 2월 : ETRI 정보보호본부 선임연구원
2021년 3월 ~ 현재 : 전북대학교 소프트웨어공학과 부교수
관심분야 : 지능형보안, 모빌리티보안, 원자력보안