원자력발전소 운전지원을 위한 인공지능 모델의 신뢰성

박재관, 구서룡
한국원자력연구원

Jkpark183@kaeri.re.kr, srkoo@kaeri.re.kr

Reliability of Artificial Intelligence Models for Supporting Nuclear Plant Operations

Jaekwan Park, Seoryong Koo
Korea Atomic Energy Research Institute

요 약

운전지원은 원자력 분야에서 인공지능 기술의 현장 적용 가능성이 가장 높은 기능 중에 하나이다. 원전 시설을 감시, 비정상 탐지, 사고 예측, 조치 지원과 같은 다양한 세부 기능에서 인공지능 기술을 적용하기 위한 연구가 수행되고 있다. 원전은 안전상의 이유로 규제기관의 신뢰성 검토를 통한 허가를 받은 기능만 현장에서 운용 가능하다. 그러나, 인공지능 기술은 전통적인 소프트웨어와는 다른 특성들 때문에 기존 방식으로는 신뢰성을 평가하기 어렵고 이것은 현장 적용을 더디게 하는 요소가 된다. 이 논문은 국내외 문헌조사를 통해 원전분야에서 인공지능 모델의 신뢰성을 평가하기 위한 항목들을 검증 요건으로 제안한다. 이러한 검증 요건에 대한 정성적, 정량적 평가는 원전 인공지능 모델의 신뢰성을 평가하는데 도움이 될 것으로 기대된다.

1. 서론

원자력발전소는 제어실의 운전원들이 조작하는 보호, 제어, 감시를 위한 디지털 시스템에 의해 운영된다. 최근, 운전원들을 지원하기 위해서 인공지능 기술을 활용한 다양한 운전지원시스템들이 연구되고 있다.

원자력분야에 사용되는 안전등급의 디지털 시스템은 RG 1.168[1]에 명시된 Software integrity, Software reliability, QA, SW development tool, V&V tasks 와 같은 규제 지침(Regulatory positions)을 만족해야 한다. 운전지원과 같은 비안전등급의 시스템이더라도 최소 SW reliability 와 V&V Tasks 에 대한 표준인 IEEE 1012[2]에 따른 확인 및 검증 활동이 요구된다. 그런데, Zhang et. al.[3]과 ISO/IEC TR 29119-11[4]에 소개된 바와 같이, 인공지능 모델은 전통적인 소프트웨어와는 다르게 높은 복잡도, 명세 부족, 비 결정적 행동과 같은 독특한 특성들을 가지고 있다. 이러한 특성은 기존의 전통적인 방법들로는 인공지능 모델을 검증하는 것을 어렵게 만든다.

최근 인공지능 모델에 대한 특성을 정의하고 신뢰성을 확보하기 위한 연구들이 진행되고 있다. 먼저, 인공지능 모델에 대한 다양한 관점을 정리한 Zhang et. al.[3]에서는 인공지능 시스템에 대하여 Correctness, Generalization, Robustness 의 3 가지 기능적 요건과 Interpretability, Security, Efficiency, Fairness, Data Privacy 의 5 가지 비-기능적 요건 관점에서 평가할 필요가 있음을 기술하였다. TTA.KO-10.1497[5]에서는 IT 분야에서 인공지능 시스템의 신뢰성을 확보하기 위해서 데이터, 모델, 시스템 관점에서 15 가지 요구사항을 정의하였다. 또한, 국제표준기관에서 발행한 기술보고서인 ISO/IEC TR 29119-11[4]은 인공지능 시스템의 신뢰성을 보장하기 위해서 알고리즘과 모델의 정보를 제공하고 개발자와 독립적인 시험, 적대적 사례 시험과 같은 구체적인 방법들을 소개하였다. 국제 공학기술연구소의 보고서[6]에서도 항공, 원자력과 같은 안전분야에 사용되는 인공지능 시스템에 대하여 모델(Model), 명세(Specification), 확인 및 검증(Verification and validation), 보안(Security)과 같은 주요 평가 항목을 제안하였다. 또한, 시험 관점에서는 모델의 학습 입력(input domain), 운영 도메인(operational domain), 고장 도메인(failure domain), 적대적 도메인(adversarial domain)으로 나누어 수행될 필요가 있음을 기술하였다.

본 논문은 이러한 인공지능 신뢰성 관련 문헌들을

기반으로 원전 분야에 필요하다고 판단되는 요건들을 선별하고 운영 환경을 고려하여 적절히 변경함으로써 검증 요건들을 정의한다.

2. 원전 인공지능 모델의 신뢰성 검증 요건

본 논문에서 설정한 원전 인공지능 모델 검증 요건은 표 1 과 같다.

<표 1> 원전 의사결정지원을 위한 인공지능 검증 요건

요구사항 ID	요구사항 명칭
REQ-1	오픈소스 라이브러리의 보안성 및 호환성 확보
REQ-1-1	오픈소스 라이브러리의 안전성 확인
REQ-1-2	오픈소스 라이브러리의 위험요소 관리
REQ-1-2-1	사용 중인 오픈소스 라이브러리 라이선스 준수사항 이행
REQ-1-2-1	사용 중인 오픈소스 라이브러리의 호환성 및 보안취약점 확인
REQ-2	인공지능 모델 공격에 대한 방어 대책 수립
REQ-2-1	모델 추출 공격에 대한 방어 방안 수립
REQ-2-2	모델 회피 공격에 대한 방어 방안 수립
REQ-3	인공지능 모델 명세 및 추론 결과에 대한 설명 제공
REQ-3-1	인공지능 모델 명세를 투명하게 제공
REQ-3-2	신뢰도 제공이 필요한 인공지능 모델 출력결과에 대한 신뢰도를 제공
REQ-3-3	사용자가 모델 추론 결과의 도출 과정을 수용할 수 있도록 근거 제공
REQ-4	인공지능 모델 성능 평가
REQ-4-1	인공지능 모델의 정확도 성능 평가
REQ-4-2	인공지능 모델의 강건성 성능 평가

첫 번째 요건은 오픈소스 라이브러리의 보안성 및 호환성 확보이다. 인공지능 모델 개발에서는 다양한 오픈소스 라이브러리가 활용된다. 원전은 안전이 매우 중요한 산업분야이므로 사용되는 시스템의 보안성과 라이브러리간의 호환성은 가장 기본적인 요건이다. 따라서, 안전성, 보안성 측면에서 신뢰수준, 안정적인 업데이트 여부, 라이브러리의 보안 취약점 등의 관점에서 해당 오픈소스를 확인하여 운영 및 보안상 위험요소를 점검할 필요가 있다.

빈번하게 업데이트가 진행되는 인공지능 라이브러리의 특성상 특정 버전의 코드로 장기간 검토하는 기존의 품질 보증 방법으로 안전성을 평가하는 것은 어렵다. 또한, 장기간의 작동 경험으로 그러한 품질 신뢰도를 보여주는 것도 어렵다. 다만, 전세계 많은 개발자들이 동시에 사용하고 있고, 확인되는 문제점들이 지속적으로 개선되어 사용되고 있는 널리 알려진 오픈소스는 안전성을 가진 것으로 판단하는 것이 적용 가능한 합리적인 기준이 될 수 있다. 이러한 관점에서 [5]의 오픈소스 안전성 요건과 [7]의 오픈소스 체크항목을 기반으로 오픈소스 라이브러리에 대하여 아래와 같은 검토 항목을 도출하였다.

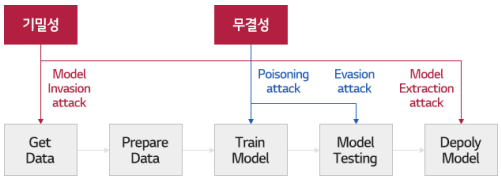
- ✓ 오픈소스 라이브러리 식별 및 목록 관리 여부
- ✓ 오픈소스 기술지원(패치 및 버전 업데이트) 여부 확인
- ✓ 오픈소스 커뮤니티 활성화 여부
- ✓ 오픈소스 목록 관리의 자동화 활용 여부

오픈소스 위험요소 관리 요건에는 먼저 라이선스 준수가 있다. 기본적으로 오픈소스는 무료로 사용할 수 있지만, 라이선스별로 준수사항은 별도로 규정된다. 따라서, 라이선스 사항을 확인하여 발생가능 한 법률적 위험을 최소화할 필요가 있다. 예를들어, 주요 오픈소스 라이브러리인 Tensorflow, Keras, PyTorch 는 자유로운 재배포, 소스코드 공개는 의무가 아니며, 개인 및 단체에서 차별없이 사용 가능하다.

또한, 위험요소 관리요건에는 오픈소스 라이브러리의 호환성 및 보안취약점 확인이 요구된다. 라이브러리의 버전 변경 과정에서 다른 라이브러리 버전과 호환되지 않는 호환성 문제를 초래할 수 있다. 따라서, 라이브러리간 의존성을 파악하여 호환성을 확인해야 한다. 오픈소스는 코드가 공개되어 있으므로 공격자가 공개된 코드를 통해 결함을 발견하고 악의적인 공격을 할 수 있다. 따라서, 사용하는 오픈소스에 보고된 보안취약점이 있는지 확인이 필요하다.

두 번째 요건은 인공지능 모델 공격에 대한 방어 대책 수립이다. 인공지능 모델의 신뢰성에 대한 가장 큰 우려 중에 하나는 적대적 공격(Adversarial attack)에 의한 오작동이다. 원전 계측제어시스템은 폐쇄망으로 구성되는 사이버보안 규정이 적용되기 때문에 인터넷에 직접 연결되는 IT 분야와 달리 공격의 가능성은 매우 낮다. 그러나, 원전 시스템에서 오작동이 발생할 경우, 운전원의 인적오류를 유발하여 발전소 운영에 악영향을 미칠 수 있기 때문에 인공지능 모델 개발에서 적대적 공격에 대응하기 위한 방법이 고려될 필요가 있다.

인공지능 모델의 학습과정에서 기밀성과 무결성을 공격하는 적대적 공격에는 학습과정에서 악의적인 학습 데이터를 주입하여 모델을 망가뜨리는 데이터 중독 공격(Data Poisoning Attack), 인공지능 모델의 입력 데이터를 교란하여 모델을 속이는 회피 공격(Evasion Attack), 역공학을 이용해 모델이나 학습 데이터를 탈취하는 모델 추출 공격(Model Extraction Attack)과 학습 데이터 추출 공격(Inversion Attack)이 있다[5].

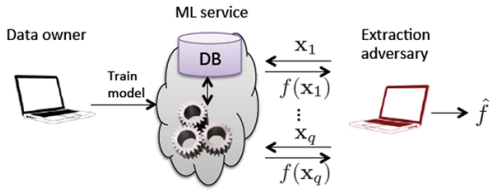


(그림 1) 적대적 공격 4 가지 유형 (출처: Kundu[8])

인공지능 모델 관점에서는 (1) 모델이 탑재되는 시스템의 운영환경을 분석하고, (2) 모델 추출 공격, 모델 회피 공격에 대한 방지 또는 완화 방안을 수립해 적대적 의도로 실행되는 공격에 대비할 필요가 있다.

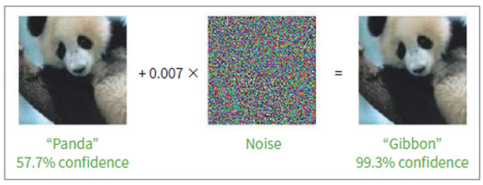
모델 추출 공격은 개발된 인공지능 모델을 제 3 자

가 탈취하는 공격이다. 공격자는 모델의 API 를 통해 쿼리를 계속 던져서 반환되는 예측값과 신뢰도값을 분석하여 유사한 모델을 만들어낸다. 이를 통해, 공격자는 모델을 탈취하거나 모델 회피 공격과 같은 2 차 공격을 위한 도구로 활용할 수 있다.



(그림 2) 모델 추출 공격 (출처: Tramer et. al.[9])

모델 회피 공격은 입력 데이터에 최소 변조를 가하여 인공지능 모델을 속이는 기법이다. 아래와 같이, 팬더 이미지에 사람의 눈으로는 분간하기 어려운 노이즈를 추가하면 인공지능 모델이 이를 긴팔원숭이로 착각하게 만드는 것이 가능하다. 이러한 공격은 모델을 오작동하게 만들기 때문에 이에 대한 대책이 필요하다.



(그림 3) 모델 회피 공격 (출처: Goodfellow et. al.[10])

세 번째 요건은 인공지능 모델 명세 및 추론 결과에 대한 설명 제공이다. 제어실 사용자가 인공지능 모델의 추론 결과를 신뢰하기 위해서는 인공지능 모델이 제공하는 판단 또는 추론 결과의 도출 과정을 이해할 수 있어야 하며, 사용자에게 모델에 대한 정보와 함께 추론 결과에 대한 설명 및 근거를 제시하는 것이 바람직하다[5].

이를 위해, 인공지능 모델의 개발 및 테스트 과정에서 발생한 다양한 정보를 문서로 제공하는 것이 필요하다. 이러한 문서를 모델 명세서[11]라고 한다. 모델 명세서가 제공될 경우, 사용자는 모델의 입출력 및 성능에 대한 사항을 참조할 수 있다. 즉, 개발자는 인공지능 모델의 개발과정 및 시험에서의 다양한 결과를 기록한 모델 명세서를 작성하여야 한다. 이 문서에는 최소한 모델의 목적, 모델의 구조, 입출력 정보, 성능 등의 결과를 포함해야 한다.

또한, 모델의 추론 결과가 참값(Ground Truth)과 일치할 확률을 사용자에게 제공함으로써 인공지능 모델이 도출한 결과를 얼마나 확신하는지를 나타내는 불확실성을 모델 추론 결과에 대한 설명으로 제공할 수 있다. 이는 사용자가 더 정확한 판단을 할 수 있도록 지원하면서 신뢰성을 높이는 방법이다. 인공지능 모델 출력에 대한 신뢰도 제공이 필요할 경우, 다음에 따라 신뢰도 정보를 제공한다.

- ✓ 모델의 추론 결과가 참값과 일치할 확률을 계산할 수 있다면, 이것을 모델의 최종 의사결정에 대한 설명으로 사용할 수 있다. 확률 변수의 분산 크기로 인공지능 모델이 도출한 결과를 얼마나 확신하는지를 나타내는 불확실성을 모델 추론 결과에 대한 설명으로 고려해 볼 수 있다.
- ✓ 모델의 예측 확률과 모델 출력의 불확실성의 두 가지 정보 관점에서, 예측 확률이 임계치 보다 낮은 경우 또는 불확실성이 높은 경우에는 사용자가 이를 인지하도록 모델 추론 결과에 대한 설명을 반드시 제공하여야 한다. 이것은 기대하는 또는 유지되어야 하는 품질 수준보다 낮은 상황을 포함한다.
- ✓ HMI 로 표시할 경우, 예측 확률은 수치로 표현하고 불확실성은 이미지화 하여 표현하는 방법이 사용될 수 있다.

또한, 사용자가 모델 추론 결과의 도출 과정을 수용할 수 있도록 근거 제공할 필요가 있다. 인공지능 모델의 성능은 우수하지만, 설명가능성은 낮다. 낮은 설명가능성은 모델 예측에 대한 신뢰도를 낮출 수 있기 때문에, 사용자가 납득 가능한 모델 출력 결과의 근거를 확보할 필요가 있다. 이를 위해, 최근 연구되고 있는 입력 요인, 모델 내부, 설명 변수 분석 등과 같은 다양한 설명가능한 인공지능(XAI, eXplainable AI) 기술의 도입을 고려할 수 있으며, 이를 통해 모델 추론 및 출력 결과의 근거를 시각화하여 제시할 수 있다.

- ✓ XAI 기술 적용이 가능한 경우, 추론 결과 설명을 위한 기법 적용을 검토한다. 추론 결과의 도출 과정에 대한 이해(모델의 작동방식, 최근 추이, 판단 근거 등)할 수 있도록 관련 정보를 사용자에게 제시한다.
- ✓ XAI 기술 적용이 불가능한 경우, 대안을 마련하여 적용한다. XAI 기술 적용이 어려우면, 차선책을 고려할 수 있다. 딥러닝 모델이 복잡성으로 인해 현재 기술 수준으로는 XAI 기술의 적용이 어려울 수 있다. 예를들면, 역전파 기반 XAI 를 적용한다고 하더라도, 추론 결과에 어느 신호들이 영향을 미치는지 일부 표시가 가능하지만 사용자가 보기에는 해석성이 떨어질 수 있다. 대안으로써 의사 결정 트리와 같은 전통적인 방법을 도입하는 것도 고려해 볼 수 있다.

네 번째 요건은 인공지능 모델의 성능 평가이다.

인공지능 모델 개발의 최종 목표는 목표기준(또는 허용기준) 이상의 성능을 갖추는 것이다. 그리고, 개발 과정에서 노출되지 않은 오류 및 불확실성을 확인하기 위해서 제 3 자 시험을 수행하여 모델 성능 결과에 대한 신뢰성을 높일 수 있다. 또한, 현장에서 발생할 수 있는 다양한 변동 상황을 검토하고 그러한 상황에서 모델이 강건하게 작동할 수 있는지 평가할 필요가 있다[4]. 본 논문에서는 이를 종합적으로 고려하여, 독립된 시험자에 의해 인공지능 모델에 대한 정확도 평가와 강건성 평가로 분류한다.

먼저, 인공지능 모델의 정확도 성능은 다음과 같은 절차를 통해 평가할 수 있다.

- ✓ 개발자가 제공한 인공지능 모델의 성능 지표를 확인하고 검토한다.
- ✓ 학습 데이터세트와 시험 데이터세트를 확보하고 랜덤 추출로 시험 케이스를 생성하고 시험을 수행하여 개발자의 성능결과와 비교한다.
- ✓ 개발자가 제공한 성능지표가 재현되는지 확인한다.
- ✓ 추가적인 지표가 필요하다고 판단될 경우, 해당 지표의 성능을 측정한다.

또한, 인공지능 모델의 강건성 평가는 다음과 같은 사항을 고려하여 수행한다.

- ✓ 운전지원을 위한 인공지능 모델은 원전 시뮬레이터에서 생산된 데이터세트로 학습하지만, 원전 운영 환경에서는 그러한 클린 데이터(Clean data)만 존재하는 것은 아니다.
- ✓ 장기간 고온, 고압, 방사선과 같은 가혹 환경에서 운영되는 계측 장비들은 서서히 노화가 발생하고 고장 또는 교체 주기에 이르기 전까지는 정상 작동된다.
- ✓ 이러한 자연스러운 계측채널 노화로 인한 일정 범위내의 드리프트(drift)는 원전 설계단계부터 고려되어 허용된다.
- ✓ 따라서, 학습 데이터에 드리프트 변화를 주입하여 시험 케이스를 생성하고 시험함으로써 인공지능 모델의 현장 적용 강건성을 확인할 수 있다.

3. 결론 및 향후 연구

인공지능 기술이 성숙함에 따라, 원자력 분야에서도 인공지능 기술 적용을 위한 연구가 활발히 이뤄지고 있다. 인공지능 기술의 현장 적용성의 공감대가 이뤄지고 있는 부분은 제어실 운전지원 기능이다. 하지만, 기존 소프트웨어와 달리, 높은 복잡도, 명세 부족, 비 결정적 행동과 같은 특성으로 인해 신뢰성 검증이 어려운 문제가 있고, 이것은 현장 적용을 위해서 거쳐야 할 규제 기관의 인허가를 이끌어 내기 어렵게 한다. 본 논문은 국내외 관련 문헌들을 기반으로 공통적으로 중요하게 기술된 항목들 중에서 원전에서 중요하게 고려될 필요가 있는 신뢰성 항목을 검증 요건으로 제안하였다. 또한, 각 항목별로 중요 고려사항을 기술하여 평가 방향을 제시하였다. 향후, 이 요건을 기반으로 실제 모델에 대한 심층 평가 사례 연구를 수행할 계획이다.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. RS-2022-00144150) and the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry, & Energy (MOTIE) of the Republic of Korea (No. 20224B10100130)

참고문헌

[1] US NRC, RG 1.168, Verification, Validation, Reviews, and Audits for Digital Computer Software used in Safety Systems of Nuclear Power Plants, 2013

[2] IEEE, IEEE 1012, Standard for Software Verification and Validation, 2004

[3] Zhang et. al., “Machine Learning Testing: Survey, Landscapes and Horizons”, IEEE Transactions on Software Engineering, Vol. 48, Issue 1, pp. 1-36, 2022

[4] ISO/IEC, ISO/IEC TR 29119-11, "Guidelines on the testing of AI-based systems”, 2020

[5] 한국정보통신기술협회, TTAK.KO-10.1497, “인공지능 시스템 신뢰성 제고를 위한 요구사항”, 2023

[6] The Institution of Engineering and Technology, "The Application of Artificial Intelligence in Functional Safety", 2024

[7] 정보통신산업진흥원, “기업 공개소프트웨어 거버넌스 가이드”, 2021

[8] Sandip Kundu, “Security and Privacy of Machine Learning Algorithms”, International Symposium on Quality Electronic Design, 2019

[9] Florian Tramer et. al., "Stealing Machine Learning Models via Prediction APIs", Proceedings of the 25th USENIX Conference on Security Symposium, pp. 601-618, 2016

[10] Ian Goodfellow et. al., "Explain and harnessing adversarial examples", ICLR, 2015

[11] 한국정보통신기술협회, TTAK.KO-11.0280, “검증용 데이터세트의 밸런스 기반 인공지능 소프트웨어 신뢰성 평가 방법 - 제 3 부: 시계열 타입 밸런스 데이터 설계”, 2021