Journal of the Korea Academia-Industrial cooperation Society
Vol. 25, No. 8 pp. 559-565, 2024

https://doi.org/10.5762/KAIS.2024.25.8.559
ISSN 1975-4701 / eISSN 2288-4688

AI 적용 체계 데이터 품질 점검 및 운용유지단계 모니터링 실증 연구

이영민*, 강지훈, 정민경, 박주영
국방기술품질원 첨단미래기술센터 AI·사이버팀

A Study on Data Quality Inspection and Monitoring at Operation and Sustainment Stage of AI Application System

Yeong-Min Lee*, Ji-Hoon Kang, Min-Kyung Jeong, Joo-Young Park
AI · Cyber Team, Advanced Technology Center, Defense Agency for Technology and Quality

요 약 전세계적으로 인공지능(AI)의 발전과 함께 다양한 AI 기술이 개발되고 민수 분야 현업에 적용되고 있는 상황이다. 우리나라 국방 분야에서도 국방혁신 4.0 등을 통해 다양한 AI 관련 정책 제도와 AI 적용 체계 획득이 추진되고 있다. AI 적용 체계의 품질을 확보하기 위한 방안으로 데이터의 품질 점검과 지속적으로 진화되는 AI의 특성에 맞게 지속 확인하는 업무도 필요하다. 본 논문에서는 AI 성능 평가와 모니터링 업무의 기본 개념을 설명하고, 실제 운용 중인 AI 적용 체계를 대상으로 품질 점검 및 성능 개선을 수행한 실증 결과를 제시하였다. 4차 학습 및 성능 평가 실증을 통해 객체별 성능지표 α 및 β의 향상도를 확인 할 수 있었으며, 보완해야 할 사항에 대해서도 파악할 수 있었다. 이러한 모니터링 업무를 통해 획득된 AI 체계의 품질을 개선 시킬 수 있으며, 개발 및 양산 단계로의 피드백을 통해 후속 AI 체계의 성능 고도화까지 가능할 것으로 사료된다.

Abstract With the development of artificial intelligence (AI), various AI technologies have been developed worldwide and are being applied to private demand. In the Korean defense sector, various AI-related policy systems and AI application systems are also being acquired through Defense-innovation 4.0. To secure the quality of the AI application system, it is also necessary to check the quality of data and continue to check according to the characteristics of continuously evolving AI. This paper explains the basic concepts of AI performance evaluations and monitoring and reports the results of a quality inspection and performance improvement for the AI application system under actual operation. Improvements in performance indicators α and β for each class and the matters that need to be supplemented could be confirmed by demonstrating the 4th learning and performance evaluation. The quality of the acquired AI system can be improved by such monitoring, and feedback on the development and mass production stages will improve the performance of subsequent AI systems.

Keywords : AI, Data Quality, AI Application System, AI Monitoring, Quality Management, Quality Inspection

*Corresponding Author : Yeong-Min Lee(DTaQ)
email: ymlee@dtaq.re.kr
Received May 31, 2024 Revised June 2, 2024
Accepted August 2, 2024 Published August 31, 2024

1. 서론

최근 미국을 비롯한 다양한 선진국에서는 인공지능(Artificial-Intelligence, 이하 AI) 기술에 집중적으로 투자하고 있으며, 특히 국방 분야에서도 전투원들의 생존성을 향상시키고, 전투 효율을 극대화하기 위하여 AI 기술이 적용된 무기체계를 도입하고 있다[1]. 하지만 AI 기술은 데이터 편향에 따른 차별적 결과, 딥페이크와 같은 윤리적 문제, 오탐 등 아직까지 한계점은 분명히 존재하고 있는 상황이다. 특히 생명과 직결될 수 있는 국방 및 의료 분야에서는 AI의 지속적인 품질을 확인하고 관리하는 것이 매우 중요하다고 할 수 있다.

AI 시스템은 기본적으로 소프트웨어와 함께 구현된다. 따라서 AI 시스템의 품질은 전반적으로 소프트웨어 시스템의 품질 체계를 따르게 된다. AI 시스템의 품질은 ISO/IEC 25010[2]에서 다루어지고 있는 품질 속성에 AI 관점의 하위 속성(Sub-characteristcis)을 추가함으로써 그 정의를 확장하고 있다[3]. ISO/IEC 25010에서 제시하고 있는 소프트웨어 시스템 품질 모델에서 소프트웨어 시스템 품질 하위 속성에 기능적합성, 호환성, 신뢰성(Reliability) 등이 포함되어 있다.

미 국방부(DoD)에서도 2020년 AI의 설계, 개발 및 배포에 대한 5가지 윤리원칙을 채택하고, 2022년에는 이러한 윤리원칙에 부합하는 역량 구축을 위해 '책임있는 AI 전략 구현 경로'를 배포하는 등 AI의 신뢰성을 매우 중요한 요소로 판단하고 있다[4,5]. 국내에서는 국방부, 방위사업청을 비롯한 국방 유관기관 또한 AI가 적용될 체계의 신뢰성 확보를 위한 여러 준비를 하고 있는 실정이다. 방위사업청 출연기관인 국방기술품질원에서도 AI 적용 체계의 전순기 품질관리를 위한 연구를 통해 데이터 품질관리 매뉴얼을 작성하는 등 AI 적용 품질관리를 위한 활동을 수행하고 있으나 현재 체계 개발을 통해 도입된 AI 적용 체계가 적어 AI 데이터 품질 관리 및 성능 평가 관련 상세한 실증 사례가 없는 상황이었다. 따라서, 본 논문에서는 우리나라에서 시범적으로 진행되었던 AI 사업을 대상으로 AI 적용 무기체계의 성능 확인을 위한 데이터 품질관리 및 운용유지 단계 모니터링 실증 연구 수행 결과와 한계점을 제시하고 그 후속 연구 방향성까지 함께 제안하고자 한다.

2. 본론

2.1 AI의 특성 및 성능 확인 절차

기존 여러 무기체계들에 탑재가 되는 소프트웨어의 경우 소프트웨어 산출물을 통하여 요구사항 달성 정도를 설계, 코드 검증 등을 통하여 수행할 수 있었다. 하지만 무기체계에 적용되는 AI를 검증하기 위해서는 기존 전통적인 소프트웨어 검증 방식과는 다른 접근 방식이 필요한데, 그 근본적인 이유는 Fig. 1과 같이 AI가 기존 전통적인 소프트웨어와는 개발 방법과 동작 특성에서 차이점이 존재하기 때문이다. 전통적 소프트웨어 개발은 정해진 입력과 출력을 산출하기 위해 알려진 절차와 방법을 통하여 구현하며 코드를 수정하기 전까지 항상 동일한 결과를 산출하는 결정론적 특성을 가진다. 반면 AI 기능 개발의 경우 원하는 출력을 얻기위한 모형을 기계적으로 생성하게 되는데 이 모형은 동적으로 변화되며, 확률적 결과를 제시하는 비결정론적 특성을 가진다[6].

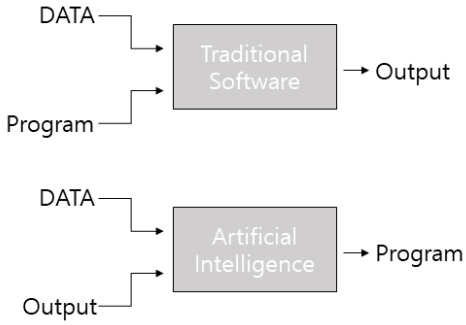

Fig. 1. Comparison of Traditional SW and AI Operation Characteristic[6]

ISO/IEC TR 29119-11 기술보고서[7]에서도 AI 적용 체계(AI-based systems)의 경우 시스템이 지속적으로 변화하기 때문에 평가하기가 어렵다는 'Self-learnig system' 문제와 확률론적 특성 때문에 결과를 예측하기 어렵고 매번 동일한 결과를 도출하지 못한다는 것을 의미하는 'Probabilistic and non-deterministic systems' 문제 등 다양한 AI 적용 체계의 평가의 어려운 사항들에 대해서 언급하고 있다.

따라서 AI의 성능을 확인하기 위하여 단순한 산출물 검토가 아닌 AI 특성에 맞는 별도 '성능 평가'가 필수적이라고 할 수 있으며, 꾸준한 학습이 요구되는 AI 특성에 따라 개발단계에서 반복적인 성능평가를 수행하는 것이 중요하다.

이러한 절차를 거쳐서 전력화되어 사용자(무기체계의 경우 소요군)가 사용하고 있는 운용 단계에서 또한 지속적인 재학습과 평가를 통한 성능 향상이 필요하며 본 논

문에서는 이 절차를 운용유지단계 '모니터링' 업무라고 지칭한다. ISO/IEC TR 29119-11 기술보고서에서도 'Evolution' 문제 언급을 통해 진화되는 시스템을 주기적인 평가를 통해 유지·보수하여 정확도(Accuracy), 정밀도(Precision) 등의 성능 목표와 편향성들을 확인해야 한다고 언급하고 있다.

본 절의 내용을 요약하자면 AI의 성능을 확인하기 위한 별도의 절차로 '성능평가'와 '모니터링'이라는 업무가 필수적이라고 생각되며 본 논문에서 성능평가 및 모니터링의 상세 절차와 실제 시범 적용 사례에 대해 제시할 예정이다.

2.2 AI 성능 평가

AI 적용 무기체계의 성능 확인을 위한 첫 번째 업무인 '성능평가'는 Fig. 2와 같이 단위기능 평가와 시스템 평가로 구분될 수 있다[1].

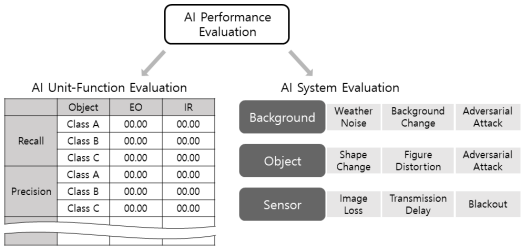

Fig. 2. Classification of AI Performance Evaluation

AI 단위기능 평가는 AI 관련 요구사항에 따른 기능적 정합성을 만족하는지 확인하는 시험으로, 기존 민간분야에서 사용되는 혼동 매트릭스를 기반으로 한 성능지표 달성 여부를 판단하는 것을 의미한다. 탐지/추적 분야에서는 재현율(Recall), 정밀도(Precision), 정확도, F1 Score, mAP 등의 성능지표 들을 데이터 기반의 시험을 통해 정량적인 수치를 확인하는 것이 대표적인 예시이다.

AI 시스템 평가는 AI 기능 외적 요소에 대한 영향을 확인하는 것이라고 할 수 있으며, 외적 요소에 대한 중점은 시스템의 강건성 확인이다. 여기에서 강건성 확인 방법으로는 ISO/IEC TR 29119 AI 적용 체계 블랙박스 테스팅 파트에서 언급되고 있는 기법인 '조합 테스팅, 백투백 테스팅, A/B 테스팅, 변성 테스팅' 4가지 중 변성 테스팅을 적용하는 것을 고려하고 있다. 변성 테스팅 기법을 적용한 AI 시스템 평가는 다양한 외부 요인으로 인한 위협 상황에서도 신뢰할 수 있는 시스템인지 확인하는 것을 목적으로 하며, 외부 요인의 예시는 날씨 변화, 학습 시 예상하지 못한 상황, 적대적 공격 등이 있다. AI 시스템 평가의 개념은 일반 무기체계 중 HW의 내구도 시험, 혹한기나 혹서기와 같은 가혹 조건에서의 시험과 유사한 개념이라고 할 수 있다. 해당 시스템 평가에는 학습 시 사용된 데이터를 기반으로 외부 요인을 반영한 데이터를 평가 데이터로 사용하여야 하므로 합성데이터의 사용이 필수적이다.

2.3 운용유지 단계 모니터링

AI 기술이 적용되는 무기체계의 경우 군에 납품이 된 후 실운용 환경에서 기대에 미치지 못하는 성능을 내는 경우가 있다. 한 예로 AI 관련 사업 요구성능이 모델의 특성에 맞지 않는 성능지표를 선정하여 시험평가를 수행한 사례가 대표적이라고 할 수 있다. 또한, 데이터의 수량 부족, 데이터의 보안 사항으로 인한 제한적 학습 등의 이유로 모델이 완전하게 학습되지 않은 상태에서 실물에 의한 시험평가가 수행되는 것도 실운용 환경에서의 성능

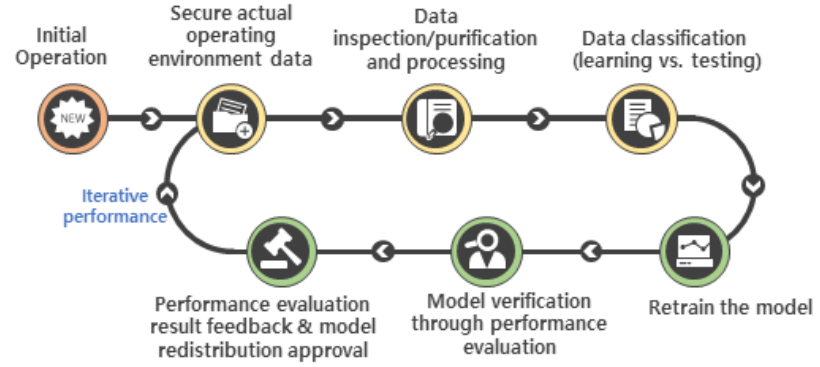

Fig. 3. AI Monitoring Procedures at Operation and Sustainment Stage

저하의 원인이라고 할 수 있다. 따라서, 납품 전 사업관리 단계에서 AI 성능평가(AI 기능 평가 및 시스템 평가)를 수행하였더라도, 전력화 이후 실제 배치된 운용 환경과의 적합성, 지속적인 모델 최신화 등의 수행을 통해 성능 저하를 없애는 '모니터링' 절차가 필요하다.

모니터링 업무의 목적은 실 운영단계에서 확보되는 데이터를 이용한 재학습 · 모델 최신화 활동 수행이다. 주요 과업은 Fig. 3에 제시된 바와 같이 최초 전력화 이후 데이터의 수집 · 관리, 데이터의 검수 · 정제 · 가공, 데이터 분류, 모델 재학습, 성능 평가, AI 모델의 재배포 6가지로 구분된다.

모델 검증 단계 까지는 실운용환경에서의 데이터를 이용한다는 점만 제외하고는 개발단계 성능평가 단계에서 데이터를 이용한 단위 기능 평가를 수행하는 점에서 유사하다고 할 수 있다. 모델의 검증(최신화)이 이루어진 후 에는 그 결과를 체계 운용 부대, 소요부대, 개발기관 등 유관 기관에 피드백을 수행한다. 피드백의 범위로는 단순 성능 개선 수치 뿐만 아니라 모델의 유효성을 검증하는 업무의 일환으로 개발, 양산 단계에서 활용할 수 있는 데이터 환류, 모델 개선 필요 방향, 현 모델의 한계점 등 여러 요소들이 포함된다.

2.4 모니터링 업무 실증 사례

본 절에서는 운용유지 단계 모니터링 업무의 필요성, 타당성, 개선 필요사항 등을 확인하기 위하여 실제 군에서 운용 중인 특정 AI 기술이 적용된 OOO 체계를 대상으로 모니터링 시범 적용을 수행한 내용을 제시하고자 한다. 모니터링 업무 중 '데이터 품질 점검과 분류와 관련된 기술 지원(Case 1)'과, 점검 및 분류가 완료된 데이터를 이용하여 '모델 재학습, 성능 평가를 통한 모델 검증 및 피드백 과정(Case 2)'으로 크게 두 가지 과업 범위로 나눠서 사례를 제시하고자 한다. 체계명과 성능 지표 및 데이터 수량 등의 구체적인 수치는 보안 관계 상 가림처리하거나 다른 표기법(ex. 성능지표 α, β 등)으로 제시하였다.

2.4.1 데이터 품질 점검 및 분류 기술지원(Case 1)

수요군의 AI 기술 적용 OOO 체계 데이터 품질 점검 및 분류 기술지원 요청에 의거 '22년~'23년 대상 객체(A/B/C)들의 배율 별 데이터 수량 점검 및 대상 상황에 대한 세분류 과업을 수행하였다.

Table 1. Data Quality Control Indicator (Data Compatibility Part)

Classification	Indicator	Description
Data Compatibility	criterion compatibility	Metrics that measure diversity, reliability, sufficiency, and factuality in order to select criteria for whether the secured data is suitable for learning purposes
	Technical suitability	a measure of file format, resolution, clarity, color, size, length, sound quality, etc. to technically determine whether the secured data is suitable for learning purposes
	statistical diversity	an indicator that measures class distribution diagrams, instance distribution diagrams, sentence length, vocabulary, etc. to prevent data bias

이 데이터 품질 점검 과업 수행을 통해 데이터 셋의 구성, 분포를 확인할 수 있었으며 본 연구 저자의 소속기관에서 제정한 'AI 적용 무기체계 데이터 품질관리 매뉴얼' 내 데이터 적합성-통계적 다양성 지표(Table 1)에 대한 확인을 수행했다고 볼 수 있다. AI 모델의 성능지표 달성 확인을 위하여 학습 모델의 훈련용/검증용/시험용 데이터셋으로 구분 분류하는 것이 필요하다. 본 사업에서도 학습데이터셋(훈련용/검증용)은 업체에 제공하고, 시험데이터셋은 모델의 학습에 사용되지 않도록 별도 분리 보관하여 모델의 성능을 검증하는 것이 필요하다는 취지하에 데이터 분류 기술지원을 수행하게 되었다. 데이터셋 구성 비율은 학습데이터와 시험데이터가 약 8 : 2가 되도록 전체 데이터 수량을 기준으로 분류를 수행하였다.

Fig. 4. Classification of small data (example)

Fig. 4에 소규모 데이터 분류(예시) 그림을 제시하였는데 랜덤 샘플링 수행 시 학습 데이터와 시험 데이터의 상관관계가 높다고 판단되는 경우(개발 체계의 운용환경상 데이터의 수량이 적거나 유사 데이터 그룹이 많은 경우)는 관련기관 협의 하에 특정 범위를 시험 데이터로 활용하는 절차가 필요하다. 본 사업에서도 동일한 환경에서 취득된 유사 데이터의 수량이 많아 단순 랜덤 샘플링을 수행하게 된다면 중복 데이터들로 인하여 성능지표가 과하게 높게 나타날 가능성이 식별되었다. 따라서 특정 환경 구역에서 취득된 데이터들을 시험 데이터셋으로 분류하여 모델 성능 척도를 확인하도록 협의하여 진행하였다.

분류가 완료된 데이터는 별도의 PC에 분리 보관하고 학습이 완료된 후 로그를 확인하여 시험데이터가 사용되지 않았는지를 다시 한 번 확인하였다.

2.4.2 모델 재학습, 성능 평가를 통한 모델 검증 및 피드백 과정(Case 2)

Table 2에 제시된 바와 같이 AI가 적용된 OOO 체계 사업을 통해 3차에 걸쳐 개선된 학습 모델이 탑재되어 납품되었고, 납품 시점 이후에 운용부대에서 취득된 실 운용 환경 데이터를 대상으로 정제/가공 업무 등을 수행하여 4차 학습을 진행하였다. 모델의 성능 개선 정도를 확인하는 절차는 기존 사업 간 진행되었던 성능 평가 방식과 동일하게 고정된 시험데이터를 이용하여 모델의 성능지표 α와 성능지표 β을 비교하는 방식이다. 만약 시험데이터가 변수가 된다면 모델간의 정확한 성능 비교가 어렵기 때문에, 1~4차 모델 합산 시험데이터를 기준으로 수행하였다.

Table 2. Training and Test Data Set Configuration

Traning Model	Training Data		Test Data	
	Quantity	Total	Quantity	Total
1st	000	000	000	000
2nd	000	000	000	000
3rd	000	000	000	000
4th (Monitoring pilot application)	000	000	000	000

Table 3. Results of performance improvement comparison (1st ~ 4th comparison)

Class	Training Model	Test Results (Based on 1st to 4th Test Data)			
		Performance metric α		Performance metric β	
		IR	CCD	IR	CCD
A	1st	00.00	00.00	00.00	00.00
	2nd	00.00	00.00	00.00	00.00
	3rd	00.00	00.00	00.00	00.00
	4th	00.00	00.00	00.00	00.00
B	1st	00.00	00.00	00.00	00.00
	2nd	00.00	00.00	00.00	00.00
	3rd	00.00	00.00	00.00	00.00
	4th	00.00	00.00	00.00	00.00

Fig. 5. Class A Performance Comparison Results (Performance Metric α(Top) & β(Bottom))

Table 3과 Fig. 5, 6에 실사업 대상 성능 평가 수행 결과를 제시하였다. 동일한 시험데이터를 기준으로 학습 모델의 성능을 IR/CCD로 구분하여 확인하였다. 체계에 적용되는 객체들 중 특정 객체 A, B에 대해서 모니터링 업무를 수행하였다.

객체 A의 경우 4차 학습 모델에서 성능지표 A와 B 모두 미세하게 성능이 개선되는 것을 확인하였다. 다만, 객체 A 대상 성능 비교 결과에서 IR 대비 CCD의 성능이 떨어지는 결과를 확인할 수 있었는데, 객체 A를 인식하는 CCD, IR의 자체 성능에서의 차이로 인해 발생했을 수 있다. 만약 기본 자체 성능 차이가 아닌 학습과 관련된 이슈라면 향후 CCD 데이터 추가 확보 및 모델 재학습을 통한 개선이 필요할 수도 있을것으로 사료된다.

객체 B의 경우 4차 학습 모델에서 거의 성능이 동일하거나, 오히려 미세하게 감소하는 경향을 나타내었다. 3차 학습 모델에서 성능이 가파르게 증가한 것을 보아 사업 초기 제공된 객체 B 데이터 대비 3차 데이터 확보 환경(시범 운용 기간)에서의 객체 B 데이터 비중이 매우 높았을 것으로 생각되며, 4차 모델이 3차 환경의 객체 B 데이터에 과적합(Overfitting)되었을 가능성도 있다고 생각된다. 향후 3차 데이터 구성을 재확인하고 실 운용 환경 취득 데이터를 늘리는 등 전반적인 데이터 재구성 업무 수행이 필요할 것으로 판단된다.

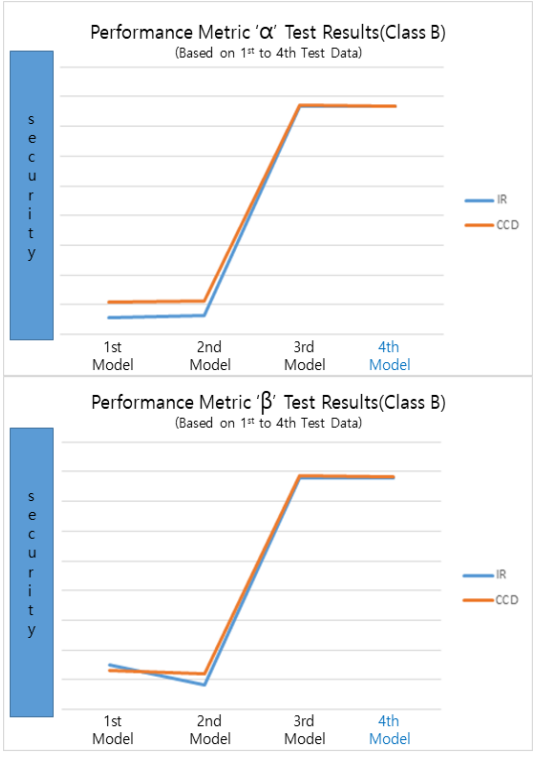

Fig. 6. Class B Performance Comparison Results (Performance Metric α(Top) & β(Bottom))

3. 결론

본 논문에서는 AI 적용 무기체계의 운용유지단계에서의 성능 확인을 위한 단위기능 평가 및 시스템 평가로 구분되는 성능 평가와 실운용 데이터를 활용한 모니터링 절차에 대한 실증 연구 결과를 제시하였다. 모델 성능 비교를 통한 AI 모니터링 실사업 적용 가능성을 확인할 수 있었다. 또한, 체계 AI 모델의 보완이 필요한 점을 식별하여 수요군에 환류하고, 유사 AI 사업에 대한 평가 사례와 향후 소요 제기와 관련된 정책 제도적 제언을 제공하였다.

다만, 본 시범 적용 간 한정된 지역에서 데이터 확보가 수행되었고 모니터링 업무 수행을 위한 독립예산이 미 편성되어 있어 체계적 업무 수행 불가하다는 등의 제한사항이 있었다. 따라서, 향후 업무의 정책적 제도 정립을 위하여 모니터링 대상 선정, 계획 수립, 분석 주기 등에 대한 후속 연구를 지속 수행할 계획이다. 향후 실제 정식적인 업무 절차가 정립된다면 실운용 환경에서 품질 개선(성능 고도화)을 통해 국방 전력 증강 및 실 운용부대의 무기체계 활용능력 확대가 가능할 것으로 생각된다. 또한, 개발, 양산 단계로 모니터링 업무 결과를 지속 환류하여 후속 AI 획득 사업 간 성능이 개선되고 AI 운용 범위가 확장되는 등의 효과를 기대할 수 있을 것이다.

References

[1] Y. M. Lee, "A Study on the Performance Enhancement of AI-Applied Weapon Systems at Operation and Sustainment Stage", Proceedings of The Korea Institute of Military Science and Technology Conference, 2023.

[2] ISO/IEC, "ISO/IEC 25010 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models", 2011.

[3] J. H. Kwak, "Data quality and reliability of artificial intelligence systems", TTA Journal, Vol. 201. 2022. (https://www.tta.or.kr/)

[4] DoD(U.S. Department of Defense), "Adopts 5 Principles of Artificial Intelligence Ethics", 2020.

[5] DoD(U.S. Department of Defense), "Responsible Artificial Intelligence Strategy and Implementation Pathway," Policy Report, the U.S. Department of Defense(DoD), United States of America, Jun. 2022.

[6] H. G. Kim, "Need and Status of Defense Artificial Intelligence Quality Management", DQS Magazine. (https://www.dtaq-media.kr/)

[7] ISO/IEC, "ISO/IEC TR 29119-11 Software and systems engineering — Software testing — Part 11:Guidelines on the testing of AI-based systems", Technical Report, ISO/IEC, Nov. 2020.

이 영 민(Yeong-Min Lee) [정회원]

- 2016년 8월 : 금오공과대학교 전자공학부 (공학사)
- 2020년 8월 : 금오공과대학교 전자공학과 (공학석사)
- 2021년 7월 ~ 현재 : 국방기술품질원(DTaQ) 연구원

〈관심분야〉
국방, 전자공학, 전파공학, 인공지능

강 지 훈(Ji-Hoon Kang) [정회원]

- 2013년 2월 : 경상대학교 전자공학과 (학사)
- 2015년 8월 : 경상대학교 전자공학과 (석사)
- 2016년 8월 ~ 2019년 7월 : 한국산업기술시험원(KTL) 연구원
- 2019년 8월 ~ 현재 : 국방기술품질원(DTaQ) 연구원

〈관심분야〉
국방, 전자공학, 인공지능, 시험평가

정 민 경(Min-Kyung Jeong) [정회원]

- 2020년 2월 : 부산대학교 나노메카트로닉스공학과 (공학사)
- 2019년 12월 ~ 현재 : 국방기술품질원(DTaQ) 연구원

〈관심분야〉
국방품질경영, 전자회로, 반도체 공정기술

박 주 영(Joo-Young Park) [정회원]

- 2015년 2월 : 숭실대학교 전기공학과 (공학사)
- 2019년 1월~ 현재 : 국방기술품질원 연구원

〈관심분야〉
국방, 전기, 전자, 인공지능, 사이버보안