Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2023.46.4.229

Vol. 46, No. 4 : 229-237, December 2023
ISSN : 2005-0461 / EISSN : 2287-7975

Test and Evaluation Procedures of Defense AI System linked to the ROK Defense Acquisition System

Yong-Bok Lee† · Min-Woo Choi · Min-ho Lee

Korea National Defense University

국방획득체계와 연계한 국방 인공지능(AI) 체계 시험평가 방안

이용복† · 최민우 · 이민호

국방대학교 국방과학학과

In this research, a new Test and Evaluation (T&E) procedure for defense AI systems is proposed to fill the existing gap in established methodologies. This proposed concept incorporates a data-based performance evaluation, allowing for independent assessment of AI model efficacy. It then follows with an on-site T&E using the actual AI system. The performance evaluation approach adopts the project promotion framework from the defense acquisition system, outlining 10 steps for R&D projects and 9 steps for procurement projects. This procedure was crafted after examining AI system testing standards and guidelines from both domestic and international civilian sectors. The validity of each step in the procedure was confirmed using real-world data. This study's findings aim to offer insightful guidance in defense T&E, particularly in developing robust T&E procedures for defense AI systems.

Keywords : Defense Test and Evaluation(T&E), Data-based Performance Evaluation Procedure, Defense AI System

1. 서 론

사회의 다양한 분야에 인공지능(AI)이 활용되면서 인공지능(AI)에 대한 충분한 시험평가 부족 등의 이유로 오류가 발생하거나, 사회적·윤리적인 위협이 등장하고 있다.

테슬라 자동차는 눈이 오지 않는 온화한 지역에서 충분히 시험할 수 없었던 강설 조건에서는 자율주행 기능이 비정상적으로 작동하는 오류가 있었다[30]. 최근에는 가상공간에서 무인기의 인공지능(AI) 기능을 시험할 때 인공지능(AI) 알고리즘 오류로 무인기를 통제하는 사람을 공격한 사례도 있었다. 이러한 인공지능(AI)의 오류는 전투력을 다루는 국방 분야에서는 민간 분야보다 심각한 물질적·인적 피해를 일으킬 수 있다.

인공지능(AI)이 적정 수준의 성능과 품질을 확보하기 어려운 이유는 데이터의 수량과 품질 확보 문제, 알고리즘의 효율성과 취약성 문제, 인공지능(AI)과 인간 또는 인공지능(AI)과 다른 시스템간의 상호운용성 문제 등을 들 수 있다.

특히 국방 인공지능(AI) 체계는 이러한 문제 이외에 운용환경, 개발환경, 시험평가 측면에서 민간 분야에 비해 제한사항이 많으며 세부내용은 다음과 같다.

첫째, 운용환경 측면에서 국방 인공지능(AI) 체계가 활용되는 전투환경은 인공지능(AI)의 기술적 한계가 두드러지게 나타날 수 있는 열악한 조건이다. 인공지능(AI)는 학습 데이터에서 경험하지 못한 잡음(Noise), 특이값(Outlier) 등 입력값의 변화가 많은 전투환경에서는 추론 결과의 신뢰성이 낮아질 가능성이 높다. 예를 들면, 민간 자율주행 차량의 인공지능(AI)는 도로 표식의 색깔과 종류, 신호, 포

Received 9 November 2023; Finally Revised 24 November 2023; Accepted 27 November 2023
† Corresponding Author : miliman@naver.com

장된 노면상태 등 정제된 환경에서 차량 제어를 판단한다. 그러나, 군용 자율주행 차량의 인공지능(AI)은 비포장도로, 인원과 차량이 혼재된 주행환경, 도로파손 등이 우발적으로 빈번하게 발생하는 조건에서 차량 제어를 판단해야 하므로 신뢰성을 유지하기가 매우 어렵다.

둘째, 개발환경 측면에서 국방 인공지능(AI) 체계는 성능이 검증된 알고리즘을 활용하여 군 요구사항에 적합한 데이터를 학습하여 튜닝하는 절차로 개발한다. 그러나 인공지능(AI) 학습을 위한 데이터가 군사보안 문제 등으로 공개가 제한되는 경우가 대부분이다[30].

셋째, 시험평가 측면에서 인공지능(AI)의 성능은 특정 시기에 완성되지 않고, 전체 수명주기(Total life cycle) 기간 동안 축적되는 학습 데이터의 변화에 따라 성능이 변화하는 특징이 있다. 따라서 특정 시기에 성능이 완성되어 전투용 적합 여부를 판단하는 일반적인 무기체계와는 다른 시험평가 절차가 필요하지만, 현재까지 제도적으로 정립된 절차는 없다. 한편 인공지능(AI) 성능 결정의 주요 요소인 인공지능(AI) 알고리즘, 데이터 품질, 컴퓨터 성능에 대해 군 내부적으로는 전문적인 시험 및 검증 인프라가 없으며, 군 외부적으로도 국방 인공지능(AI) 체계에 대한 성능시험 성적서를 발급할 수 있는 공인기관이 없다.

본 연구에서는 이러한 제한사항을 해소하기 위해 국방 인공지능(AI) 체계 시험평가 방안을 국방획득체계와 연계하여 제안했다. 국방 인공지능(AI) 체계는 단독으로 획득되기 보다는 'AI기반 무인응급처치체계' 등의 형태로 현재 무기체계 획득절차를 적용하기 때문에 국방획득체계와 연계하여 연구가 필요하다. 제안한 방안은 일반적인 무기체계 시험평가의 핵심인 실물에 의한 현장 시험평가 이전에 인공지능(AI) 모델의 성능을 독립적으로 시험할 수 있는 데이터 기반 성능평가를 선행하는 개념과, 이를 구현하기 위한 세부절차이다. 실물에 의한 현장 시험평가는 국방 시험평가 분야에서 오랜 기간 동안 충분한 경험을 통해 절차가 잘 정립되어 있기 때문에 현재의 절차를 적용한다. 데이터 기반 성능평가 절차는 국방획득체계의 사업유형에 따라서 연구개발사업과 구매사업으로 구분했다.

본 연구에서 제안한 개념과 절차는 실제 데이터를 기반으로 실증하면서 정제하였다.

2. AI 시스템 시험 관련 문헌연구

국방 인공지능(AI) 체계의 시험평가는 매우 중요하다. 이러한 시스템들은 높은 신뢰성, 안정성, 보안을 요구한다. 따라서 인공지능(AI) 시스템을 평가할 때에는 성능 지표, 시뮬레이션, 보안, 국제 규정 및 기준 준수, 데이터 관리 등을 고려해야 한다[25]. 본 연구의 주제인 국방 인공지능(AI) 체계 시험평가와 관련된 시험평가 절차, 성능평가 지표, 시뮬레이션 기반 시험평가, 데이터 품질 측면의 연구동향을 살펴본다.

먼저, 성능평가 지표와 관련된 연구를 살펴보면 다음과 같다. 인공지능(AI)의 성능을 정확하게 측정하고 평가할 수 있는 지표를 개발하는 것이 중요한 연구 주제로 부상하고 있다. 특히, 국방 인공지능(AI) 체계는 임무를 정확하고 신뢰성 있게 수행할 수 있어야 하기 때문에 인공지능(AI) 체계 시험평가 연구가 더욱 필요하다. 그러나 현재 인공지능(AI) 체계의 시험평가 절차는 아직 제도적으로 확립되지 않았다. 이러한 문제의식을 바탕으로 일부 연구에서는 제도적 발전의 필요성을 지적하고, 운용 환경에 적합한 데이터 세트의 구축 및 유지 보수, 군 인공지능 영상 감시 체계의 특성을 고려한 시험평가 절차와 인공지능 모델의 특성과 전체 수명주기를 고려한 시험평가 제도를 개념적으로 제안했다[4, 18].

다음은 성능평가 척도 연구로, 2022년에는 인공지능(AI) 기반 감시 시스템 사례를 통해 인공지능(AI) 시험평가 대상 체계의 본질적인 성능을 정확하게 평가할 수 있는 성능평가 척도 적용의 필요성을 사례 중심으로 제시하였다. 이 연구는 성능 지표 관련 용어의 해석 차이를 줄이기 위해 용어를 명확히 정의하고, 모델 요구사항에 부합하는 성능지표를 제안했다[21].

다음으로 시뮬레이션 관련 연구는 다음의 이유로 중요하다. 인공지능(AI) 시스템이 실제 군사 작전에 배치되기 전에는 다양한 실제 조건과 시나리오를 반영한 시뮬레이션 검증이 필수적이다. 이 과정은 인공지능(AI) 모델이 훈련 데이터에만 특화되는 과적합을 방지하고, 모델의 일반화 능력을 평가하는 데 있어 필수적인 단계이다. 국방 분야에서 인공지능(AI) 시스템의 시험평가를 위한 시뮬레이션 활용에 대한 구체적인 사례는 현재까지 명확히 제시되지 않았지만, 시스템의 안정성과 신뢰성을 확보하기 위한 시뮬레이션 평가의 필요성은 지속적으로 강조되고 있다[3, 24]. 특히, 시뮬레이션을 통해 모델을 학습시키는 인공지능(AI) 개발 절차에서는 이러한 평가가 더욱 중요하다[5, 20]. 최근에는 국내의 한 소프트웨어 기업이 자율주행 무기체계 시험평가 시뮬레이터를 개발하여 시뮬레이션에서 인공지능(AI) 무기체계의 시험평가 가능성을 입증하였다[9]. 또한, 시뮬레이션을 통한 시험평가는 자율주행 자동차 개발과 산업용 로봇의 자동화 연구에서도 활발히 이루어지고 있기 때문에 국방 분야에서의 적용 가능성이 점차 증가하고 있다[6, 17]. 이와 같은 연구들은 시뮬레이션을 통한 시험평가가 실제 환경에서의 결과와 유사한 효율성을 가질 수 있음을 입증하고 있다.

마지막으로, 데이터 품질은 인공지능(AI) 모델의 성능에 결정적인 영향을 미친다. 이는 국방, 의료, 자율주행 차

량, 드론 등 다양한 분야에서 인공지능(AI)의 성능에 직접적으로 기여하는 요소이다[2]. 인공지능(AI) 시스템의 발전과 함께 데이터 품질의 중요성이 강조되면서 데이터 품질 측정, 요구사항 관리에 대한 국제적인 기준이 제시되고 있다. 특히, 다양한 산업 및 제품에 대한 국제표준을 설정하는 조직인 ISO와 IEC[1]는 5259와 25024 문서를 통해 데이터 품질 평가 기준을 제시하고 있다.

ISO/IEC 5259는 인공지능 분석 및 머신러닝을 위한 데이터 품질과 관련된 용어정의 및 예시를 비롯한 인공지능(AI) 데이터 품질에 대한 개념적 이해와 데이터 품질 측정, 요구사항 관리 및 데이터 처리 프레임워크를 제공한다[2, 10, 11, 12, 13, 14, 27]. ISO/IEC 5259는 자동차의 자율주행 시스템(ADS, Autonomous Driving System) 표준 프레임워크 개발시 참조되었으며[26], 머신러닝에서 데이터 품질을 평가하는데 사용하였다[28].

ISO/IEC 25024는 소프트웨어 공학과 시스템 공학의 품질 요구사항과 평가(SQuaRE, Software Quality Requirements and Evaluation)에 대한 국제 표준 시리즈인 ISO/IEC 25000 시리즈의 일부이다. 이 시리즈는 소프트웨어 제품의 품질을 측정하고 평가하는 데 사용되는 모델, 용어, 정의 및 절차를 제공한다[30]. 특히, 데이터의 다양한 특성에 맞춘 정량적 품질 측정 기준을 제공한다. 이는 데이터 수명주기를 통틀어 적용되며, 데이터 품질 측정 방법과, 데이터 품질 요구사항 정의를 위한 지침을 명확히 한다. 구조화된 데이터에 적용 가능한 이 표준은 데이터 관리자, 개발자, 유지보수 담당자, 품질 관리자 등 데이터 품질 측정이 필요한 모든 이해 관계자에게 유용하다[9]. 이 표준은 정부가 공개하는 다양한 데이터 세트의 품질을 평가하는데 효과적으로 사용할 수 있는 가능성이 확인되었으며[10], 이 표준을 기초로 개발한 데이터 평가 프레임워크가 제안되었다[1]. 국내에서도 데이터 품질과 관련된 정부의 노력으로 인공지능 학습용 데이터 품질관리 가이드라인을 제시하였다[7, 22]. 가이드라인에서는 인공지능 학습용 데이터를 구축함에 있어서 구축계획 수립단계부터 데이터획득/수집, 정제, 가공 등에 대한 절차, 산출물 및 품질관리 활동을 제시하고 있다[23, 29].

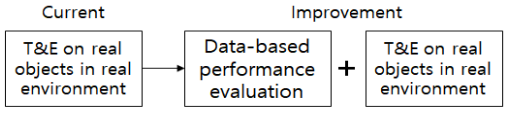
지금까지 인공지능(AI) 시스템의 신뢰성과 안정성 보장을 위한 정확한 성능지표 사용의 필요성, 시뮬레이션 검증 방법 적용의 중요성, 인공지능(AI) 성능에 결정적인 영향을 미치는 데이터 품질과 관련된 표준 및 가이드라인에 대해 알아보았다. 이처럼, 인공지능(AI) 체계의 시험평가에 관한 연구들이 분야별 필요성과 요구에 맞춰 연구되고 있다. 그러나 국방획득체계와 연계하여 국방 인공지능(AI) 체계의 시험평가 절차를 데이터를 기반으로 실증해서 구체적으로 제시한 사례는 없었다.

3. 국방 인공지능(AI) 체계 시험평가 개념

국방전력발전업무훈령에 따르면 시험평가 목적은 협의적으로는 시제품의 성능이 시험평가 기준을 충족하는지 여부를 확인하는 것이지만, 광의적으로는 시제품의 성능이 동일한 규격으로 양산되어 전력화되었을 때 전장 환경에서 유사하게 발휘될 수 있는 일반화(Generalization) 가능성을 확인하는데 있다. 시험평가 절차는 이러한 목적달성을 위해 실물에 의한 현장평가를 원칙으로 한다.

그러나 국방 인공지능(AI) 체계를 일반 무기체계와 같이 현장에서 실물에 의한 현장평가만 하는 것은 인공지능(AI)이 전투환경에서 직면할 수 있는 다양한 상황을 충분히 조성하는 것이 곤란하기 때문에 시험평가의 충분성 측면에서 부족하다.

이를 해소하기 위해 Lee et al.(2022)은 <Figure 1>과 같이 데이터 기반 성능평가 이후에 실물에 의한 현장 시험평가를 수행하는 개념을 제안했다[7]. 이 개념에서 현장 시험평가는 현재 잘 정립되어 있는 일반 무기체계 시험평가 절차를 적용하면 된다.



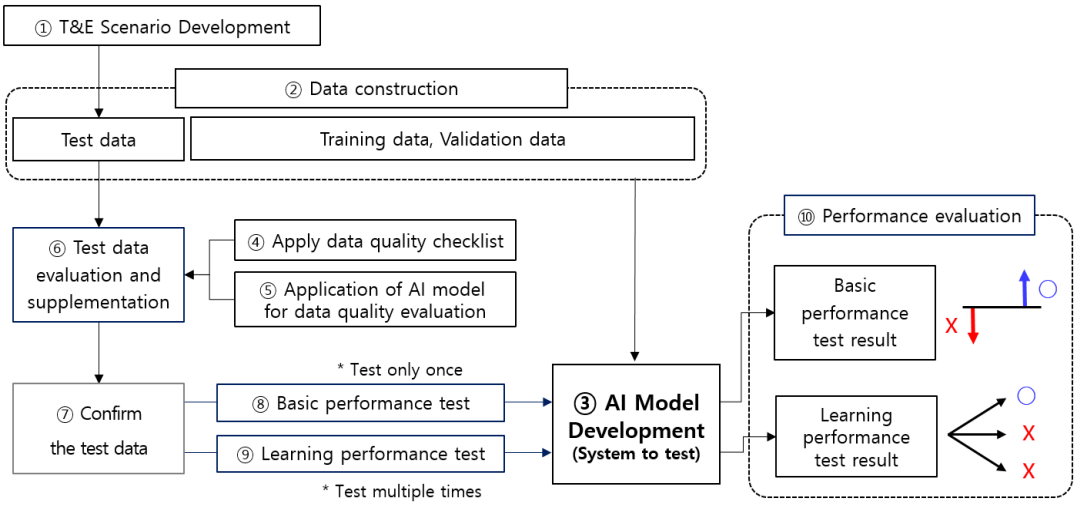
<Figure 1> Defense AI System T&E Concept

일반 무기체계 시험평가 절차를 인공지능(AI)에 적용할 경우, 인공지능(AI)의 고유한 특성인 학습에 따른 성능변화 또는 인공지능(AI) 시스템의 복잡성을 충분히 평가하는 것이 곤란할 수 있다. 또한, 시험 결과의 기준 충족 또는 미충족의 원인이 인공지능(AI)과 다른 하부 시스템과 연계되서 복합적으로 나타나기 때문에 원인 파악이 곤란할 수 있다. 그러나, 이 절차는 오랜 기간 동안 실무적으로 검증되었기 때문에 시험평가의 구조적 효율성과 일관성을 제공하여 인공지능(AI) 시험평가를 위한 별도의 인프라 구축의 필요성을 줄일 수 있는 장점이 있다.

그러나 데이터 기반 성능평가는 개념만 제시되어 있기 때문에 정의와 세부 절차에 대한 정립이 필요하다. 따라서 본 연구에서는 국방 인공지능(AI) 시스템의 시험평가 관점에서 데이터 기반 성능평가와 실물에 의한 현장 시험평가의 개념을 다음과 같이 정의하였다.

첫째, 데이터 기반 성능평가는 소요제기서에 명시된 인

1) ISO: International Organization for Standardization), IEC: International Electrotechnical Commission.



<Figure 2> Data-based Performance Evaluation Procedure in AI System R&D Project

공지능 모델(AI Model)에 대한 요구 성능 충족 여부를 주 체계와 분리된 인공지능(AI) 모델을 대상으로 데이터를 이용하여 평가하는 절차로서 정확도(Accuracy), 재현율(Recall) 등 인공지능(AI) 성능평가 척도를 사용한다.

둘째, 실물에 의한 현장 시험평가는 소요제기서에 명시된 인공지능 체계(AI aided system)에 대한 요구 성능 충족 여부를 데이터 기반 성능평가를 통해 성능이 확인된 인공지능(AI) 모델이 탑재된 체계를 대상으로 시험평가 현장에서 평가하는 절차로서 인공지능(AI) 모델과 다른 구성품 성능이 융합된 성능을 평가할 수 있는 탐지율(Detection rate), 오경보율(False alarm rate) 등의 척도를 사용한다.

국방획득체계는 연구개발사업과 구매사업으로 구분한다. 따라서 본 연구에서는 데이터 기반 성능평가 절차를 두 가지 경우로 구분하여 제안한다.

4. 연구개발사업에서 데이터 기반 성능평가 절차

데이터 기반 성능평가를 위해서는 시험 데이터 품질의 완전성을 점검해야 한다. 시험 데이터 품질에 따라 인공지능(AI) 모델 성능평가 결과의 신뢰도가 결정되기 때문이다. 이를 위해 본 연구에서는 '시험평가 시나리오'와 '시험 데이터 품질 체크리스트', '시험 데이터 품질 평가용 인공지능(AI) 모델'을 활용하여 시험 데이터의 품질 향상방안을 제안했다. 세부 절차는 <Figure 2>와 같이 10단계로 구성되어 있다.

① 시험평가 시나리오 개발 단계에서는 소요결정 관련 문서·운영개념서·교범 등의 문헌을 연구하여 시험평가 대상 체계의 환경 조건과 발생 가능한 이벤트를 목록화한다. 현실에서는 모든 발생 가능한 경우에 대해 데이터 수집과 시험평가를 할 수 없다. 따라서 환경 조건과 이벤트를 각각 데이터 수집 가능성, 발생 빈도, 시험평가 자원 가용성 등을 고려하여 <Table 1>의 예시처럼 필수 조건과 개별 조건으로 분류한다. <Table 1>로부터 필수 환경 및 이벤트 조건에 해당하는 요인(Factor)과 수준(Level)을 조합하여 인공지능(AI) 모델 개발 및 시험 데이터 구축 범위를 한정하기 위한 시험평가 시나리오를 개발한다.

<Table 1> Example of Factor Grouping for a T&E Scenario Development

Condition	Factor	Levels
essential environment condition (8 items)	time (4 items)	dawn, daytime, evening, night
	season (4 items)	spring, summer, fall, winter
	- skipping -	
	filming distance (2 items)	standard distance, long distance
essential event condition (9 items)	no. of people (2 items)	1 person, 2 people
	- skipping -	
	act (11 items)	concealment, infiltration, stop, escape, walk, etc.
individual condition (5 items)	object size(5 Level), magnification(3 Level), etc.	

② 데이터 구축 단계에서는 시나리오를 기초로 데이터를 수집한다. 데이터 수집은 일반적인 인공지능(AI) 개발 과정과 같다. 이 단계에서는 인공지능(AI) 모델에 대한 데이터 기반 성능평가를 위해 수집한 전체 데이터에서 시험 데이터(Test data)를 분리한다. 시험 데이터를 분리하는 방

법은 Hold-Out Method, K-fold Cross Validation 등이 있다[22, 23]. 이해관계자가 많은 국방 시험평가의 특수성을 고려 시 국방 분야에서 데이터 분리방법은 인공지능(AI) 개발용 데이터와 시험용 데이터가 독립적으로 분리될 수 있는 Hold-Out Method가 적절하다. Hold-Out Method를 적용할 때는 일반적으로 통용(Rule of thumb)되는 '훈련 데이터(Training data):검증 데이터(Validation data):시험 데이터=8:1:1 비율[22]' 등을 적용하되 전체 데이터의 분포와 국방 환경의 특수성을 고려해서 이해관계자가 동의하는 합리적인 데이터 분리 방법을 적용하는 것이 타당하다.

③ 인공지능(AI) 모델 개발 단계에서는 학습 및 검증 데이터를 이용해서 인공지능(AI) 모델을 개발한다. 현재 국방 분야에서 인공지능(AI) 개발은 알고리즘을 자체 개발하기보다는 Image Detection 분야의 YOLO 알고리즘 등 국제적으로 성능이 검증된 알고리즘에 운용목적에 적합한 데이터를 학습하는 일반적인 절차를 따른다. 따라서 본 연구에서는 세부 절차를 제안하지 않는다. 다만, 장기적으로는 인공지능(AI) 알고리즘에 종속되지 않기 위해 자체 개발할 수 있는 역량을 갖출 필요가 있다.

④ 데이터 품질 체크리스트 적용 단계에서는 시험 데이터의 품질 수준을 다양한 관점에서 점검할 수 있는 체크리스트를 개발하여 시험 데이터에 적용한다. 체크리스트는 '2. 인공지능(AI) 시스템 시험 관련 문헌연구'에서 제시한 문헌 연구결과를 종합하여 시험평가 대상 AI 체계에 적합하도록 개발한다. 다수 문헌에서 식별한 체크리스트에 포함되어야할 공통적인 요소는 <Table 2>와 같이 체크리스트 분류 기준, 체크리스트 항목, 확인 시기, 확인 주체, 확인 방법 등이 있다.

<Table 2> Example of Checklist Items

Item	Description
checklist classification criteria	classify the checklist and assign a classification number
checklist item	describe the purpose, method, and target of each checklist item in a clear sentence.
check time	check time for each checklist item
check performer	classified as tester or testee for each checklist item
check type	Classified as quantitative or qualitative method

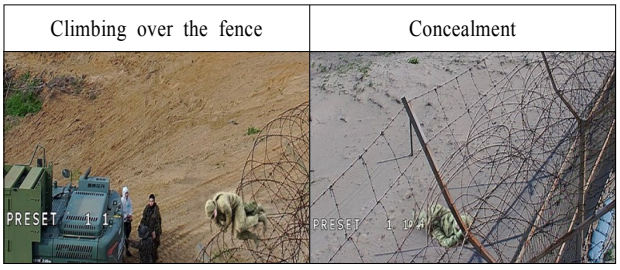
체크리스트 항목은 예를 들면, 데이터 출처 신뢰성을 점검하기 위해 '수집한 데이터의 출처는 신뢰할 수 있는가?'라는 질문에 대해 ISO_25024에서는 <Table 3>에서처럼 데이터가 신뢰할 수 있는 출처로부터 수집된 비율(Degree to which values are provided by a qualified organization)을 계산하는 수식을 이용해서 점검하는 것을 권장하고 있다[30].

<Table 3> Example of Data Source Reliability Checklist from ISO 25024

$X = A/B$
A = number of data values provided or validated/certified by a qualified organization
B = number of data values for which source credibility can be defined

⑤ 데이터 품질 평가용 인공지능(AI) 모델 적용 단계에서는 시험평가 대상 체계가 아닌 성능이 검증된 인공지능(AI) 알고리즘을 이용해서 시험 데이터의 유효성 측면에서 품질을 평가한다. 이 단계에서는 인공지능(AI) 유형별(분류, 탐지, 인식, 질의응답, 문서요약 인공지능(AI) 등) 성능과 인공지능(AI) 모델의 유연성, 신뢰성, 공평성을 적절히 평가할 수 있도록 시험 데이터가 구성되었는지 여부를 평가한다. 유효성 평가 방법은 성능이 검증된 인공지능(AI) 모델에 시험 데이터를 적용하여 적절한 성능(Accuracy, AUC, Fβ-Score 등)이 발휘되는지 여부를 평가하는 방법, MS-COCO 등 잘 알려진 데이터셋과 시험 데이터를 동일한 인공지능(AI) 모델에 적용한 결과를 비교하는 방법 등이 있다.

⑥ 시험 데이터 평가 및 보완 단계에서는 체크리스트와 인공지능(AI) 모델에 의한 데이터 품질 평가 결과를 기초로 시험 데이터의 품질 완성도를 높인다. 이 단계에서는 체크리스트를 통해 군사보안・개인정보보호법 준수를 위한 비식별화처리 결과 보완, 라벨링 결과 수정 등을 통해 데이터 구축간 전처리한 결과를 보완한다. 또한, 체크리스트와 인공지능(AI) 모델을 통해 식별한 완전성(Completeness), 다양성(Diversity), 대표성(Representativeness) 등 데이터 품질 측면에서 미흡한 결과를 보완한다. 예를 들어, 대표성 측면에서 시험평가 시나리오와 시험 데이터의 클래스(Class)와 인스턴스(Instance) 분포를 비교하여 부족한 경우 데이터 보완소요를 도출한다. 데이터 보완은 시험평가 대상 인공지능(AI) 체계가 일반적이지 않은 군사적 환경에서 운용되는 특성을 고려하여 실제 환경에서 추가 수집을 원칙으로 하되 안전・군사보안 등의 이유로 실제 환경에서 수집이 제한되는 경우에는 영화세트장과 같은 유사 환경을 조성하여 데이터를 수집하거나 <Figure 3>에서처럼 합성 데이터(synthetic data)를 생성하여 보완한다.



<Figure 3> Example of Synthetic Data Generation

⑦ 시험 데이터 확정 단계에서는 시험평가조직, 해당 체계 소요제기기관 등 군 내부의 이해관계자들이 시험 데이터 품질 완전성과 신뢰성 등에 대해 검토 및 협의하여 시험 데이터를 확정한다. 이 단계에서는 특히 <Table 4>와 같이 이전 단계에서 데이터 추가 수집, 합성 데이터 생성 등에 의한 시험데이터 변화에 따른 인공지능(AI) 모델 성능의 변화 정도의 적절성을 확인하는데 중점을 둔다.

<Table 4> Example of Comparison before and after Supplementation of Test Data

Metric		Test data supplementation	
		Before	After
Diversity	time characteristics	1 class (daytime)	2 classes (daytime, evening,)
	outfit characteristics	3 classes (civilian uniform, combat uniform, ghillie suit)	4 classes (civilian uniform, combat uniform, ghillie suit, poncho raincoat)
	behavioral characteristics	5 classes (stop, escape, crawl, walk, conceal)	6 classes (concealment, stop, escape, crawl, walk, conceal)
	- omitted below -		
effectiveness	F1-Score	0.93	0.93
	Precision	0.99	0.95
	Recall	0.90	0.93

⑧ 기본성능 시험 단계에서는 인공지능(AI) 모델의 군 요구성능 충족여부를 소요제기서에 명시된 정확도, Fβ-Score 등 인공지능(AI) 성능평가 척도를 사용하여 시험한다. 기본성능 시험은 인공지능(AI) 모델이 시험 데이터를 학습하지 않도록 특정 시기에 1회만 수행한다.

⑨ 학습성능 시험 단계에서는 데이터 품질에 따라 성능이 변하는 인공지능(AI)의 특성을 고려하여 ROC curve, PR curve 등 학습능력 평가에 적합한 척도를 사용하여 시험평가 기간 중 여러 번 시험한다.

⑩ 성능평가 단계에서는 인공지능(AI) 모델의 기본성능과 학습성능 시험결과를 종합적 평가한다. 기본성능 평가는 인공지능(AI) 모델의 군 요구 성능 충족여부를 기준 충족 또는 기준 미 충족으로 평가한다. 학습성능 평가는 군 요구 성능 충족여부와 별개로 인공지능(AI) 모델의 성능 변화 추세를 판단한다. 인공지능(AI) 모델은 학습용 데이터 변화와 모델 튜닝(tuning) 수준에 따라 성능이 향상되거나, 변하지 않거나, 성능이 저하되는 3가지 경우로 구분된다. 따라서 학습 성능 평가결과는 성능향상 가능 또는 성능향상 불가능으로 평가한다. 마지막으로 기본성능과 학습 성능 시험 결과를 종합적으로 평가하여 실물에 의한 현장 시험평가 진입여부를 판단한다. 평가결과는 <Table 5>와 같이 4개의 경우로 구분한다.

<Table 5> Performance Evaluation Results

Test result		Enter the next stage*
Basic performance	Learning performance	
satisified	possible to improve performance	entrance
	impossible to improve performance	conditional entrance (need to check generalizability)
unsatisfied	possible to improve performance	conditional entrance (need to build data)
	impossible to improve performance	not entrance

* on-site T&E using the actual AI system.

첫째, 기본성능이 기준을 충족하고, 학습을 통한 성능향상이 가능한 경우에는 실물에 의한 현장 시험평가로 진입이 가능하다.

둘째, 기본성능은 기준을 충족했지만, 학습을 통한 성능향상이 불가능한 경우에는 인공지능(AI) 모델이 데이터에 대해 과적합(Over fitting)되어 있을 가능성이 높다. 따라서 실물에 의한 현장 시험평가 단계에서 충분한 시험을 통해 일반화(Generalization) 가능성을 확인한다.

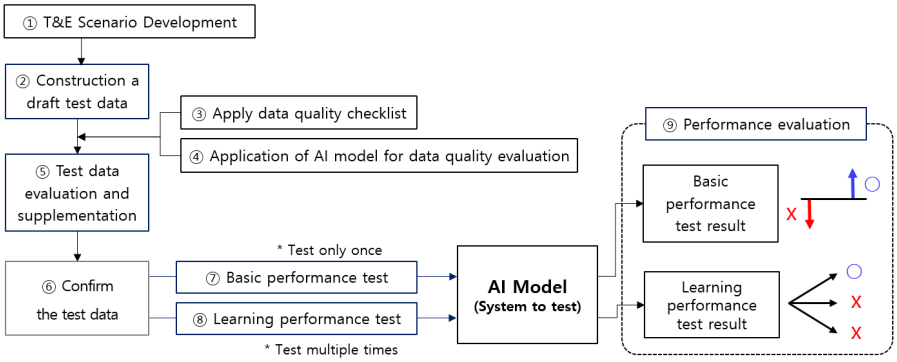
셋째, 기본성능이 기준을 충족하지 못했지만, 학습을 통한 성능향상이 가능한 경우에는 군사보안 등의 이유로 인공지능(AI) 모델 학습 및 검증 데이터의 수량과 범주가 충분하지 않을 가능성이 높다. 따라서 실물에 의한 현장 시험평가 진입 전까지 추가 데이터 수집이 가능한 경우에는 조건부 진입으로 평가한다. 추가 데이터 수집이 불가능한 경우에는 실물에 의한 현장 시험평가로 진입이 불가능한 것으로 평가한다.

넷째, 기본성능과 학습 성능이 모두 기준을 충족하지 못한 경우에는 실물에 의한 현장 시험평가로 진입이 불가능한 것으로 평가한다.

상기의 10단계 절차를 실제 데이터에 적용하여 데이터의 품질 향상 정도를 확인한 결과 인공지능(AI) 모델 성능이 <Table 6>과 같이 향상 되었다.

<Table 6> Performance Results in Procedure in AI System R&D Project

Contents	F1-Score	Precision	Recall
Before	0.74	0.77	0.72
After	0.93	0.99	0.9



<Figure 4> Data-based Performance Evaluation Procedure in AI System Purchasing Project

4. 구매사업에서 데이터 기반 성능평가 절차

연구개발사업은 인공지능(AI) 개발을 위한 데이터를 구축해서 개발한 인공지능(AI) 체계를 시험평가 하지만, 구매사업은 구매한 완성된 인공지능(AI) 체계를 시험평가 한다. 데이터 기반 성능평가 측면에서는 인공지능(AI) 학습 및 검증 데이터 구축 여부와 인공지능(AI) 체계 개발 여부의 차이가 있다.

따라서 구매사업의 데이터 기반 성능평가는 연구개발사업과 많은 분야가 동일하다. 본 절에서는 구매사업의 데이터 기반 성능평가 절차를 <Figure 4>와 같이 9단계로 제안하고, 연구개발사업과 개념적으로 차이가 있는 시험 데이터 초안 구축 단계에 대해서만 기술한다.

① 시험평가 시나리오 개발과, ③ 데이터 품질 체크리스트 적용부터 ⑩ 성능평가 단계는 연구개발사업과 동일하다.

② 시험 데이터 초안 구축 단계에서는 시험평가 시나리오를 기초로 시험 데이터를 구축한다. 시험 데이터는 군이 시험 데이터를 보유하고 있는 경우와 보유하지 않은 경우로 구분한다.

시험 데이터를 보유하고 있는 경우에는 해당 데이터의 시험평가 대상 인공지능(AI) 체계의 운용개념에 대한 적합성을 확인하여 보완 여부를 판단한다. 시험 데이터를 보유하고 있지 않은 경우에는 인공지능(AI) 모델의 과적합을 방지하기 위해 인공지능(AI) 학습 및 검증 데이터와 독립된 출처로부터 시험 데이터를 수집하고, 공개하지 않는다.

5. 결론 및 향후 연구방향

5.1 결론

현재의 국방 시험평가는 실물에 의한 현장 시험평가 중심으로 발전되어 있다. 그러나 인공지능(AI)는 추론과정의 불확실한 Black box적인 특성 때문에 인공지능(AI) 모델의 신뢰성을 충분히 시험할 수 있는 별도의 절차가 필요하지만, 구체적으로 정립되지 않았다.

본 연구에서는 이러한 제한사항을 해소하기 위해 국방 인공지능(AI) 체계 시험평가 절차를 제안했다. 제안한 개념은 인공지능(AI) 모델의 성능을 독립적으로 시험할 수 있는 데이터 기반 성능평가 절차를 선행한 후에 실물에 의한 현장 시험평가를 수행한다. 데이터 기반 성능평가 절차는 국방획득체계의 사업추진 유형으로 구분하여 연구개발사업은 10단계, 구매사업은 9단계로 제안했다. 특히 그동안 인공지능(AI) 시험 분야에서 상대적으로 덜 중요하게 다루어진 시험 데이터의 품질 완전성 향상 절차를 구체화했다.

본 연구에서 제안한 절차는 국내외 민간 분야에서 제시한 인공지능(AI) 체계 시험 관련 표준과 가이드라인을 광범위하게 분석하여 개발했으며, 실제 데이터를 기초로 각 단계의 적절성을 검증했다. 이러한 과정으로 데이터 기반 성능평가 후에 SW 신뢰성 시험 수행에 대한 고려가 필요하다.

본 연구에서 제안한 개념과 절차는 실제 데이터를 기반으로 실증하면서 정제했기 때문에 국방 인공지능(AI) 시험평가 절차 정립이 필요한 국방 분야에 유용한 방향성을 제공할 수 있을 것이다.

5.2 향후 연구방향

본 연구결과는 국방 인공지능(AI) 체계 시험평가에 대한 전략적 수준에서 방향이 결정된 것을 가정했다. 향후에는 본 연구결과를 방법론으로 활용할 수 있는 국방 인공지능(AI) 체계 시험평가에 대한 전략과 프레임워크, 제도개선 방안에 대한 연구를 수행할 예정이다.

제안한 절차를 실증하기 위한 데이터 처리하는 과정은 복잡하고 방대하다. 따라서 본 연구에서는 지면의 제한으

로 전체적인 절차를 제시하는데 중점을 두고 각 절차 중에서 핵심적인 사례만 포함했다. 향후에는 제안한 절차를 기초로 세부 과정을 기술적인 측면에서 구체적으로 제시할 예정이다.

References

[1] Calabrese, J., Esponda, S., and Pesado, P.M., Framework for Data Quality Evaluation Based on ISO/IEC 25012 and ISO/IEC 25024, VIII Conference on Cloud Computing, Big Data & Emerging Topics (Modalidad virtual, 8 al 10 de septiembre de 2020), 2020.

[2] Chang, W., ISO/IEC JTC 1/SC 42 (AI)/WG 2 (data) data quality for analytics and machine learning (ML), 2022, Information Technology Laboratory.

[3] Cho, K.T., Lee, S.Y., Lee, H.M., Kim, S.H., and Jeong, H.M., Enhancing the Efficiency and Reliability for M&S based Test and Evaluation System Development, *Journal of the Korea society for Simulation*, 2012 21.1 pp. 89-96.

[4] Cho, Q., Han, M., Ryu, H., Lee, J., and Kim, S., A Study on Development of Test and Evaluation Methods for AI-based Image Surveillance Systems in Defense, *Journal of Applied Reliability*, 2023 Vol. 23, No. 1, pp. 97-104.

[5] Choi, M., Choi, E., Song, Y., Kim, J., Park, S.T., Kwon, O., and Cho, N., Simulation Based Reinforcement Learning for the Intelligence Behavior of Autonomous Weapon System, *Journal of the Korea Society For Simulation*, 2023, Vol. 32, No. 2, pp. 91-111.

[6] Clegg, A., Wijmans, E., Lee, S., Savva, M., Chernova, S., and Batra, D., Sim2real Predictivity: Does Evaluation in Simulation Predict Real-World Performance?, *IEEE Robotics and Automation Letters*, 2020, Vol.5, No. 4. pp. 6670-6677.

[7] Géron, A., Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, H. S. Park Ed., *Republic of Korea: Hanbit Publishing Network*, 2020, pp. 61.

[8] https://www.etnews.com/20231012000315.

[9] https://www.iso.org/standard/35749.html.

[10] ISO/IEC JTC 1/SC 42/WG2, 《ISO/IEC WD -5259-5:202X(E) (Artificial intelligence-Data quality for analytics and machine learning(ML)-Part 5: Data quality governance framework≫, 2022.

[11] ISO/IEC JTC 1/SC 42/WG2, 《ISO/IEC WD 5259-1:202X(E) (Artificial intelligence-Data quality for analytics and machine learning(ML)-Part 1: Overview, terminology, and example≫, 2022.

[12] ISO/IEC JTC 1/SC 42/WG2, 《ISO/IEC WD 5259-2:202X(X) (Artificial intelligence-Data quality for analytics and machine learning(ML)-Part 2: Data quality measures≫, 2022.

[13] ISO/IEC JTC 1/SC 42/WG2, 《ISO/IEC WD 5259-3:20##(X) (Artificial intelligence-Data quality for analytics and machine learning(ML)-Part 3: Data quality management requirements and guidelines≫, 2022.

[14] ISO/IEC JTC 1/SC 42/WG2, 《ISO/IEC WD 5259-4:202#(E) (Artificial intelligence-Data quality for analytics and machine learning(ML)-Part 4: Data quality process framework≫, 2022.

[15] ISO/IEC JTC 1/SC 7, 《ISO/IEC 25024:2015 Systems and software engineering: Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of data quality≫, 2015.

[16] Kim, B.-S. and Yun, K., A study on the test evaluation of automated-vehicles based on artificial intelligence, Spring Conf, of KSAE, 2019, pp. 722-722.

[17] Kim, B.S. and Yun, K., A Study on the Test Evaluation of Automated-Vehicles Based on Artificial Intelligence, *The Korean Society of Automotive Engineers, Annual Spring Conference*, 2019, pp. 722-722.

[18] Kim, M.Y. and Noh, S.C., Acquisition Process of Military Weapon System Using Artificial Intelligence, *Proceedings of Symposium of the Korean Institute of Communications and Information Sciences*, 2022.

[19] Kohavi, R., A study of cross-validation and bootstrap for accuracy estimation and model selection, in *Proc. 14th Int. Joint Conf. on Artif. Intell.*, 1995, pp. 1137-1143.

[20] Kurrek, P., Jocas, M., Zoghlami, F., Stoelen, M., Salehi, V., Ai motion control–a generic approach to develop control policies for robotic manipulation tasks, *Proceedings of the Design Society: International Conference on Engineering Design*, 2019, Vol. 1, No. 1. Cambridge University Press.

[21] Lee, Y., Jeon, I., and Kim, S., A Study on the Development of Artificial Intelligence Weapon System Test and Evaluation Method: Focusing on the Performance Evaluation of the Classification Model, *Reliability Applied Research*, 2022, Vol. 22, No. 1, pp. 1-9.

[22] Ministry of Science and ICT, Korea Information and Communication Technology Association, Trustworthy Artificial Intelligence Development Guide (Plan), 2023.

[23] National Information Society Agency (Vol 3.0), Republic

of Korea: Guide for Constructing Artificial Intelligence Training Data, 2023.

[24] Park, J.H., A Study on the V&V Process of M&S for the Test and Evaluation, *Journal of the Korea Academia-Industrial Cooperation Society*, 2019, Vol. 20, No. 9, pp. 397-404.

[25] Porter, D.J. and Dennis, J.W., Test & evaluation of ai-enabled and autonomous systems: *A literature review, Institute for Defense Analyses*, 2020.

[26] Priestley, M., O'Donnell, F., and Simperl, E., A survey of data quality requirements that matter in ML development pipelines, *ACM Journal of Data and Information Quality*, 2023, Vol. 15, No. 2, pp. 1-39.

[27] Rangineni, S., An Analysis of Data Quality Requirements for Machine Learning Development Pipelines Frameworks, *International Journal of Computer Trends and Technology*, 2023, Vol. 71, No. 9, pp. 16-27.

[28] Schnelle, S. and Favaro, F. M., ADS Standardization Landscape: Making Sense of its Status and of the Associated Research Questions, arXiv preprint arXiv, 2023, 2306.17682.

[29] Shin, J.H., Data quality verification method for artificial intelligence learning, Journal of Electronic Engineering, 2021, Vol. 48, No. 7, pp. 28-34.

[30] Torchiano, M., Vetrò, A., and Iuliano, F., Preserving the benefits of Open Government Data by measuring and improving their quality: An Empirical Study, *IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, 2017, Vol. 1. IEEE.

ORCID

Yong-Bok Lee | https://orcid/org/0000-0002-0338-9977
Min-Woo Choi | https://orcid.org/0000-0002-2699-3579
Min-ho Lee | https://orcid.org/0000-0003-4549-8933