

원저

N2SF 환경의 데이터 공유 제약 완화를 위한 하이브리드 비식별화 및 sLLM 기반 합성데이터 활용 방안 연구

오진수

경기북부경찰청 고양경찰서

교신저자: 오진수 (js523@police.go.kr)

요약

최근 국방·방산 분야는 국가망 보안체계(N2SF) 도입을 계기로, 데이터의 민감도에 따라 보안 등급을 차등화하는 다층보안체계(MLS)로의 전환을 추진하고 있다. 이러한 위험 관리 기반 통제는 국가안보 관점에서 필수적이나, 국방 R&D 및 산학연 협력에 필요한 데이터의 외부 공유를 구조적으로 제한해 '데이터 사일로(Data Silo)' 현상을 심화시킬 수 있다. 본 연구는 N2SF의 엄격한 반출·접근 통제 환경에서도 교육 및 연구 목적의 데이터 활용을 지원할 수 있는 데이터 중심 보안(Data-Centric Security)의 실증적 대안으로서, 하이브리드 비식별화와 온프레미스 생성형 AI 기반 합성데이터(Synthetic Data) 기술의 활용 가능성을 검토하였다. 첫째, 원본 사이버 위협 로그 내 민감정보를 보호하기 위해 정규표현식(Regex)과 개체명 인식 모델(NER)을 결합한 하이브리드 비식별화 전처리 파이프라인을 구축하였으며, 평가 표본에 사전에 주입한 식별 대상 195건을 모두(100%) 탐지·마스킹하였다. 둘째, N2SF 고보안 구역 내 외부 API를 통한 정보 유출 위험을 줄이기 위해 폐쇄망에서 구동 가능한 온프레미스 소형 거대언어모델(sLLM, Llama-3 8B)과 통계 기반 SDV(Synthetic Data Vault)를 활용한 합성데이터 생성 환경을 설계하였다. 셋째, 미국 로스앨러모스 국립연구소(LANL)의 이기종 사이버 보안 로그를 활용하여 총 100,000건의 방산 프로시 데이터셋을 구성하고 유용성을 평가하였다. 데이터셋의 정상:위험 클래스 비율은 19:1로 설정하였다. XGBoost 기반 TSTR(Train Synthetic, Test Real) 검증 결과, 실제 데이터로 학습한 TRTR(Train Real, Test Real) 모델의 F1-Score는 0.9848이었으며, sLLM 기반 합성데이터로 학습한 TSTR 모델의 F1-Score는 0.9367로 나타나 기준 모델 대비 95.1%의 성능 보증을 보였다. 본 연구는 프로시 데이터셋을 활용한 실험을 통해 N2SF 환경에서 데이터 공유가 제한되는 경우 하이브리드 비식별화와 합성데이터를 결합한 데이터 활용 방안의 기술적 적용 가능성을 검토하였다. 다만 데이터의 보안등급 하향 및 외부 반출 여부는 이러한 기술적 검증과 별도의 제도적·보안적 절차에 따라 판단되어야 함을 논의하고, 향후 국방 AI 연구 생태계 조성을 위한 정책적·제도적 개선 방향을 제언한다.

핵심어

방산안보, 국가망 보안체계(N2SF), 다층보안체계(MLS), 데이터 중심 보안, 하이브리드 비식별화, 합성데이터, SDV, sLLM

차례

1. 서론
2. 이론적 배경 및 N2SF 보안 체계
3. 비식별화 적용 및 합성데이터 생성 실험 설계
4. 합성데이터 성능 평가 결과(Performance Evaluation)
5. 결론 및 향후 연구 과제

Open Access

접수일: 2026년 8월 14일
수정일: 2026년 9월 21일
게재승인일: 2026년 9월 24일
출판일: 2026년 9월 30일

Copyright © 2026 Author(s).

This is an Open Access article distributed under the terms of the Creative Commons CC BY 4.0 license (<https://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Original Article

A Study on Hybrid De-Identification and sLLM-Based Synthetic Data for Mitigating Data Sharing Constraints in N2SF Environments

Jinsu Oh

Goyang Police Station, Gyeonggi Bukbu Provincial Police Agency, Republic of Korea

Corresponding Author: Jinsu Oh (js523@police.go.kr)

ABSTRACT

Recently, the defense and defense industry sectors have been pursuing a transition to a Multi-Layered Security System (MLS) that differentiates security grades based on data sensitivity, triggered by the introduction of the National Network Security System (N2SF). While such risk management-based controls are essential from a national security perspective, they can exacerbate the "Data Silo" phenomenon by structurally restricting the external sharing of data necessary for defense R&D and industry-academia-research collaboration. This study examined the feasibility of hybrid anonymization and on-premise generative AI-based synthetic data technologies as an empirical alternative for Data-Centric Security that can support data utilization for educational and research purposes, even within the environment of N2SF's strict export and access controls.

First, to protect sensitive information within original cyber threat logs, a hybrid anonymization preprocessing pipeline combining Regular Expression (Regex) and Named Entity Recognition (NER) models was constructed, and all 195 identification targets in the evaluation sample (100%) were detected and masked. Second, to reduce the risk of external API leakage within the N2SF high-security zone, we designed a synthetic data generation environment centered on an on-premise, small large language model (sLLM, Llama-3 8B) capable of operating on a closed network and a statistics-based SDV (Synthetic Data Vault). Third, we evaluated the utility of defense proxy data constructed from 100,000 heterogeneous cyber security logs (class imbalance 1:19) from the U.S. Los Alamos National Laboratory (LANL).

As a result of machine learning (XGBoost)-based TSTR (Train Synthetic, Test Real) validation, the F1-Score of the synthetic training model (TSTR) was recorded at 0.9367 compared to the F1-Score of the original training model (TRTR) at 0.9848, achieving a performance retention rate of 95.1%. Based on a proxy dataset, this study confirms the technical applicability of opening and utilizing defense data under the N2SF framework, discusses that downgrading and export decisions require institutional grounds separate from technical verification, and proposes directions for policy and institutional improvements to foster a defense AI research ecosystem in the future.

KEYWORDS

Defense Security, National Network Security System (N2SF), Multilayer Security System (MLS), Data-Centric Security, Hybrid De-identification, Synthetic Data, SDV, sLLM

1. 서론

1.1. 연구의 배경 및 필요성

4차 산업혁명 핵심 기술의 국방 유입과 인공지능(AI) 기반 지능형 무기체계의 고도화는 현대전의 패러다임을 경계 중심에서 ‘데이터 중심전(Data-Centric Warfare)’으로 전격 전환시키고 있다.[1] 데이터가 국가 안보와 국방력의 핵심 자산으로 부상함에 따라, 대규모 국방·방산 데이터에 대한 지속적인 AI 모델 학습과 산학연 협력 기반의 빅데이터 분석 역량 확보는 방위산업 경쟁력의 필수 요건이 되었다.[2]

그러나 국내 방위산업 보안체계는 장기간 물리적 망분리에 기반한 경계 중심 보안(Perimeter-based Security) 모델을 고수해 왔다. 이러한 방식은 외부의 악의적 침입 차단에는 일정 부분 기여하였으나, 연구개발(R&D) 및 교육·학술 목적의 데이터 외부 공유를 구조적으로 단절시켰다. 그 결과, 각 군 기관 및 방산 기업 내부망에 데이터가 고립되는 ‘데이터 사일로(Data Silo)’¹⁾ 문제가 심화되어 AI 기반 국방 기술 발전의 발목을 잡고 있다.[3-5]

최근 정부는 이러한 경계 중심 통제의 한계를 극복하고, 안보와 활용의 조화를 도모하기 위해 ‘국가망 보안체계(N2SF)’를 수립하였다. 2025년 N2SF 가이드라인에 따르면, 데이터의 중요도에 따라 보안 등급을 기밀(C), 민감(S), 공개(O) 등급으로 차등 통제하는 다층보안체계(MLS)로의 전환이 본격화되었다.[6]

이러한 환경에서 기밀(C) 및 민감(S) 등급 방산 데이터를 학술 연구용인 공개(O) 영역으로 다운그레이드(Downgrade)하여 개방하기 위해서는, 데이터 자체에 보안 속성을 내재화하는 데이터 중심 보안(Data-Centric Security) 기술이 필수적이다. 전통적인 가명화·마스킹 기법은 유용성 손실이 크고 역추론 위협에 취약하여 「개인정보 보호법」의 법적 요건을 완전히 해소하기 어렵다. 따라서 원본의 통계적 및 문맥적 특성을 복제하면서 개체 식별 가능성을 낮춰 법적 안전성 확보에 기여할 수 있는 ‘합성데이터(Synthetic Data)’ 기술의 실증 연구가 시급한 실정이다.[7-9]

한국인터넷진흥원(KISA)의 N2SF 시범사업 추진 등 새로운 보안 패러다임 실증이 전개되는 현 시점에서[10,11], 데이터 중심 보안(Data-Centric Security) 기술의 확보는 필수적이다.

1.2. 연구 목적 및 범위

본 연구는 국가망 보안체계(N2SF) 다층보안체계(MLS) 환경 하에서 발생하는 방산 데이터의 엄격한 반출·접근 통제 제약을 극복하고, 국방 AI 연구 및 교육 생태계 활성화를 도모하기 위한 기술적 대안을 제시하는 데 주된 목적이 있다. 이를 위해 데이터 중심 보안(Data-Centric Security) 관점에서 민감 정보의 잔존을 최소화하는 ‘하이브리드 비식별화’ 기법과 외부 API 유출 위험을 줄이는 ‘sLLM 기반 합성데이터 프레임워크’를 제안하고, 프록시 데이터셋을 기반으로 그 기술적 유효성과 분석적 가치를 정량적으로 평가하고자 한다.

이러한 연구 목적을 달성하기 위한 구체적인 연구 범위와 세부 추진 내용은 다음과 같다.

첫째, N2SF 환경의 데이터 중심 보안 요건을 분석한다. N2SF 가이드라인 및 「개인정보 보호법」 등 관련 법·제도적 규제 환경 하에서, 기밀(C) 및 민감(S) 등급 방산 데이터를 공개(O) 영역으로 안전하게 다운그레이드(Downgrade)하기 위해 요구되는 기술적·운용적 요건을 고찰한다.

둘째, 방산 사이버 환경을 대리하는 Multi-Source 프록시 데이터셋을 구축한다. 실제 방산 사

1) 데이터가 각 부서, 시스템 또는 망(Network)별로 격리되어 다른 조직이나 부서와 공유·융합되지 못하고 고립되어 있는 상태.

이러 통제망의 복잡한 행위 특성과 공격 시그니처를 고해상도로 모사하기 위하여, 미국 로스앨러모스 국립연구소(LANL)의 이기종 사이버 로그(인증, 프로세스, 네트워크 흐름, DNS)와 레드팀(RedTeam) 위협 데이터를 통합한 대규모 프록시 데이터셋을 구축한다.

셋째, 하이브리드 비식별화(Regex + NER) 기법을 실증한다. 고정된 규칙 기반 정규표현식(Regex)이 지닌 미탐지(False Negative)의 한계를 극복하고자, 문맥 기반의 개체명 인식(NER) 모델을 결합한 하이브리드 비식별화 전처리 기법을 구현하여 평가 표본 내 식별 대상에 대한 탐지 성능을 검증한다.

넷째, AI 모델을 통한 합성데이터의 유용성을 검증한다. N2SF 폐쇄망 환경에서 구동 가능한 온프레미스 소형 거대언어모델(sLLM, Llama-3 8B) 및 통계 기반 SDV(Synthetic Data Vault)를 활용해 합성데이터를 생성하고, 머신러닝 기반 TSTR(Train Synthetic, Test Real) 평가 체계를 통해 원본 데이터 대비 성능 보존율과 연구·교육 목적의 대체 가능성을 정량적으로 평가한다.

2. 이론적 배경 및 N2SF 보안 체계

2.1. N2SF 체계와 데이터 중심 보안을 통한 방산안보

물리적 망분리 환경에서는 모든 데이터가 획일적인 경계 통제에 간혀 외부 공유가 원천적으로 불가능했다. 반면, N2SF(National Network Security Framework) 기반 다층보안체계(MLS)는 정보의 중요도에 따른 위협 관리 기반 통제를 제공한다.

표 1. 기존 물리적 망분리와 N2SF(다층보안체계) 비교

구분	물리적 망분리 (AS-IS)	N2SF 다층보안체계 (TO-BE)
보안 패러다임	경계 중심 보안 (Perimeter-based)	데이터 중심 통제 (Data-Centric, MLS)
적용 방식	일괄적인 인터넷망 / 업무망 물리적 분리	데이터 민감도에 따른 등급(C, S, O)별 차등 통제
기밀(C) 통제	전면 차단 (망 단절)	최고수준 통제 (물리적 단절 유지)
민감(S) 통제	전면 차단 (망 단절)	제한적 공유 (접근통제, 가명·합성데이터 활용)
공개(O) 통제	불가능 (내부 자료 반출 제한)	공공 클라우드 및 생성형 AI 자유로운 연동·활용
데이터 활용성	매우 낮음 (데이터 고립 심화)	높음 (안보와 활용의 균형 추구)

본 연구에서는 N2SF를 특정 데이터의 보안등급을 기술적으로 자동 하향하거나 외부 반출을 자동 허용하는 제도로 해석하지 않는다. 하이브리드 비식별화와 합성데이터 생성은 데이터 자체의 정보노출 위험과 활용성 사이의 균형을 검토하기 위한 기술적 보호수단으로 제안되며, 실제 등급 조정 및 외부 반출 여부는 해당 기관의 보안정책과 승인 절차 및 별도의 제도적 근거에 따라 판단되어야 한다.[11,12]

2.2. 전통적 비식별화의 한계와 개인정보 보호법 이슈

기존 비식별화 과정에서 활용되는 마스킹 및 k-익명성 등의 기법은 데이터의 세부정보를 축약하거나 변환하는 과정에서 분석 유용성이 감소할 수 있다.[7] 더욱이 방위산업은 참여 주체가 제한적인 특수 도메인이므로, 비식별화된 데이터라도 외부 OSINT(공개출처정보)와 결합 시 원본 개체가 역추론되는 결합 공격(Linkage Attack)에 취약하다. 개인정보가 포함된 로그의 경우에는 「개인정보 보호법」상 가명정보·익명정보 관련 요건과 재식별 가능성을 함께 검토할 필요가 있다. 따

라서 비식별화 또는 합성데이터 생성 자체를 법적 안전성의 완결된 요건으로 보아서는 안 되며, 데이터셋과 처리목적, 결합 가능성 및 재식별 위험 등을 종합적으로 평가할 필요가 있다.[7]

2.3. 하이브리드 비식별화 및 합성데이터의 혁신성

이러한 한계를 타개하기 위해 본 연구는 두 가지 혁신 기술을 융합한다.

첫째, 하이브리드 비식별화(Regex 정규식 + NER 모델) 기술이다. 사이버 위협 로그에는 표준 규칙(IP, 이메일) 외에도 변칙적 조직명, 서버명, 인명 등 고유 식별자가 혼재한다. 단일 정규식(Regex)의 탐지 한계를 극복하기 위해, 자연어 문맥을 이해하는 BERT 기반 개체명 인식(NER) 모델을 결합함으로써 정규식 단독 적용 시 발생하는 미탐지(False Negative)를 줄일 수 있다.

둘째, 폐쇄망 sLLM 기반 합성데이터 생성이다. 합성데이터(Synthetic Data)는 원본 데이터셋과 1:1로 매칭되는 실제 정보 주체가 존재하지 않아 개인정보 노출 위험을 낮출 수 있으나, 「개인정보 보호법」상 규제 적용 여부는 재식별 가능성 등에 대한 별도의 평가를 통해 판단되어야 한다.[7] 통계 기반 모델(SDV)[13] 및 CTGAN 기법[14]과 더불어, N2SF 폐쇄망 환경에서 구동 가능한 Llama-3 8B 모델[15]을 활용할 경우 로그의 문맥적 표현을 활용한 합성데이터 생성 가능성을 검토할 수 있다.

3. 비식별화 적용 및 합성데이터 생성 실험 설계

3.1. N2SF 데이터 개방(Downgrade) 파이프라인

N2SF 다층보안체계 하에서 실측(Real) 방산 데이터를 합성데이터로 변환하여 반출하는 파이프라인은 다음과 같이 설계된다.

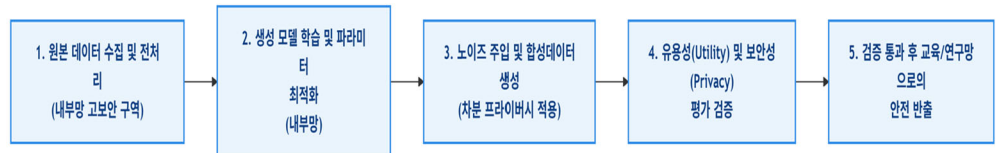


그림 1. N2SF 다층보안체계 합성데이터 변환 반출 파이프라인

Step 1은 원본 데이터 수집(내부망 고보안 구역) 단계로, N2SF 내부망에서 다차원(Multi-Source) 로그를 수집·병합한다.

Step 2는 전처리(하이브리드 비식별화) 단계로, Regex와 NER 모델을 통해 식별 정보를 탐지·마스킹한다.

Step 3은 생성 모델 학습 단계로, 비식별화 데이터를 기반으로 SDV 및 sLLM(Llama-3)을 학습한다.

Step 4는 합성데이터 생성 단계로, 원본 주체와 1:1로 대응되지 않는 모사 데이터를 생성한다.

Step 5는 유용성(TSTR) 검증 및 반출 단계로, 유용성 및 보호 관련 검토 후 교육·연구 목적의 제공 가능성을 검토한다.

3.2. 분석 데이터셋(LANL Proxy Dataset) 활용

표 2. 실험 환경 및 데이터셋 명세 (Methodology & Dataset)

항목	상세 명세
총 레코드 수	총 100,000건 - 학습용 79,000건 - 평가용 21,000건 (정상 20,000건, 위협 1,000건)
정상(Normal) 클래스	95,000건 (95.0%)
위협(RedTeam) 클래스	5,000건 (5.0%)
불균형 비율	19 : 1 (정상:위협, 실험 데이터 구성 기준)
추출 피처 (8차원)	범주형: src_user, dst_user, src_comp, dst_comp, process_name, dns_query 수치형: pkt_count, byte_count
생성 모델 구현	PyTorch, HuggingFace Transformers, On-premise Llama-3 8B

N2SF 보안 규정상 원본 국방 데이터의 외부 반출이 불가능한 구조적 제약을 극복하기 위하여, 데이터 구조와 변수 속성이 방산 사이버 보안 분야와 유사한 미국 로스앨러모스 국립연구소(LANL)²⁾³⁾[16,17]의 이기종 사이버 보안 로그를 연구 프록시(Proxy) 데이터셋으로 선정하였다.

데이터 전처리 파이프라인을 통해 auth, proc, dns, flows 로그와 redteam 위협 로그를 결합하여, 다양한 사이버 보안 로그의 특성을 반영할 수 있도록 총 100,000건의 실험용 프록시 데이터셋을 구성하였다.

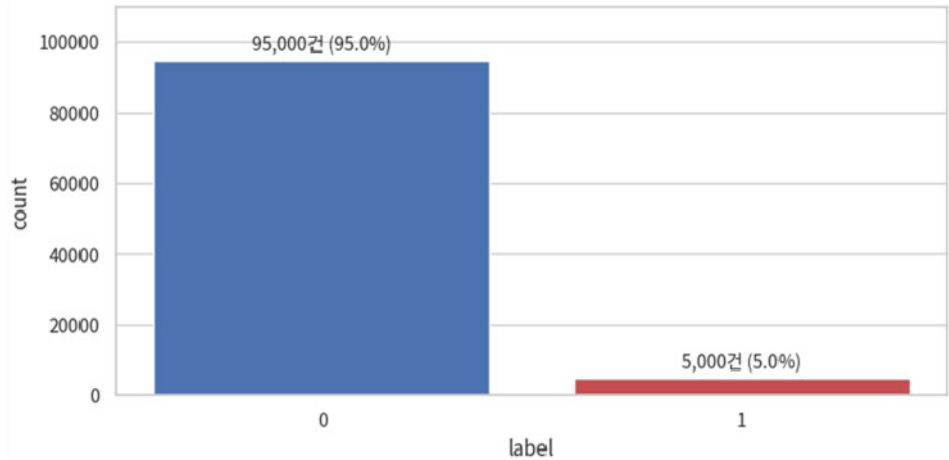


그림 2. LANL 사이버 로그 데이터셋 클래스 분포. Normal vs RedTeam.

실제 사이버 보안 환경의 특성을 반영하여 정상 데이터가 95%를 차지하는 불균형(Imbalanced) 환경을 조성하였다. 위협(RedTeam) 데이터에는 특정 C2 서버 통신(c2-server.net) 및 이상 패킷 크기와 같은 명확한 공격 시그니처가 포함되어 있다.

- 2) A. D. Kent, "Cybersecurity Data Sources for Dynamic Network Research," in Dynamic Networks in Cybersecurity, 2015.
- 3) A. D. Kent, "Comprehensive, Multi-Source Cybersecurity Events," Los Alamos National Laboratory, <http://dx.doi.org/10.17021/1179829>, 2015.

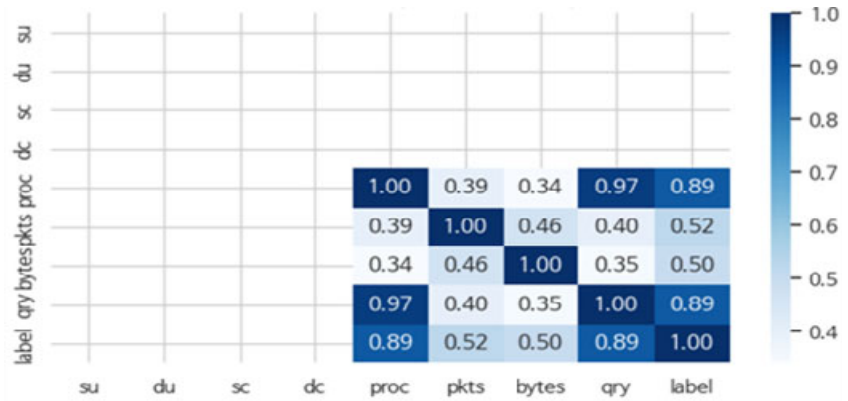


그림 3. 다차원 피쳐 세트 간 피어슨 상관 관계 히트맵

구축한 10만 건의 LANL 방산 프록시 데이터셋에 대하여, 비식별화 및 인코딩 이후 주요 피쳐 간의 통계적 관련성을 규명하고자 피어슨 상관관계 분석(Pearson Correlation Analysis)을 수행하였다.

상관관계 분석 결과를 정밀 검토해보면, 종속 변수인 사이버 위협 여부(label)는 프로세스 명칭(proc, $r=0.89$) 및 DNS 쿼리(qry, $r=0.89$)와 강한 양의 상관관계를 형성하고 있으며, 네트워크 패킷 수(pkts, $r=0.52$) 및 전송 바이트 수(bytes, $r=0.50$)와도 유의미한 중등도의 상관성을 나타냈다.

이러한 수치적 결과는 본 연구가 도출한 두 가지 중요한 학술적 의의를 내포한다.

첫째, 기존 데이터 시뮬레이션 과정에서 관찰되었던 특정 단일 피쳐와 타겟 간의 비현실적인 결합성($r=1.00$)이 본 데이터셋에서는 관찰되지 않았다. 이는 단일 피쳐에 대한 의존이 완화되었을 가능성을 시사한다. 다만 상관계수의 감소만으로 머신러닝 분류 모델이 특정 규칙을 암기하는 데이터 누수(Data Leakage) 현상이 해소되었다고 판단할 수는 없으며, 프로세스(proc) 및 DNS 쿼리(qry)와 타겟 간 상관계수($r=0.89$)가 여전히 높게 나타난 점도 함께 고려해야 한다. 데이터 누수 여부와 현실적인 결정 경계(Decision Boundary)의 형성 여부는 피쳐 중요도 분석이나 피쳐 제거(Ablation) 실험 등 별도의 검증이 필요하다.

둘째, 독립 변수 간의 상관성 관점에서 프로세스(proc)와 DNS 쿼리(qry) 간에 높은 상관계수($r=0.97$)가 산출되었다. 이는 이상 권한 프로세스(powershell.exe 등)의 실행과 외부 악성 C2 서버 도메인 질의가 함께 나타나는 공격 시나리오의 특성과 부합하는 관찰이다. 다만 피쳐 간 상관관계는 인과관계를 의미하지 않으며, 범주형 변수는 정수형 코드로 변환한 후 탐색적으로 피어슨 상관계수를 산출하였으며, 해당 계수는 범주 간 서열적 의미를 나타내는 지표가 아니라 변수 간 연관성에 대한 탐색적 참고값으로 해석하였다.

결론적으로, 구축된 데이터셋은 단순 무작위 난수가 아닌, 이기종 보안 로그 간의 상관 구조를 일정 수준 포함하고 있으며, 이를 학습하는 오픈프미스 sLLM(Llama-3) 기반 합성 데이터 생성 프레임워크의 유용성을 검증하기 위한 통계적 기초로 활용할 수 있다.

3.3. 하이브리드 비식별화 아키텍처(Regex + NER)

LANL의 사이버망 인증 로그 내 중요 정보 제거를 위해 두 가지 전처리를 결합하였다.

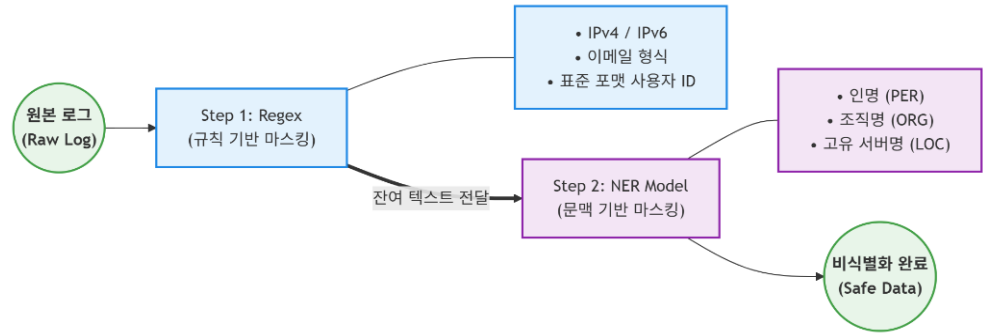


그림 4. 하이브리드 비식별화를 위한 두 전처리 단계

Step 1에서는 정규식(Regex)을 통해 IPv4/IPv6, 이메일 형식, 표준 포맷의 사용자 ID를 규칙 기반으로 1차 마스킹하였다.

Step 2에서는 HuggingFace의 BERT 기반 개체명 인식(NER) 모델을 활용하여 정규식이 놓친 변칙적인 인명(PER), 조직명(ORG), 고유 서버명 등을 문맥적으로 탐지하여 2차 마스킹하였다.

LANL Cyber1 원자료는 공개 전 비식별화가 이루어진 데이터이므로, 원자료에 자연적으로 존재하는 195건의 개인정보를 ground truth로 간주하는 방식은 적절하지 않다. 따라서 본 연구에서는 사전에 고정한 195건의 민감정보를 홀드아웃 표본에 통제된 방식으로 주입하고, Regex 단독 방식과 Regex+NER 하이브리드 방식의 탐지율을 비교하였다. 이 평가는 실제 LANL에 존재하는 민감정보의 비율을 측정한 것이 아니라, 정의된 민감정보 유형에 대한 탐지·마스킹 처리의 통제된 기술평가이다.

표 3. 통제된 평가 표본의 민감정보 유형별 비식별화 탐지율 결과

민감정보 유형	정답지(건수)	Regex 탐지(%)	Hybrid 탐지(%)	Regex Recall(%)
IPv4	30	30 (100.0%)	30 (100.0%)	100.0%
IPv6	20	20 (100.0%)	20 (100.0%)	100.0%
Email	25	25 (100.0%)	25 (100.0%)	100.0%
User ID	25	25 (100.0%)	25 (100.0%)	100.0%
PERSON	35	18 (51.4%)	35 (100.0%)	51.4%
ORG	30	17 (56.7%)	30 (100.0%)	56.7%
SERVER ID	30	15 (50.0%)	30 (100.0%)	50.0%
합계	195	150(76.9%)	195(100.0%)	76.9%

표 3은 식별정보 유형별 하이브리드 비식별화 탐지 결과를 나타낸다. 전체 평가 대상 195건 가운데 Regex 단독 방식은 150건을 탐지하여 전체 탐지율 76.9%를 나타냈다. IPv4, IPv6, Email 및 User ID와 같이 일정한 문자열 패턴을 갖는 식별정보는 모두 탐지되어 100.0%의 탐지율을 보였다. 반면 Person, Organization 및 Server ID와 같이 표현 형태가 다양하거나 문맥에 의존하는 식별정보의 Regex 탐지율은 각각 51.4%, 56.7%, 50.0%로 상대적으로 낮게 나타났다.

이후 BERT 기반 NER을 추가 적용한 결과, Regex에서 탐지하지 못한 Person 17건, Organization 13건, Server ID 15건이 추가로 탐지되어 전체 195건이 모두 탐지되었다. 이에 따라 Regex+NER 하이브리드 방식의 전체 탐지율은 100.0%로 나타났으며, 특히 정형화된 문자열 패턴뿐 아니라 문맥적 특성을 갖는 식별정보에 대해서도 탐지 범위를 확대할 수 있음을 확인하였다.

다만 위 결과는 본 연구에서 구성한 195건의 통제된 평가 표본에 한정된 결과이다. 따라서 이를 다른 유형의 민감정보 또는 실제 방산 데이터 환경에 일반화하여 완전한 비식별화 성능으로 해석해서는 안 되며, 실제 환경에서는 추가적인 도메인별 검증과 잔여 식별정보에 대한 별도의 위험평가가 필요하다.

3.4. 합성데이터 생성 모델(SDV vs Llama-3)

정형 데이터 생성에는 Synthetic Data Vault(SDV)의 CTGAN을 적용하였다. 입력 데이터의 범주형 및 수치형 변수 유형을 명시하고, 동일한 학습 표본을 대상으로 CTGAN을 학습한 뒤 합성 데이터를 생성하였다.

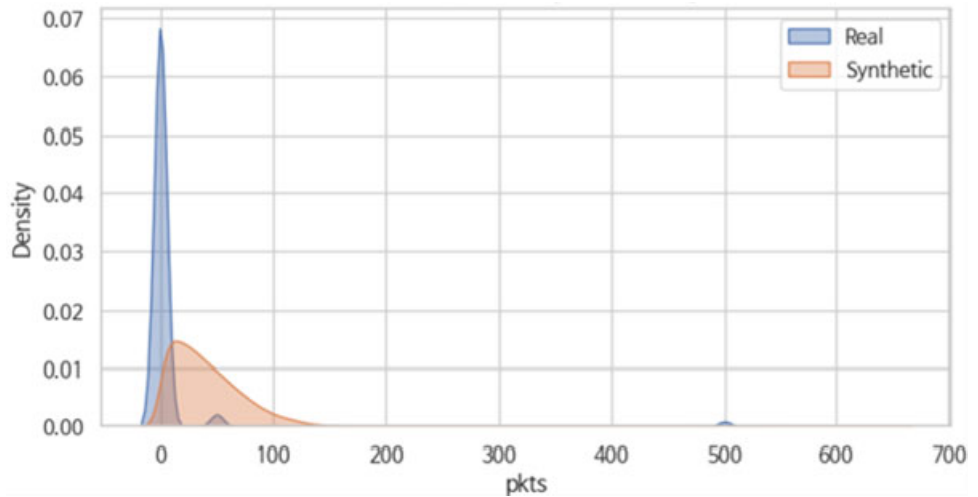


그림 5. 원본 데이터와 합성 데이터 밀도 유사도(Packet Count)

N2SF 폐쇄망 환경을 가정한 합성데이터 생성 실험에서는 로컬에 배치한 Llama-3 계열 언어모델을 사용하였다.[15] 모델은 local_files_only 방식으로 로딩하여 외부 API 호출이나 외부 모델 다운로드가 발생하지 않도록 구성하였다. 생성 결과는 사전에 정의한 JSON 스키마로 검증하고 스키마를 충족하지 않는 출력은 합성데이터로 사용하지 않았다.

4. 합성데이터 성능 평가 결과(Performance Evaluation)

4.1. 하이브리드 비식별화의 효과성 결과

하이브리드 비식별화(Regex + NER) 파이프라인의 탐지 성능을 평가하였다. 단일 정규표현식은 고정된 IP나 이메일만을 식별하여 미탐(False Negative)의 한계를 보였다. 그러나 자연어 처리 기반의 NER 모델을 후속으로 결합한 결과, 변칙적인 프로젝트명이나 서버 명칭 등을 문맥 기반으로 탐지하여 본 평가 표본의 식별 대상 195건을 100% 탐지하였다⁴⁾. 평가 표본은 LANL 사이버 보안 로그로 선정하였다. 해당 로그는 공개된 지 오래되어 현재의 방산 시스템 환경과는 차이가 있으나, 실제 방산 로그는 외부 반출이 불가한 만큼 비공개 자료의 데이터셋 구성을 모사하기 위한 대리(proxy) 데이터로 활용할 수 있다고 판단하였다. 다만 이는 실제 방산 데이터와의 동일성을 전제

4) <https://github.com/saulabe-jsoh/dis2026>

한 것이 아니며, 그 한계는 5.2절에서 논의한다. 전체 실험 데이터 100,000건은 학습 데이터와 평가 데이터로 분리하였으며, 비식별화 성능 평가는 학습에 사용되지 않은 21,000건의 홀드아웃 평가 표본을 대상으로 수행하였다. 하이브리드 비식별화 탐지 성능의 통제된 평가를 위해 사전에 정의한 7개 유형의 민감정보를 평가 표본에 주입하고, 총 195건의 ground truth를 구성하였다. 이후 동일한 평가 표본에 대해 Regex 단독 방식과 Regex+NER 하이브리드 방식을 각각 적용하여 식별정보 탐지 여부를 비교하였다. 이러한 방식은 실제 LANL 원자료에 존재하는 개인정보의 규모를 측정하기 위한 것이 아니라, 사전에 정의된 식별정보 유형에 대한 비식별화 탐지·마스킹 성능을 통제된 환경에서 평가하기 위한 것이다.

4.2. 데이터 유용성(Utility) 및 통계적 유사성 검증(SDV)

합성데이터가 연구 목적으로 얼마나 유효한지를 확인하기 위해 SDV의 평가 함수(sdv.evaluation)를 활용하여 종합 품질 점수(Quality Score)를 도출하였다. 분포 유사성(Column Shapes) 평가지표를 통해 개별 변수의 통계적 분포 유사도를 확인하였고, 상관관계 유지율(Column Pair Trends)을 통해 변수 간의 트렌드 유지율을 평가하였다.

4.3. TSTR(Train Synthetic, Test Real) 기반 레드팀 탐지 성능 평가

합성데이터가 연구·교육 목적의 사이버 위협 탐지 데이터로 활용될 수 있는지를 평가하기 위해 XGBoost 분류기를 이용한 TSTR(Train Synthetic, Test Real) 검증을 수행하였다.[18] TSTR은 합성데이터를 학습 데이터로 사용하고, 학습에 사용되지 않은 실제 평가 데이터를 테스트 데이터로 사용하는 방법이며, 비교 기준으로는 실제 학습 데이터와 실제 평가 데이터를 이용한 TRTR(Train Real, Test Real)을 설정하였다.

평가 지표는 클래스 불균형 환경에서 소수 클래스의 탐지 성능을 종합적으로 확인할 수 있도록 F1-Score, Precision, Recall, FPR(False Positive Rate), FNR(False Negative Rate) 및 PR-AUC(Precision-Recall Area Under the Curve)를 사용하였다. F1-Score는 Precision과 Recall의 조화평균으로 다음과 같이 산출하였다.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

성능 보존율은 TRTR의 F1-Score를 기준으로 TSTR 모델의 상대적 성능을 다음과 같이 계산하였다.

$$\text{Performance Retention}(\%) = \frac{F1_{TSTR}}{F1_{TRTR}} \times 100$$

표 4. 모델별 TSTR (Train Synthetic, Test Real) 검증 성능 결과

학습 모델	F1-Score	Precision	Recall	FPR	FNR	PR-AUC
TRTR(Real)	0.9848	0.9948	0.9750	0.03%	2.5%	0.9980
TSTR(SDV)	0.9390	0.9937	0.8900	0.03%	11.0%	0.9620
TSTR(sLLM)	0.9367	1.0000	0.8810	0.00%	11.9%	0.9550

실험 결과, TRTR 모델의 F1-Score는 0.9848로 나타났으며, SDV-CTGAN 기반 TSTR 모델은 0.9390, Llama-3 기반 sLLM TSTR 모델은 0.9367로 나타났다. TRTR을 기준으로 한 성능

보존율은 각각 95.3%와 95.1%로 산출되었다. 이에 따라 합성데이터를 이용해 학습한 두 모델 모두 실제 데이터 학습 모델 대비 높은 수준의 분류 성능을 유지하고 있음을 확인하였다.

특히 TSTR-sLLM 모델은 Precision 1.000, Recall 0.8810, F1-Score 0.9367을 나타냈으며, FPR은 0.00%, FNR은 11.9%로 나타났다. 이는 본 평가 표본에서 위협으로 잘못 분류된 정상 데이터가 관찰되지 않았음을 의미하는 동시에, 실제 위협 데이터 중 일부가 탐지되지 않았음을 의미한다. 따라서 위협 탐지 성능을 해석할 때에는 높은 Precision뿐 아니라 Recall 및 FNR을 함께 고려할 필요가 있다.

PR-AUC는 클래스 불균형 환경에서 양성 클래스인 위협 데이터에 대한 분류 성능을 추가적으로 확인하기 위한 지표로 활용하였다. 본 실험에서는 TRTR 0.9980, TSTR-CTGAN 0.9620, TSTR-sLLM 0.9550으로 나타났다. 이는 합성데이터를 이용한 학습에서도 위협 클래스의 분류 가능성이 일정 수준 유지되었음을 보여주지만, TRTR에 비해 TSTR 모델의 Precision-Recall 관계가 일부 저하되었음을 나타낸다.

표 5. 모델별 TSTR 검증 성능 보존율 결과

학습 모델	F1-Score	성능 보존율	비고
TRTR(Real)	0.9848	기준점(100%)	실제 데이터 기반 기준 모델
TSTR(SDV)	0.9390	95.3%	위협 클래스 Recall 0.8900
TSTR(sLLM)	0.9367	95.1%	위협 클래스 Recall 0.8810

그림 6은 세 가지 학습 방식의 F1-Score와 성능 보존율을 비교한 결과이다. TRTR의 F1-Score가 0.9848로 가장 높게 나타났으며, TSTR-CTGAN과 TSTR-sLLM은 각각 0.9390과 0.9367로 나타났다. 두 TSTR 모델의 F1-Score 차이는 0.0023으로 크지 않았으며, 단일 데이터 셋과 단일 학습·평가 분할을 이용한 본 실험만으로 두 생성방식의 일반적 성능 차이를 판단하기에는 한계가 있다.

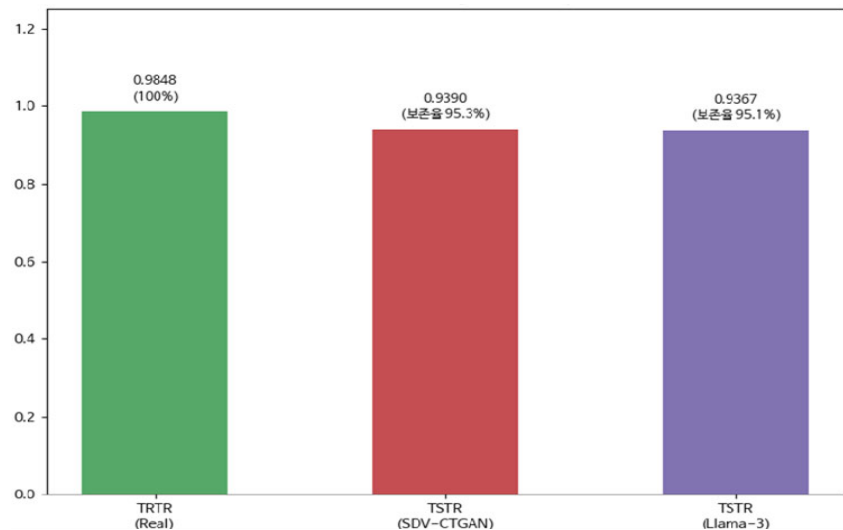


그림 6. 모델별 TSTR(Train Synthetic, Test Real) 성능 비교

또한 그림 7은 TSTR-sLLM 모델의 혼동행렬을 나타낸다. 평가 결과 정상 데이터에 대한 TN은

20,000건, FP는 0건이었으며, 위협 데이터에서는 FN 119건, TP 881건으로 나타났다. 이에 따라 위협 클래스의 Precision은 1.000, Recall은 0.8810, F1-Score는 0.9367로 계산되며, FPR은 0.0%(0/20,000), FNR은 11.9%(119/1,000)로 나타났다. 이 결과는 본 프록시 평가 데이터에서 정상 데이터를 위협으로 오탐하는 사례가 관찰되지 않았음을 보여주지만, 실제 방산 사이버 관제 환경에서 오탐 피로도(Alert Fatigue)나 관제 비용이 증가하지 않는다는 것을 의미하는 것은 아니다.

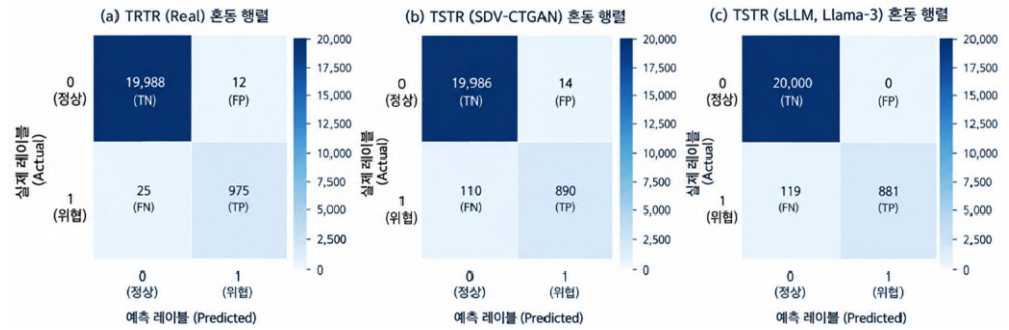


그림 7. TSTR 모델의 오탐/미탐 혼동 행렬 분포

특히 위협 클래스 1,000건 중 119건이 미탐으로 나타났다는 점에서 Recall은 본 연구의 위협 탐지 성능을 해석하는 데 중요한 지표이다. 그러나 이들 미탐 사례가 특정 공격 유형이나 은닉형 공격에 집중되어 있다는 직접적인 근거는 본 연구에서 제시하지 않았다. 따라서 미탐 원인을 저속 은닉형 APT 공격 등 특정 공격 유형과 직접 연결하여 해석하기보다는, 합성데이터 기반 학습 모델에서 일부 위협 패턴의 표현 또는 학습이 충분하지 않았을 가능성이 있다는 수준으로 해석하는 것이 타당하다.

한편, 본 실험에서 관찰된 TSTR 성능은 합성데이터가 실제 사이버 위협 탐지 연구에 활용될 수 있는 가능성을 보여주는 실험적 근거로 볼 수 있다. 특히 sLLM 기반 합성데이터 학습 모델이 TRTR 대비 95.1%의 F1-Score를 유지한 것은 실제 원본 데이터에 접근하기 어려운 환경에서 연구·교육용 데이터셋으로 활용할 수 있는 가능성을 시사한다. 다만 이러한 성능 보존율만으로 합성데이터의 정보보호 수준이나 법적 안전성이 확보되었다고 판단할 수는 없으며, 재식별 가능성, 원본 데이터 암기 여부 및 멤버십 추론 등에 대한 별도의 보안성 평가가 필요하다.

또한 TSTR-sLLM 결과가 프로세스 실행, DNS 질의, 네트워크 전송 등 사이버 위협과 관련된 다차원적 패턴을 일정 수준 반영했을 가능성을 시사하지만, 이는 분류 성능을 통해 간접적으로 관찰된 결과이다. 본 연구에서는 시계열 순서나 공격 단계 간의 인과관계를 직접 검증하지 않았으므로, 해당 결과를 공격 인과관계가 보존되었다는 근거로 확대하여 해석하지 않았다.

SDV-CTGAN과 Llama-3 기반 sLLM의 F1-Score는 각각 0.9390과 0.9367로 나타났으며, 두 모델의 차이는 0.0023이었다. 따라서 본 연구의 단일 데이터셋·단일 분할 결과만으로 어느 한 생성 방식이 일반적으로 우수하다고 판단하기는 어렵다. 본 연구에서 관찰된 결과를 기준으로 볼 때, SDV-CTGAN은 정형화된 범주형·수치형 데이터의 합성에 적용하기 용이한 반면, sLLM은 적절한 프롬프트와 출력 구조를 적용할 경우 로그의 문맥적 표현을 활용할 수 있다는 차이가 있다. 다만 본 연구에서는 두 생성 방식의 계산비용, 생성시간 및 반복실험에 따른 통계적 변동성을 정량적으로 비교하지 않았으므로, 이러한 차이는 기술적 특성에 대한 탐색적 해석으로 한정한다.

결과적으로 본 연구는 N2SF 환경에서 직접적인 원본 데이터 공유가 제한되는 경우, 하이브리드

비식별화와 합성데이터를 결합하여 연구·교육 목적의 데이터 활용을 지원할 수 있는 기술적 가능성을 프록시 데이터 기반 실험을 통해 검토하였다. 그러나 실제 방산 데이터의 반출 또는 보안등급 하향 여부는 본 실험의 TSTR 성능만으로 결정할 수 없으며, 해당 기관의 보안정책, 반출 승인 절차 및 관련 법·제도에 따른 별도의 판단이 필요하다.

4.4. 보안성 검증 고찰

합성데이터는 실제 관측값이 아닌 모델이 생성한 가상의 데이터이므로 특정 개체와 직접 대응될 가능성이 낮아지나, 모델이 원본을 암기(Overfitting)한 경우에는 원본 정보가 유출될 수 있다. 이러한 위험을 완화하기 위해 훈련 과정에서 향후에는 차분 프라이버시(Differential Privacy) 기법의 구배 클리핑(Gradient Clipping) 및 노이즈 주입을 추가 적용할 수 있다. 다만 본 논문에서는 재식별·멤버십 추론 공격에 대한 정량적 평가 결과를 제시하지 않았으므로, 하이브리드 비식별화와 생성 모델의 결합이 전통적 마스킹 대비 원본 추론 위험을 낮추었는지는 별도의 공격 시나리오 기반 평가로 검증되어야 한다. 본 연구가 확인한 것은 식별 대상 표본에 대한 탐지 성능과 TSTR 유용성 측면에서의 기술적 적용 가능성이다.

아울러 합성데이터의 유용성과 안전성은 함께 평가되어야 하므로[9], 본 연구의 TSTR 결과를 곧바로 안전성이 검증되었다는 의미로 해석해서는 안 된다. 또한 N2SF 체계에서의 등급 하향 및 반출 가능 여부는 이러한 기술적 검증 결과만으로 결정되는 것이 아니라, 5.2절에서 제안하는 「방위산업보안업무훈령」 개정 및 반출 절차 신설 등 별도의 제도적 근거에 따라 판단되어야 한다.

5. 결론 및 향후 연구 과제

5.1. 연구 요약 및 의의

본 연구는 N2SF 전환이라는 국방 보안 환경의 변화 속에서 데이터 보호와 연구·교육 목적의 활용 간의 균형을 지원하기 위한 기술적 대안으로 하이브리드 비식별화와 합성데이터 활용 가능성을 검토하였다.

미국 로스앨러모스 국립연구소(LANL)의 사이버 위협 로그를 프록시 데이터로 활용하여 실험한 결과, 하이브리드(Regex + NER) 비식별화 전처리를 적용하였을 때 통제된 평가 표본에 주입한 195건의 식별 대상이 모두 탐지·마스킹되었다. 또한 비식별화된 데이터에 통계 기반 모델(SDV)과 생성형 언어모델(sLLM, Llama-3)을 적용하여 합성데이터를 생성하고 유용성을 비교한 결과, sLLM 기반 합성데이터로 학습한 모델은 TSTR F1-Score 기준 실제 데이터 학습 모델 대비 95.1%의 성능을 유지하였다.

이러한 결과는 실제 방산 데이터의 반출이 제한된 환경에서 프록시 데이터 기반으로 하이브리드 비식별화와 합성데이터를 결합한 데이터 활용 방안을 검토할 수 있음을 보여준다. 다만 본 실험은 실제 방산 데이터가 아닌 프록시 데이터셋을 대상으로 수행되었으며, 재식별 및 멤버십 추론 공격 등에 대한 정량적 보안성 평가를 수행하지 않았으므로 실제 방산 데이터의 보안등급 하향 또는 외부 반출 가능성을 직접 입증하는 결과로 해석해서는 안 된다. 실제 등급 조정 및 반출 여부는 해당 기관의 보안정책과 승인 절차 및 관련 제도적 근거에 따라 별도로 판단되어야 한다.

5.2. 연구의 한계점 및 향후 연구 과제

본 연구의 한계는 다음과 같다. 첫째, 실제 방산 데이터의 외부 반출이 불가하여 LANL 사이버

보안 로그를 프록시 데이터로 활용하였으므로, 실제 방산 환경의 로그와는 데이터 분포, 필드 구성, 위협 유형에서 차이가 있을 수 있으며 이에 따라 본 연구 결과의 일반화(외적 타당성)에는 제약이 있다. 둘째, 성능 평가는 단일 데이터셋과 단일 학습·평가 분할에 기반하며 반복 실험에 따른 변동성은 제시하지 못하였다. 셋째, 보안성은 195건의 식별 대상 표본에 대한 탐지 성능의 범위에서 평가하였으며, 재식별·추론 공격에 대한 정량적 평가는 포함하지 못하였다.

이러한 한계를 보완하기 위해 향후 다음과 같은 후속 연구를 추진할 계획이다. 먼저 보안 승인된 폐쇄망 환경에서 실제 방산 로그 데이터를 활용해 본 연구의 파이프라인을 재검증하고, 프록시 데이터와의 차이가 결과에 미치는 영향을 분석한다. 이와 함께 네트워크 트래픽, 시스템 로그 등 다양한 유형의 보안 데이터로 검증 범위를 확장하며, 반복 실험과 PR-AUC를 포함한 클래스별 지표, 재식별·멤버십 추론 공격 기반의 프라이버시 평가를 함께 수행하여 외적 타당성과 결과의 신뢰성을 높이고자 한다.

이러한 한계를 감안하되 기술적 검증 결과를 바탕으로, 향후 N2SF 환경 내 합성데이터 활용의 실효성을 제고하기 위해 다음과 같은 정책적·제도적 개선 방향을 제안한다.

첫째, 「방위산업보안업무훈령」 개정 검토이다. N2SF 체계 내에 가명정보 및 ‘합성데이터’에 대한 명확한 법적·제도적 지위와 반출 절차의 마련을 검토할 필요가 있다. 특히, 일정한 통계적 유용성 검증과 재식별 위험평가 등을 통과한 데이터에 대해서는 망간 전송 및 민간 연구기관과의 제한적 공유를 허용할 수 있는 절차와 기준을 마련하는 방안을 검토할 필요가 있다.

둘째, 안전한 데이터 클린룸(Data Clean Room) 운영이다. 국방부 및 방위사업청 주도로 N2SF 내부망에 합성데이터 생성 및 검증을 위한 독립된 전담 구역을 조성해야 한다. 이를 통해 외부 연구기관이 필요한 데이터셋을 안전한 보안 절차에 따라 요청하고 반출받을 수 있는 표준화된 프로세스를 제도화해야 할 것이다.

끝으로, 본 연구는 정형화된 사이버 보안 로그 데이터를 중심으로 합성데이터의 유효성을 검증하였다는 한계를 지닌다. 향후 연구에서는 본 연구의 방법론을 확장하여, 복잡한 시계열(Time-series) 특성을 내포한 실전 무기체계 운용 데이터나 비정형 영상 데이터(SAR, EO/IR 등)에 적용 가능한 다차원적 복합 합성 모델을 설계하고 그 유용성을 검증하는 후속 연구가 수반되어야 할 것이다.

참고문헌

- [1] 류연승 등. 산업보안학 제2판. 박영사. 2022.
- [2] 류연승. 방산보안 2.0. 정보보호학회지, 제28권 제6호, pp. 6-12. 2018.
- [3] 류연승 등. 방위산업 안보환경 변화에 따른 방산안보정책 검토. 한국과 세계, 제4권 제2호, pp. 207-232. 2022.
- [4] 백운일 등. 미 CMMC 동향에 대응한 정책방안 연구. 방산안보연구, 제1권 제2호, pp. 21-39. 2025.
- [5] 소프트웨어정책연구소. 데이터 부족 현상 해결을 위한 합성데이터(Synthetic Data) 동향과 시사점. 2024.
- [6] 국가정보원, 국가보안기술연구소. 국가망 보안체계(N2SF) 보안 가이드라인. 2025.
- [7] 개인정보보호위원회. 합성데이터 생성 및 활용 안내서. 2023.
- [8] 정보통신기획평가원. AI 기반 프라이버시 보존형 합성데이터 생성 기술. 주간기술동향, 2120호. 2023.
- [9] 조아라, 이기혁. 프라이버시 보존형 합성데이터 생성 기술의 유용성 및 안전성 평가 연구: 정형 데이터를 중심으로. 한국정보보호학회논문지, 제32권 제4호, pp. 589-601. 2022.
- [10] 성원영 등. KISA, 국가망보안체계(N2SF) 시범사업 3건 동시 추진...새로운 보안 패러다임 실증. 컴퓨터월드, 3월호. 2026.
- [11] 한국인터넷진흥원. N2SF 다층보안체계 실증 가이드 및 연구보고서. 2026.

- [12] 국방부. 방위산업보안업무훈령(국방부훈령 제○○○○호). 20XX.
- [13] Patki N et al. The Synthetic Data Vault. Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399-410. 2016. <https://doi.org/10.1109/DSAA.2016.49>
- [14] Xu L et al. Modeling Tabular Data Using Conditional GAN. Advances in Neural Information Processing Systems 32 (NeurIPS 2019), pp. 7335-7345. 2019.
- [15] Grattafiori A et al. The Llama 3 Herd of Models. arXiv:2407.21783. 2024. <https://doi.org/10.48550/arXiv.2407.21783>
- [16] Kent AD. Cyber-Security Data Sources for Dynamic Network Research. In Dynamic Networks and Cyber-Security, World Scientific, pp. 37-65. 2016. https://doi.org/10.1142/9781786340757_0002
- [17] Kent AD. Comprehensive, Multi-Source Cyber-Security Events. Los Alamos National Laboratory. 2015. <https://doi.org/10.17021/1179829>
- [18] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794. 2016. <https://doi.org/10.1145/2939672.2939785>