인공지능 기반 네트워크 침입 탐지 및 사이버보안 기술 동향

AI-based Intrusion Detection System and Cybersecurity Technology Trends

박노삼 (N.S. Park, siru23@etri.re.kr) 지능형네트워크보안연구실 책임연구원
이종훈 (J.H. Lee, mine@etri.re.kr) 지능형네트워크보안연구실 책임연구원
한미경 (M.K. Han, mkhan@etri.re.kr) 지능형네트워크보안연구실 책임연구원
이훈기 (H.K. Lee, lhk@etri.re.kr) 지능형네트워크보안연구실 책임연구원
신용윤 (Y.Y. Shin, uni2u@etri.re.kr) 지능형네트워크보안연구실 책임연구원

ABSTRACT

This paper explores the evolving trends of AI in network intrusion detection system and cybersecurity. It analyzes various ML paradigms, including supervised, unsupervised, and semi-supervised learning, outlining their respective advantages and limitations in cybersecurity applications. Furthermore, the analysis extends to the latest technological trends, including self-supervised learning, which addresses the challenge of data labeling, and the integration of advanced techniques like federated learning and generative AI. These emerging approaches aim to improve model efficiency in hybrid and multi-cloud environments, enhance the detection of unknown attacks, and bolster the overall resilience of cybersecurity frameworks. The findings underscore the critical role of AI in building adaptive and intelligent security systems capable of defending against dynamic and complex cyber threats.

KEYWORDS AI, Cybersecurity, Network Threats, Self-Supervised Learning

I. 서론

디지털 기술의 급속한 발전은 전 세계 산업계 전반에 걸쳐 큰 변화를 불러왔다. 인공지능(AI), 사물인터넷(IoT), 클라우드 컴퓨팅, 5G 네트워크 등의 기술은 기업 운영, 정부 행정, 의료, 교육 등 모든 영역에서 데이터의 연결성과 활용성을 극대화하고 있다. 그러나 이러한 기술 발전은 동시에 다양한 형태의 사이버 위협을 수반하게 되었으며, 사이버보안은 디지털 사회의 핵심 인프라로 자리 잡고 있다.

* DOI: https://doi.org/10.22648/ETRI.2025.J.400505
* 이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구 결과임[No. RS-2024-00397469, 특화망·기업망 통합보안을 위한 5G 특화망 보안 기술개발].

최근 사이버 공격은 특정 개인이나 기업을 넘어서 국가 기반 시설이나 글로벌 공급망까지 영향을 미치는 수준으로 발전하고 있으며, 그 공격 방식도 정교하고 다양화되고 있다. 전통적인 시그니처 기반 보안 솔루션으로는 최신 위협에 효과적으로 대응하기 어려워졌으며, 특히 암호화된 네트워크 트래픽을 포함한 복잡한 데이터 환경에서는 탐지의 한계가 더욱 두드러진다.

생성형 AI 등 기술의 발전과 클라우드 환경 및 제로트러스트 보안 모델의 확산, AI 기반 신규 위협 등장 등은 사이버보안 지형의 변화를 일으키고 있다. 사이버보안 기술은 단순한 방어 중심에서 벗어나 자율성과 선제 대응 능력을 갖춘 방향으로 진화 중이다.

1. 사전방어형 사이버보안

글로벌 사이버보안 환경은 기존의 탐지 후 대응 중심의 사이버보안 체계를 넘어 위협 발생 전에 공격을 능동적으로 방지하는 사전방어형(Preemptive) 사이버보안으로 빠르게 전환하고 있다[1].

새롭게 부상하는 생성형 AI 기반 위협은 기존의 탐지 및 대응 전략에 도전장을 내밀고 있다. 선제적 사이버보안 기술은 기존 보안 제어를 크게 강화하고 사이버 방어 역량을 향상할 수 있다는 시사점을 제시한다.

사이버보안 기술은 단독 제품 위주에서 통합된 플랫폼으로 초점이 이동하고 있다. 제로데이, 공급망 공격 등에 대한 선제적 대응 없이는 탐지 불가능한 위협이 증가하고 있다. 혁신 선도 기업들은 여러 선제적 사이버보안 기술을 결합하여 더욱 강력한 심층 방어 솔루션을 개발하고 있다.

2. 에이전트 AI

사이버보안 분야에서 주목받고 있는 동향 중 하나는 에이전트 AI(Agentic AI)이다[2]. 에이전트 AI는 인간의 개입 없이 스스로 판단하고 행동할 수 있는 자율형 AI 시스템을 의미한다. 단순히 프롬프트에 응답하는 수준을 넘어, 환경 분석 → 의사결정 → 실행까지 독립적으로 수행할 수 있는 에이전트 기반 구조이다. 이러한 자율성은 복잡한 업무를 자동 처리하며 사람의 감독 부담을 줄여준다. 마이크로소프트의 Security Copilot Agent와 같은 도입은 단일 AI 어시스턴트에서 벗어나 에이전트 팀이 서로 협력하여 완전 자율적으로 수행하는 패러다임 전환을 보여준다.

3. AI 기반 자동화 탐지 및 대응

AI는 보안 분야에서 탐지 정확도 향상과 대응 효율성을 획기적으로 강화하고 있다. 특히 네트워크 탐지 대응(NDR: Network Detection and Response)과 관리형 탐지 대응(MDR: Managed Detection and Response) 서비스는 지속적인 위협 모니터링, 탐지, 분석 및 대응의 자동화를 가능하게 한다.

가트너의 연구에 의하면, 조직의 30%만이 MDR 공급 업체를 활용하고 있는 것으로 나타났으나 2025년까지 50% 기업이 MDR 서비스를 도입할 것으로 예측하며, 이는 조직의 보안 인력 부족 문제해결에 효과적이라고 분석한다[3].

다크트레이스(Darktrace), 벡트라(Vectra) AI, 시스코(Cisco) Secure IDS 등은 AI 기반 자동 탐지 및 자동화 대응 기능을 제공하며, 탐지 속도와 정확성을 크게 향상시키고 있다[4]. NDR은 비정상 행동 패턴을 분석해 내부 및 외부 침입 행위를 탐지하며, AI 기반 정책으로 자동 분류와 대응까지 가능하게 한다.

4. 자가학습 기반 인텔리전스

AI 기반 사이버보안 기술에서 주목받는 분야 중 하나는 자가학습(Self-Learning)을 통한 자율적 보안 시스템이다. 사이버보안 분야의 예측 위협 인텔리전스는 자가학습을 활용하여 잠재적 위협이 현실화하기 전에 사전에 식별한다. 원시 데이터를 실행 가능한 인사이트로 변환함으로써 정보에 기반한 의사결정을 내리고 사후 대응적인 보안 태세에서 사전 예방적인 보안 태세로 전환할 수 있도록 지원한다.

자가학습은 준지도학습(Semi-Supervised)의 형태로 동작하며 일부 라벨 데이터로 학습한 뒤 비라벨 데이터에 대해서도 반복적으로 학습하여 모델을 개선한다. 준지도학습의 다른 형태인 자기지도학습(Self-Supervised Learning)은 대규모의 비라벨 데이터를 활용하여 특징 표현을 학습할 수 있어, 라벨링 비용이 많이 드는 보안 데이터 환경에서 매우 효율적인 대안으로 떠오르고 있다.

사이버보안 분야의 이러한 기술 변화에 대응하기 위해 AI 기반의 네트워크 침입 탐지 시스템(IDS: Intrusion Detection System)이 대안으로 부상하고 있다. AI는 대량의 네트워크 트래픽 데이터를 분석하고, 이상 패턴을 스스로 학습하며, 새로운 유형의 공격에도 실시간으로 대응할 수 있는 능력을 갖추고 있다. 특히 지도학습, 비지도학습 기반의 머신러닝과 딥러닝 알고리즘을 이용한 자동화된 이상 탐지 기술은 보안 담당자의 부담을 줄이고 대응 효율을 높이는 데 큰 역할을 하고 있다. 또한, 네트워크 환경에서도 BYOL[5], SimCLR[6], Deep InfoMax[7] 등의 자기지도학습 프레임워크를 통해 비정형 네트워크 트래픽, 암호화된 트래픽, IoT 데이터 등의 복잡한 환경에서도 높은 탐지 성능을 보여주고 있다.

특히 네트워크 변화에 실시간으로 적응할 수 있는 온라인 학습, 도메인 간 탐지 성능 유지를 위한 도메인 적응 등의 도메인 전이 기술 등과 결합하여 실무 적용 가능성이 높아지고 있으며, 향후 사이버보안 체계 전반에서 핵심 요소로 자리 잡을 것으로 전망된다.

최근 등장한 생성형 AI는 보안 분야에도 큰 영향을 미치고 있다. 생성형 AI는 기존 데이터 기반 학습을 넘어, 새로운 데이터를 생성하거나 시뮬레이션 환경을 구축하여 보안 탐지 모델의 학습 효과를 극대화할 수 있다.

본고에서는 글로벌 사이버보안 기술의 변화에 따른 AI 기반 네트워크 침입 탐지 분야의 기술 및 연구 동향을 소개한다. 또한, 생성형 AI와 자기지도학습 기술을 활용한 차세대 침입 탐지 및 대응 전략을 제안함으로써 미래 사이버보안 체계의 방향성을 제시하고자 한다.

II. AI 기반 네트워크 침입 탐지

사이버 위협은 지능화, 정교화되고 있으며, 전통적인 시그니처 기반의 침입 탐지 시스템만으로는 고도화된 공격에 대응하기 어려워졌다. 이러한 한계를 극복하기 위해 AI 기반 네트워크 침입 탐지 기술이 필수적인 요소로 자리 잡고 있다. AI는 방대한 네트워크 트래픽 데이터 속에서 복잡한 패턴과 이상 징후를 식별함으로써 보다 능동적이고 지능적인 방어 체계를 구축하는 데 이바지한다. 본 장에서는 AI 기반 침입 탐지 기술의 핵심 방법을 지도학습, 비지도학습, 준지도학습으로 분류하고, 기술별 특성을 비교하고 적용 가능성과 한계를 논한다.

1. 지도학습 기반 침입 탐지 기술

1.1 개요

지도학습은 라벨링된 데이터를 통해 정상적인 네

트워크 행위와 악성적인 행위 간의 상관관계를 학습하고, 이를 바탕으로 트래픽의 유형을 분류한다. 이러한 접근 방식은 명확한 분류 기준을 제공하며 높은 탐지 정확도를 달성할 수 있게 한다.

장점

- **높은 정확도:** 잘 정제된 라벨링 데이터가 확보될 경우, 특정 유형의 공격(예: DDoS 공격, 특정 맬웨어)을 높은 정확도로 식별할 수 있다.
- **낮은 오탐율:** 학습된 데이터셋을 기반으로 정상 행위와 비정상 행위를 명확하게 구분하여 오탐(False Positive)을 줄이는 데 효과적이다.

한계

- **데이터 의존성:** 모델의 성능이 학습 데이터의 품질과 양에 크게 의존하며, 학습 데이터가 충분하지 않거나 편향된 경우 모델의 성능이 저하될 수 있다.
- **데이터 라벨링의 복잡성:** 대규모 네트워크 데이터에 정확한 라벨링을 수행하는 작업은 많은 시간과 리소스를 필요로 한다.
- **알려지지 않은 공격 대응 한계:** 학습 데이터에 포함되지 않은 새로운 유형의 공격에 대해서는 탐지 성능이 현저히 떨어진다.
- **성능 오버헤드:** 딥러닝과 같은 복잡한 지도학습 모델은 학습 및 예측 과정에서 높은 연산 자원을 요구할 수 있어, 실시간 네트워크 환경 적용 시 성능 이슈가 발생할 수 있다.

1.2 대표 기술 및 적용 사례

지도학습 기반의 침입 탐지 기술은 서포트 벡터 머신(SVM) 등과 같은 전통적인 머신러닝 알고리즘은 물론, CNN, RNN 등과 같은 딥러닝 모델을 널리 활용하고 있다.

많은 상용 침입 탐지 시스템은 시그니처 기반 방식과 결합하여 지도학습 모델을 활용한다. 다만, 이상 징후 및 알려지지 않은 공격 대응 한계로 인해 비지도학습과 같이 많이 사용된다.

- **NDR:** Clumit security, 벡트라 AI, 다크트레이스와 같은 NDR 솔루션은 네트워크 트래픽을 실시간으로 분석하여 정상 행위 모델을 구축하고, 이 모델과 다른 비정상 행위를 탐지한다. 이 과정에서 지도학습, 비지도학습, 준지도학습 등을 활용하여 기존 시그니처로는 탐지할 수 없었던 위협을 검출한다.
- **SIEM(Security Information and Event Management) 솔루션:** IBM QRadar, 스플렁크(Splunk)와 같은 대규모 SIEM 솔루션은 수집된 로그 및 이벤트 데이터에 머신러닝 분석을 적용하여 이상 행위를 탐지하고 위협을 예측한다. 지도학습 기반 분류 모델을 사용하여 알려진 공격 패턴을 식별한다.
- **엔드포인트 탐지 및 대응(EDR: Endpoint Detection and Response):** EDR 솔루션들도 지도학습 모델을 활용하여 알려진 악성코드를 탐지하고, 파일 및 프로세스 실행 행위에 대한 비정상적인 활동을 분석한다.

2. 비지도학습 기반 침입 탐지 기술

2.1 개요

비지도학습은 라벨링되지 않은 데이터를 분석하여 데이터 내의 숨겨진 구조, 패턴, 또는 이상치를 스스로 발견하는 방식이다. 네트워크 침입 탐지 분야에서는 정상 트래픽의 기준선(Baseline)을 학습하고, 이 기준선에서 크게 벗어나는 비정상적인 트래픽을 이상 행위로 분류하여 공격을 탐지한다.

장점

- **제로데이 공격 탐지 능력:** 비지도학습은 사전에 정의된 라벨이나 공격 시그니처 없이 데이터의 정상적인 패턴을 학습한다. 따라서 알려지지 않은 새로운 형태의 공격이나 이상 행위 탐지율을 향상할 수 있다.
- **라벨링 작업 불필요:** 지도학습과 달리, 비지도학습은 데이터에 라벨을 부여하는 과정이 필요 없으므로, 대규모의 비정형 데이터 분석에 유리하다.
- **다양한 환경에 적용 가능:** 특정 공격 패턴에 국한되지 않고 데이터 자체의 특이성을 파악하므로, 다양한 네트워크 환경이나 시스템에 유연하게 적용될 수 있다.

한계

- **오탐:** 정상적인 네트워크 환경의 변화(예: 새로운 서비스 도입, 대규모 업데이트)를 이상치로 오인하여 잘못된 경보를 발생시킬 가능성이 높다.
- **결과 해석의 어려움:** 왜 특정 행위가 이상치로 분류되었는지에 대한 명확한 근거를 제공하기 어려울 수 있다.
- **정상 데이터의 품질 의존성:** 비지도학습 모델의 성능은 정상 데이터를 얼마나 정확하게 정의하고 학습하는지에 따라 크게 좌우된다.

2.2 대표 기술 및 적용 사례

비지도학습 기반의 침입 탐지 기술에는 전통적인 클러스터링, 오토인코더와 같은 이상치 탐지 등이 포함된다.

지도학습의 한계를 극복하기 위해 많은 상용 솔루션에서는 비지도학습 기술도 같이 적용하여 새로운 위협 탐지에 활용하고 있다.

- **내부자 위협 탐지:** 비지도학습은 기업 내부 직원의 비정상적인 행동을 탐지하는 데 활용된다. 예를 들어, 평소와 다른 시간에 민감한 데이터에 접근하거나, 평소 사용하지 않던 시스템을 이용하는 행위 등을 이상치로 분류하여 내부자 위협을 조기에 감지한다.
- **클라우드 환경 보안:** 클라우드 서비스 내의 사용자 및 시스템 활동을 모니터링하여, 계정 탈취나 권한 상승과 같은 이상 행위를 탐지하는 데 비지도학습 기술이 적용된다.

3. 준지도학습 기반 침입 탐지 기술

3.1 개요

완전히 라벨 없이 구조를 학습하는 비지도학습과 달리, 준지도학습은 지도학습과 비지도학습의 특징을 혼합한 방법으로써 일부 라벨 데이터를 기반으로 학습을 수행한다.

자기지도학습은 비라벨 데이터를 가공하여 데이터 패턴을 학습한 후, 일부 라벨 데이터로 지도학습처럼 예측한다. 이러한 모델은 비라벨 데이터를 활용하여 가치 있는 표현을 학습하고 이상 징후를 탐지함으로써 더욱 강력하고 적응력 있는 보안 시스템을 개발할 수 있도록 한다.

장점

- **효율성:** 라벨링에 드는 시간과 비용을 크게 절감하면서도 비라벨링 데이터의 이점을 활용하여 성능을 높일 수 있다.
- **높은 정확도:** 소량의 라벨 데이터를 통해 초기 모델의 정확도를 확보하고, 비라벨 데이터를 통해 모델의 일반화 능력을 향상시켜 오탐 및 미탐률을 낮출 수 있다.
- **새로운 위협 탐지 능력:** 비라벨 데이터를 활용하여 기존에 학습하지 않은 새로운 유형의 공

격 패턴을 식별할 가능성이 높아진다.

한계

- **오류 전파:** 초기 모델의 예측 오류가 가상 라벨로 이어지고, 이것이 다시 모델 훈련에 사용되면서 모델 성능 저하의 원인이 될 수 있다.
- **하이퍼파라미터 튜닝의 복잡성:** 가상 라벨 부여 기준을 설정하는 것이 중요하며, 최적의 파라미터를 찾는 것이 복잡하다.

3.2 대표 기술 및 적용 사례

준지도학습은 사이버보안의 핵심 과제인 데이터 라벨링 문제와 제로데이 위협 대응의 한계 해결을 위한 대안을 제시한다. 이는 지도학습의 정확성과 비지도학습의 유연성을 결합함으로써 적은 자원으로도 강력한 탐지 시스템 구축을 용이하게 한다.

- **적응 학습 기반 위협 탐지:** 자기 학습 기반으로 네트워크 확장 및 변경 사항을 스스로 조정하며 학습하여 네트워크 위협을 탐지한다.
- **사고 대응 자동화:** 비정상 행동의 징후를 미리 파악하고 신속한 대응 접근 방식을 제공한다.

향후 사이버보안 연구는 지도학습, 비지도학습뿐만 아니라 준지도학습을 적용한 하이브리드 모델 개발에 더 집중될 것으로 전망된다. 이러한 연구는 미래의 사이버 위협에 대한 방어 체계를 더욱더 지능적이고 자율적인 방향으로 발전시킬 것이다.

III. AI 기반 IDS 기술 동향

1. 글로벌 벤더 IDS 기술 개발 동향

AI 기반 네트워크 침입 탐지 기술은 변화하는 사이버 위협 환경과 기업 IT 인프라의 변화에 대해 끊임없이 진화하며 더욱 정교한 위협에 대응한다. 이러한 배경 속에서 글로벌 IDS 벤더들은 다음과 같은 동향을 보인다.

1.1 지도학습 모델 기반 개발 동향

지도학습 모델은 다양한 사이버보안 솔루션에 사용되고 있으며, 침입 탐지 AI 모델의 성능을 강화하여 정확한 위협 탐지 및 대응하고자 한다. 지도학습의 장점에도 불구하고, 데이터 라벨링 및 알려지지 않은 공격 탐지에 대한 제약으로 인해 비지도학습이나 준지도학습과 같은 보완적인 접근 방식으로 개발이 진행되고 있다.

다크트레이스는 비지도 기반 이상 탐지 기술로 널리 알려져 있지만, 이와 함께 사건 연관성 요약 및 맥락화 단계에서 지도학습을 병행하고 있다. 고객의 이메일, SaaS, 네트워크 환경에 영향을 미치는 의심스러운 활동 감지에 지도학습 기능인 Cyber AI Analyst를 사용하여 상관관계를 분석하고 연결한다[8]. 시스코는 Secure Network Analytics의 ML 기반 분석 파이프라인에서 지도학습과 비지도학습의 조합 등 다양한 분석 기술을 이용하는 접근 방식을 보인다[9]. 팔로알토 네트웍스(Palo Alto Networks)는 파일 및 트래픽에 대한 분류 모델에서 지도학습을 이용하여 실시간 분류 및 차단 기능을 제공하고 있다[10].

1.2 비지도학습 모델 기반 개발 동향

글로벌 벤더들은 시그니처나 라벨 없이 미지의 공격과 내부 이상 징후를 탐지하기 위한 모델을 개발하고 있다. 정상적인 네트워크 행위 패턴을 학습하여 일탈하는 모든 활동을 이상 행위로 탐지한다. 이는 제로데이 공격이나 변종 악성코드 등 알려지지 않은 위협에 대한 방어력을 크게 높인다.

다크트레이스는 Active AI 보안 플랫폼에서 조직 고유의 정상동작을 라벨 없이 학습하고 편차를 이상으로 판단하는 자가학습 AI를 통한 실시간 학습

및 대응을 강조한다[11]. 스플렁크는 K-means 클러스터링을 핵심 알고리즘 중 하나로 통합하여, 데이터 세트 내 패턴을 식별하고 이상 징후를 탐지할 수 있도록 한다[12]. 엑사빔(Exabeam)은 비정상적인 사용자 행동을 식별하고 내부자 위협 패턴을 분석하기 위해 Isolation Forest 모델을 비롯한 여러 모델을 사용한다[13].

1.3 준지도학습 모델 기반 개발 동향

준지도학습은 소량의 라벨 데이터와 대량의 비라벨 데이터를 모두 사용하여 지도학습과 비지도학습의 장점을 결합하여 새로운 위협을 탐지하는 데 활용된다. 글로벌 벤더들은 준지도학습과 같은 특정 학습 방법론을 내세우기보다는 AI, 딥러닝과 같은 포괄적인 용어를 사용하여 지능적이고 자율적인 위협 대응 능력을 강조하는 경향이 있다. 기술 자료 및 연구 결과들을 분석해 보면, 이와 유사한 접근법을 활용하여 데이터 라벨링의 한계를 극복하고 알려지지 않은 위협을 탐지하려는 동향이 확인된다.

다크트레이스는 위협 탐지 과정에서 비지도학습으로 이상 징후를 식별한 후, 일부 라벨 데이터를 활용하여 인시던트 심각도를 분석하는 접근법을 사용하고 있다[14]. 팔로알토 네트웍스의 공식 기술 문서에서는 머신러닝의 종류 중 하나로 준지도학습을 명시하고 있으며, 이는 라벨 데이터가 부족한 환경에서 활용성이 높다고 설명한다[15]. 또한 마이크로소프트 연구(Microsoft Research)의 논문에서는 자기지도학습과 능동 학습(Active Learning)을 비교하는 연구 결과를 발표해서 탐지 성능을 향상할 수 있다는 연구 결과를 제시한다[16].

신규 위협과 기존 위협이 고도화함에 따라 기존의 학습 모델만으로는 한계가 발생하고 있다. 이를 해결하기 위해 생성형 AI와 결합한 모델의 필요성이 제기되고 있다. 이러한 방법은 지도학습의 정확성과 비지도학습의 탐지 능력을 결합하고, 여기에 생성형 AI를 더해 데이터 부족 문제를 해결하고 새로운 위협에 대한 예측 및 대응을 강화하는 방향으로 발전할 것으로 예상된다.

2. AI 기반 IDS 연구 동향

최근 AI 기반 네트워크 침입 탐지 기술은 고도화된 사이버 공격에 대응하기 위해 다양한 AI 기법을 통합적으로 적용하는 방향으로 진화하고 있다. 본 절에서는 현재까지 발표된 주요 연구 흐름을 바탕으로 향후 연구 방향을 분석한다.

2.1 고도화된 AI 구조 IDS 적용

고도화된 AI(Advanced AI)는 일반적 AI 기술을 넘어서는 고급 ML 기술과 모델, 알고리즘을 포함하는 개념으로 사용되고 있다. 이는 자율성 및 적응성 향상을 위한 목적의 고도로 유능한 모델을 의미한다. IDS 기술에서도 Diffusion 모델, 트랜스포머(Transformer) 등 고도화된 AI 구조의 적용이 활발히 진행되고 있으며, 이러한 모델은 시계열 정보와 네트워크 내 상관관계를 효과적으로 학습할 수 있는 것으로 파악된다.

FlowTransformer[17]는 IDS에서 트랜스포머 아키텍처를 활용하여 플로우 기반의 탐지 성능을 향상하는 데 초점을 맞추고 있다. 긴 기간의 트래픽 패턴 및 관계 학습에 초점을 두어, 모델 크기와 입력 인코딩 방식 선택으로 정확도를 유지하면서 연산 효율을 향상하였다.

향후 Diffusion 모델과 같은 생성형 AI와 멀티모달 LLM 등의 고도화된 AI 기술은 IDS 기술 개발에 더 확산할 것이다. 이는 학습 모델의 강건성을 높이고 다양한 데이터 유형을 통합하여 이해함으로써 예측 불가능한 사이버 위협에 대해 자율적이고 능

동적으로 대응할 수 있을 것으로 전망된다.

2.2 암호화 트래픽 환경에 특화된 탐지

최근 암호화된 네트워크 트래픽에서 이상 행위를 탐지하는 모델에 대한 연구가 활발하게 이루어지고 있다. TLS 1.3과 같은 최신 암호화 프로토콜의 도입으로 인해 페이로드 기반 탐지가 어려워지면서, 패킷의 크기, 방향성, 전송 간격 등 비페이로드 메타데이터를 기반으로 한 AI 이상 탐지 알고리즘이 주목받고 있다.

MIETT[18]는 암호화된 네트워크 트래픽 흐름을 Bag of Packets로 취급해 패킷 간 상호 관계를 Two-Level Attention 구조로 학습하는 트랜스포머 기반 모델을 제시하였다. 제시된 방법을 통해 암호화된 트래픽 흐름 전체의 시각적 관계 패턴을 학습함으로써 암호화된 트래픽 분류 성능이 향상됨을 보여주었다.

향후 IDS는 암호화된 트래픽 환경에서 메타데이터 분석과 AI 기반 이상 탐지를 결합해 고도화될 것으로 보인다. 이를 통해, 장기 패턴과 복잡한 상관관계를 학습하고, 운영자가 신뢰할 수 있는 결과를 제공할 수 있을 것으로 기대된다.

2.3 설명가능성(XAI)과 실무 적용성 강화

보안 운영 환경에서 수집한 실제 데이터를 기반으로 학습하고, SHAP[19], LIME[20], DeepLIFT[21]와 같은 설명 가능한 AI(XAI: Explainable AI) 기법을 통합하는 방법이 탐지 결과의 신뢰성과 실무 적용 가능성을 향상시키고 있다. 이러한 접근은 보안 운영자의 경보 이해도를 높이고, 의사결정 기반을 명확하게 제공한다.

LENS-XAI[22]는 네트워크 침입 탐지 시스템에서 경량성과 설명 가능성, 확장성이라는 세 가지 주요 목표를 달성하기 위해 지식 증류와 VAE(Variational Autoencoders)를 사용하는 방법을 제시한다. 이는 XAI 기법을 통해 탐지 결과에 대한 명확한 설명을 제공함으로써 실제 IIoT 환경 적용 가능성을 확인하였다.

AI 기반 IDS에서 설명 가능성과 실무 적용 가능성은 점점 더 중요해질 것이다. 단순히 위협이라고 판단하는 것을 넘어, '왜' 그렇게 판단했는지를 보안 전문가에게 명확하게 제시함으로써 AI의 결정을 신뢰하고 공격의 특성을 더 깊이 이해할 수 있을 것이다.

2.4 경량화/증류 기법 AI IDS 배포

자원이 제약된 환경에서도 AI 기반 IDS가 효과적으로 동작할 수 있도록 모델 경량화 및 지식 증류 기술에 관한 연구 또한 진행되고 있다. 고성능 모델의 정확도는 유지하면서도 연산 효율성을 확보하여, 엣지(Edge) 단말 환경에서도 IDS의 활용 가능성을 높이고 있다.

GraphDART[23]는 고도화된 지속 위협을 탐지하기 위해 그래프 기반 모델링과 GNN을 활용하는 동시에, GNN의 높은 계산 비용 문제를 해결하기 위해 그래프 지식 증류를 도입하는 접근 방식을 제안한다. 또한, 별도의 교사 모델 없이도 모델 자체의 지식 증류를 활용하여 복잡한 모델을 경량화하는 접근법을 제시하기도 한다[24].

경량화/증류 기법의 IDS 적용은 향후 복합 모델의 효율적 경량화와 자기 지식 증류 고도화라는 연구 방향으로 나아갈 것으로 예상된다. 복잡한 AI 구조의 경량화와 성능 향상을 동시에 달성하는 연구를 통해 실시간 변화에 즉각적으로 적응할 수 있는 기술 개발이 가능할 것이다.

2.5 민감 데이터 보안성 강화

데이터 프라이버시 침해와 확장성에 대한 문제를 해결하기 위한 연구가 IDS에서도 증가하고 있으며,

이는 연합학습(Federated Learning)과 프라이버시 보존형 탐지에 초점을 맞추고 있다. 중앙 서버에 데이터를 전송하지 않고도 공동 학습이 가능하게 하여, 민감한 정보를 보호하는 동시에 공동의 위협 대응력을 강화할 수 있는 기반을 마련한다.

데이터가 생성되는 엣지에서 침입 탐지 모델을 직접 동작하여 민감한 데이터를 엣지에서 처리하고 탐지 결과와 같이 필요한 최소한의 정보만 클라우드로 전송하거나[25], 분산된 환경에서 참여자들의 민감한 데이터를 공유하지 않는 연합학습 프레임워크를 활용한 협력적 침입 탐지 시스템에 관한 연구가[26] 이러한 동향을 뒷받침하고 있다.

AI 기반 침입 탐지 분야는 탐지 정확도의 향상뿐만 아니라 설명 가능성, 배포 가능성, 개인정보 보호 측면에서의 기술적 진보를 함께 추구하고 있으며, 이는 실질적인 보안 운영 환경에서의 적용성과 신뢰도를 향상시키는 방향으로 나아가고 있다.

IV. 생성형 AI와 자기지도학습 기반 IDS 기술 발전 전망

생성형 AI와 자기지도학습은 최근 사이버보안 기술 발전의 핵심축으로 주목받고 있다. 자기지도학습은 기존 지도학습 방식의 한계를 극복하고, 실제 네트워크 환경의 복잡성과 역동성에 대응하기 위한 대안으로 자리 잡고 있다. 생성형 AI는 데이터의 복잡한 분포를 이해하고 재현하는 능력이 우수하다. 또한 텍스트, 이미지, 오디오 등 멀티모달 데이터를 동시에 이해하고 생성하는 부분에서 강점을 보인다.

AI 기반 IDS 기술은 이 두 기술의 결합을 통해 기존 IDS의 한계를 극복하고 예측 신뢰도를 높이는 데 이바지할 것으로 보인다.

자기지도학습과 생성형 AI를 융합하고자 하는 연구는 표 1과 같은 방향으로 활발하게 진행되고 있으며, 본 장에서는 IDS 분야와 관련된 연구 동향을 중점으로 향후 AI 기반 네트워크 침입 기술의 전망에 대해 논한다.

1. Pretext Task로서의 생성형 모델

생성형 AI와 자기지도학습을 융합하는 한 가지 강력한 방식은 생성형 모델을 자기지도학습의 프리텍스트 태스크(Pretext Task)로 활용하는 것이다. 이 접근 방식은 대량의 비라벨 데이터로부터 모델이

표 1 자기지도학습과 생성형 AI 융합 연구 동향 및 전망

활용 방식	내용	기대효과 및 전망
Pretext Task로서의 생성형 모델	• Diffusion 모델 등 생성형 AI가 데이터 재구성 → SSL 학습용 보조(Surrogate) 태스크 생성	학습 시나리오 다양화, 특징 표현 강화
생성형 AI 기반 SSL 데이터 증강	• 라벨 부족 환경에서 생성형 AI로 변형/가상 샘플 생성 → 비라벨 데이터 학습에 활용 • 영상(의료, 드론), 음성 등 라벨 비용 높은 분야	데이터 다양성 증가, 학습 견고성 향상
생성 모델 기반 SSL 학습	• 노이즈 복원을 자기지도학습 태스크로 사용	자연스러운 특징 학습, 잡음에 강한 표현 획득
멀티모달 SSL	• 텍스트 → 이미지/이미지 → 텍스트 생성 등의 양방향 생성 학습을 통한 Multi-Modal Representation 학습	멀티모달 데이터 이해 강화, 양방향 학습
소량 라벨 데이터 문제 극복	• 생성형 AI로 학습 표현을 보강하여 Few-Shot 상황에서도 SSL로 고성능 달성	소량 데이터 환경에서도 우수한 성능 유지

유용한 특징 표현을 학습하도록 유도하며, 특히 이상 탐지 분야에서 그 효과가 높을 것으로 분석된다.

TGAN-IDS[27]는 주로 GAN(Generative Adversarial Networks)을 자기지도학습의 프리텍스트 태스크로 활용하여 네트워크 침입 탐지 시스템의 성능을 향상하는 데 초점을 맞춘다. 라벨이 부족하거나 없는 네트워크 환경에서 GAN을 통해 정상 트래픽의 분포를 효과적으로 학습할 수 있으며, 라벨링 작업의 부담을 낮출 수 있다.

현재는 GAN, Diffusion 등 특정 생성 모델이 주로 활용되지만 향후 데이터 특성에 맞는 최적의 태스크를 제공하기 위한 다양한 생성 모델이 활용될 것으로 예측된다. 또한, 네트워크 트래픽, 시스템 로그 등 서로 다른 종류의 데이터를 위한 멀티모달 생성형 모델을 활용함으로써 더 높은 수준의 표현 학습을 가능하게 할 것이다.

2. 생성형 AI 기반 데이터 증강

자기지도학습에서는 데이터 증강의 필요성이 강조되는데, 이를 위해 생성형 AI로 데이터를 증강함으로써 라벨이 부족한 환경에서 모델 학습의 효율성을 높일 수 있다. 특정 분야의 데이터는 수집 자체가 어렵거나, 수집하더라도 전문적인 지식을 요구하는 라벨링 과정의 비용이 높고 오랜 시간이 소요된다. 생성형 AI는 이러한 실제 소량의 비라벨 데이터로부터 유사한 특징을 가진 가상 샘플을 생성하여 학습 데이터셋의 크기와 다양성을 확장할 수 있다. 또한, 민감한 정보를 포함하는 데이터의 프라이버시 보호 및 데이터 공유 제약을 극복하기 위해, 생성형 AI는 원본 데이터의 특징을 학습하여 통계적으로 유사하지만, 원본과 직접적으로 연결되지 않는 합성 데이터를 생성할 수 있다.

GPT 기반 데이터 증강 연구[28]에서는, IIoT 네트워크의 불균형한 데이터셋 문제해결을 위해 GPT와 SMOTE 같은 데이터 증강 기법을 비교하여, 여러 클래스의 공격을 탐지하는 모델의 성능을 평가하였다. NetDiffusion[29]은 최근에 주목받고 있는 Stable Diffusion을 파인 튜닝하여 프로토콜 사양을 준수하는 고품질의 합성 네트워크 트래픽을 생성한다. 텍스트-이미지 Diffusion 모델과 유사한 방식을 차용하여, GAN과 달리 전체적인 패턴뿐만 아니라 세부적인 종속성까지 포착하고자 하였다.

생성형 AI 기술의 발전에 따라 높은 현실성과 다양성이 확보된 데이터 증강 기법은 일반화 성능을 높이는 데 이바지할 것이다. 또한, 보안 전문가의 통찰을 반영하는 등 공격 시나리오 데이터의 다양화 등 활용 분야가 넓어질 것으로 전망된다.

3. 소량 라벨 데이터 문제 극복

소량의 라벨 데이터만 이용하는 퓨샷 학습(Few-Shot Learning)과 생성형 AI를 연계함으로써 네트워크 위협 탐지의 성능을 향상할 수 있다. 이는 라벨 데이터가 부족한 네트워크 환경에서 AI 모델을 효과적으로 구축하는 데 필수적이다. 생성형 AI는 비라벨 데이터로부터 실제와 유사한 합성 데이터를 생성하고, 자기지도학습은 생성형 AI가 추출한 복잡한 구조의 강력한 표현을 학습한다. 이후, 소량의 라벨 데이터만으로 효율적으로 파인 튜닝하거나 다운스트림 태스크에 사용될 수 있도록 학습의 틀을 제공한다.

퓨샷 학습과 생성형 AI를 결합하는 통합적 연구[30]를 통해 데이터가 부족하거나 신규 데이터가 지속적으로 발생하는 실제 네트워크 환경에서 AI 모델의 성능을 향상하기 위한 방법을 제시하였다. 생성형 AI는 기존의 소량 데이터로부터 실제와 유사한 가상 데이터를 생성하여 데이터셋의 규모를 확장

하고 다양성을 높인다. 생성형 AI가 만든 가상 데이터와 소량의 실제 데이터를 활용하여, 제로샷(Zero-Shot) 또는 퓨샷 학습은 새로운 클래스에 대해서도 빠르고 효과적으로 적응할 수 있는 방법을 제시한다.

AI 기반 IDS는 자기지도학습과 생성형 AI의 융합을 통해 한층 고도화될 것이다. 앞서 기술한 자기지도학습과 생성형 모델의 융합 연구와 함께, 설명가능한 AI 기법과 결합되어 보안 운영자가 탐지 결과를 이해하고 신뢰할 수 있게 될 것이다. 이러한 발전은 암호화된 트래픽 환경에서도 실시간 탐지와 자동 대응을 가능하게 하여, 향후 IDS가 지능적이고 자율적인 보안 체계의 핵심 요소로 자리 잡도록 할 전망이다.

V. 결론

AI 기술의 발전은 사이버보안 패러다임에 획기적인 전환점을 가져오고 있다. 기존의 보안 기술이 고도화된 위협에 한계를 보이는 상황에서, AI 기반 네트워크 침입 탐지 시스템은 자율적이며 실시간 대응 가능한 대안으로 부상하고 있다.

AI 기반 네트워크 침입 탐지 기술은 지도, 비지도, 준지도학습을 통해 알려진 위협은 물론, 미지의 제로데이 공격까지 효과적으로 탐지할 방법을 제시하고 있다. 최근에는 지도/비지도학습의 장점을 결합한 하이브리드 모델은 물론, 연합학습을 통한 데이터 프라이버시 보호, 생성형 AI를 활용한 공격 데이터 증강 등 새로운 패러다임이 연구되고 있다.

설명 가능한 AI, 경량화 모델, 프라이버시 보호형 학습 기법의 적용은 실제 보안 운영 환경에서의 신뢰성과 적용성을 높이고 있다. 앞으로의 사이버보안은 단순한 탐지를 넘어, 능동적 대응과 예측, 자율화로 나아가는 지능형 보안 체계로 진화할 것이다.

생성형 AI와 자기지도학습 기술은 비라벨 데이터 기반 탐지 성능을 높이며, 데이터 부족 문제해결과 고급 위협 예측에 이바지한다. 이러한 융합은 제로데이 공격이나 내부자 위협 등 예측 어려운 위협에 효과적이다.

본고는 이러한 사이버보안 기술 변화 흐름 속에서 AI 기반 침입 탐지 기술의 현황과 발전 방향을 제시함으로써 사이버 위협을 위한 효과적인 대응 전략 마련에 이바지하고자 한다. 지속적인 기술 고도화와 함께 인간 중심의 협력적 AI 보안 체계 구축은 향후 사이버보안의 핵심 과제가 될 것이다.

용어해설

IDS 네트워크 트래픽과 디바이스에서 알려진 악의적 활동을 모니터링하는 보안 도구

NDR 네트워크상의 이상행동이나 위협을 탐지하고 신속하게 대응하기 위한 보안 솔루션

SIEM 다양한 시스템 및 장비에서 발생하는 로그와 이벤트를 분석하여 탐지 대응하는 솔루션

참고문헌

[1] Gartner Research, “Emerging Tech: Tech Innovators in Preemptive Cybersecurity,” 2025. 1. 8.

[2] AhnLab, “2025년 보안 트렌드를 읽는 10개의 키워드,” 2025. 6. 24. https://ahnlab.com/contents/content-center/35883

[3] Gartner Research, “Emerging Market Guide for Managed Detection and Response Services,” 2023. 2. 14.

[4] https://www.sangfor.com/blog/cybersecurity/13-best-network-detection-and-response-ndr-solutions

[5] J.B. Grill et al., “Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning,” in Proc. Int. Conf. Neural Inf. Process. Syst., (Vancouver, BC, Canada), Dec. 2020, pp. 21271-21284.

[6] T. Chen et al., “A Simple Framework for Contrastive Learning of Visual Representations,” in Proc. Int. Conf. Mach. Learn., Jul. 2020, pp. 1597-1607.

[7] R.D. Hjelm et al., “Learning Deep Representations by Mutual Information Estimation and Maximization,” in Proc. Int. Conf. Neural Inf. Process. Syst., (Vancouver, BC, Canada), Dec. 2019, pp. 15535-15545.

[8] https://www.darktrace.com/blog/detecting-attacks-across-email-saas-and-network-environments-with-darktraces-combined-ai-approach

[9] https://www.cisco.com/c/en/us/products/collateral/security/stealthwatch/white-paper-c11-740605.html

[10] https://docs.paloaltonetworks.com/advanced-wildfire

[11] https://www.darktrace.com/blog/how-ai-is-transforming-cybersecurity-practices

[12] https://docs.splunk.com/Documentation/MLApp/5.5.0/User/SmartClusteringAssistant

[13] https://www.exabeam.com/explainers/siem/siem-analytics/

[14] https://www.darktrace.com/blog/ai-uncovered-introducing-darktrace-incident-graph-evaluation-for-security-threats-digest

[15] https://www.paloaltonetworks.com/cyberpedia/artificial-intelligence-ai

[16] B.E. Afshar et al., “To Label or to Pseudo Label? Active Learning vs Semi-Supervised Learning for Windows Malware Prediction,” in Proc. Can. Conf. Artif. Intell., (Ontario, Canada), May, 2024.

[17] L.D. Manocchio et al., “Flowtransformer: a transformer framework for flow-based network intrusion detection systems,” Expert Syst. vol. 241, 2024, pp. 122564.

[18] X.Y. Chen et al., “MIETT: Multi-Instance Encrypted Traffic Transformer for Encrypted Traffic Classification,” in Proc. AAAI Conf. Artif. Intell. Conf. Innov. Appl. Artif. Intell. Symp. Educ. Adv. Artif. Intell., (Philadelphia, PA, USA), Feb. 2025, pp. 15922-15929.

[19] S.M. Lundberg and S.I. Lee, “A Unified Approach to Interpreting Model Predictions,” Int. Conf. Neural Inf. Process. Syst., (Long Beach, CA, USA), Dec. 2017, pp. 4768-4777.

[20] M.T. Ribeiro et al., “Why Should I Trust You?,” Explaining the Predictions of Any Classifier,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., (San Francisco, CA, USA), Aug. 2016, pp. 1135-1144.

[21] A. Shrikumar et al., “Learning Important Features Through Propagating Activation Differences,” in Proc. Int. Conf. Mach. Learn., (Sydney, NSW, Australia), Aug. 2017, pp. 3145-3153.

[22] M.A. Yagiz et al., “LENS-XAI: Redefining Lightweight and Explainable Network Security through Knowledge Distillation and Variational Autoencoders for Scalable Intrusion Detection in Cybersecurity,” arXiv preprint, 2025. doi: 10.48550/arXiv.2501.00790

[23] S.F Rabooki et al., “GraphDART: Graph Distillation for Efficient Advanced Persistent Threat Detection,” arXiv preprint. 2025. doi: 10.48550/arXiv.2501.02796

[24] S. Yang et al., “A Lightweight Approach for Network Intrusion Detection Based on Self-Knowledge Distillation,” in Proc. IEEE Int. Conf. Commun., (Rome, Italy), May. 2023, pp. 3000-3005.

[25] S. Almabdy and A. Ullah, “Optimising Intrusion Detection Systems in Cloud-Edge Continuum with Knowledge Distillation for Privacy-Preserving and Efficient Communication,” in Proc. IEEE 9th Int. Conf. Fog Edge Comput., (Tromso, Norway), May. 2025, pp. 1-5.

[26] C. Ezelu and U. Buehler, “Privacy-preserving collaborative intrusion detection systems: A federated learning framework,” in Proc. Int. Conf. Comput. Sci. Comput. Intell., (Las Vegas, NV, USA), Dec. 2022, pp. 1076-1079.

[27] X. Zhang et al., “Dual Generative Adversarial Networks Based Unknown Encryption Ransomware Attack Detection,” IEEE Access, vol. 10, 2022, pp. 900-913.

[28] F.S. Melícias et al., “GPT and Interpolation-Based Data Augmentation for Multiclass Intrusion Detection in IIoT,” IEEE Access, vol. 12, 2024, pp. 17945-17965.

[29] X. Jiang et al., “NetDiffusion: Network Data Augmentation Through Protocol—Constrained Traffic Generation,” Proc. ACM Meas. Anal. Comput. Syst., vol. 8, no. 1. 2024, pp. 1-32.

[30] V.K. Gali and Er.R. Agarwal, “Zero-Shot Learning and Few-Shot Learning with Generative AI: Bridging the Data Gap for Real-World Applications,” Integrated J. Res. Arts Humanit., vol. 5, no. 1, 2025, pp. 193-200.