Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.34, NO.5, Oct. 2024 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2024.34.5.875

AI 기반 NIDS에 대한 모델 종류 추론 공격*

안 윤 수,[1†] 김 도 완,[1] 최 대 선[2‡]
[1,2]숭실대학교 (연구원, 교수)

Model Type Inference Attack against AI-Based NIDS*

Yoonsoo An,[1†] Dowan Kim,[1] Dae-seon Choi[2‡]
[1,2]Soongsil University (Researcher, Professor)

요 약

IoT 네트워크의 증가로 인해 사이버 공격도 함께 증가하고 있으며, 네트워크 침입탐지 시스템(NIDS)의 중요성이 강조되고 있다. 전통적인 NIDS의 한계를 극복하고 더욱 고도화된 사이버 공격에 대응하기 위해 NIDS에 인공지능 모델을 도입하는 추세이다. 인공지능 모델 기반의 NIDS는 인공지능 알고리즘이 가지는 적대적 공격에 대한 취약성을 가지게 된다. 모델 종류 추론 공격은 모델 내부의 정보를 추론하는 적대적 공격의 일종이다. 본 논문은 기존 모델 종류 추론 공격을 더욱 현실적인 가정을 적용하며, NIDS 모델을 타겟으로 최적화된 모델 종류 추론 공격 프레임워크를 제안한다. 제안하는 방식으로 NIDS 모델의 종류를 추론하는 공격 모델을 약 0.92의 분류 정확도를 보이도록 훈련할 수 있었으며, 인공지능 기반 NIDS에 대한 새로운 보안 위협을 제시하며 이에 대한 방어 기술 개발의 중요성을 강조한다.

ABSTRACT

The proliferation of IoT networks has led to an increase in cyber attacks, highlighting the importance of Network Intrusion Detection Systems (NIDS). To overcome the limitations of traditional NIDS and cope with more sophisticated cyber attacks, there is a trend towards integrating artificial intelligence models into NIDS. However, AI-based NIDS are vulnerable to adversarial attacks, which exploit the weaknesses of algorithm. Model Type Inference Attack is one of the types of attacks that infer information inside the model. This paper proposes an optimized framework for Model Type Inference attacks against NIDS models, applying more realistic assumptions. The proposed method successfully trained an attack model to infer the type of NIDS models with an accuracy of approximately 0.92, presenting a new security threat to AI-based NIDS and emphasizing the importance of developing defence method against such attacks.

Keywords: Deep Learning, Network Intrusion Detection System, Adversarial Attack

I. 서 론

IoT 네트워크의 사용 증가 및 발달로 인해 여러 사이버 공격이 증가하는 추세에 따라 네트워크 침입탐지 시스템(Network Intrusion Detection System)의 중요성이 강조되고 있다[1]. 전통적 NIDS는 여러 한계점을 가지고 있다. 전통적 NIDS는 주로 이전에 정의된 규칙 혹은 피처 기반으로 공격을 탐지하기 때문에, 변형된 공격을 탐지하기 힘들다는 특징을 가지고, 정상 트래픽을 침입으로 판단하거나 침입을 탐지하지 못하는 등 오류율이 높다는 문제점이 있다[2]. 또한 네트워크의 복잡성 증가 및 대용량 트래픽으로 인해 모든 트래픽을 효과적으로 모니터링 및 분석하기 어렵다는 한계를 가진다.

Received(06. 24. 2024), Modified(1st: 07. 30. 2024, 2nd: 09. 03. 2024), Accepted(09. 03. 2024)
* 이 논문은 2024년 정부(방위사업청)의 재원으로 국방기술진흥연구소의 지원을 받아 수행된 연구임(KRIT-CT-21-037)
† 주저자, ac.yoonsoo@gmail.com
‡ 교신저자, sunchoi@ssu.ac.kr(Corresponding author)

이러한 한계를 극복하기 위해 인공지능(Artificial Intelligence) 기반의 네트워크 침입 탐지 시스템이 연구 및 도입되고 있다[3, 4]. 인공지능 기반의 네트워크 침입 탐지 시스템은 정상 및 비정상 네트워크 행위를 구분하기 위해 복잡한 패턴을 대규모 데이터에서 학습할 수 있으며, 비교적 변형된 공격을 탐지하는 것에도 유리하다는 장점을 가진다.

그러나 인공지능 기반의 네트워크 침입 탐지 시스템은 인공지능 알고리즘이 가질 수 있는 보안 취약점에 노출될 가능성이 있다[5]. 인공지능 알고리즘은 적대적 예제[6, 7], 오염 공격[8, 9], 모델 추출[10], 멤버십 추론 공격[11] 등의 여러 유형의 적대적 공격에 노출될 수 있다는 것이 대표적인 보안 취약점이다. 이러한 적대적 공격들은 AI 모델을 속여 네트워크 보안 시스템의 탐지 성능을 저하시킬 수 있으며, 정보를 유출 시키거나 타겟 모델을 불법적으로 복제해 이용 가능하게 하는 등 심각한 보안 위협을 초래할 수 있다[12, 13].

최근 연구[14]에 따르면 공격자가 AI 기반 분류 모델의 네트워크 종류를 추론할 수 있다는 것이 밝혀졌다. 모델 종류 추론 공격은 타겟 AI 모델의 구조를 공격자가 알 수 있게 하기 때문에, 공격자는 타겟 모델의 작동 방식을 파악해 해당 모델이 어떤 패턴에 의존하는지를 파악 가능하게 한다. 모델의 아키텍쳐를 공격자가 알 수 있게 되면 모델의 아키텍쳐와 파라미터를 모두 알 수 있는 화이트 박스 상황에 가까워진다. 공격자가 모델에 대한 모든 정보를 알 수 있게 되면 적대적 공격을 할 수 있게 되므로, 타겟 모델에 대한 정보를 모두 알기 위해 공격자는 모델 탈취 공격 등의 기법[15, 16]을 적대적 공격의 중간 단계에 포함시키기도 한다. 따라서 모델의 종류 정보를 추론하는 공격 역시 타겟 모델에 더욱 효과적으로 작동하는 정교한 공격 방법을 개발하는 것에 기여할 여지가 있다.

기존의 AI 모델 종류 추론 공격은 이미지 분류 모델을 타겟으로 삼고, 타겟 모델의 모든 클래스에 해당하는 데이터를 쿼리해 얻은 confidence score 벡터를 기반으로 모델의 종류를 추론하기 위해 타겟 모델이 될 수 있을 만한 구조를 가진 여러 개의 후보 모델군을 훈련시키고, 후보 모델군에 동일한 데이터셋을 입력해 얻은 출력값과 해당 출력값을 가공한 피쳐로 후보 모델군의 종류를 분류하는 공격 모델을 훈련한다. 기존 연구는 모델의 네트워크 구조에 따라 출력값이 가지는 특징이 다르며, 공격자가 그 점을 이용해 모델의 종류를 효과적으로 추론할 수 있음을 밝혔으나 후보 모델군을 훈련할 때 사용된 데이터셋을 그대로 다시 후보 모델군에 쿼리한 출력값으로 공격 모델을 훈련했기 때문에, 공격자가 타겟 모델을 훈련한 데이터의 분포를 안 상태에서 공격 모델을 훈련해야 한다는 현실적이지 못한 가정이 전제된다는 한계를 가진다.

본 연구는 기존 연구의 한계를 극복함과 동시에 NIDS 모델의 특징에 최적화된 모델 종류 추론 프레임워크를 제안한다. NIDS 모델에 쿼리 할 때 공격으로 탐지될 수 있는 네트워크 패킷이 지속적으로 입력되면 비정상적 활동이 감지될 확률이 높아지며 모델의 반응을 지속적으로 관찰하기 어려워진다. 따라서 NIDS 모델을 타겟으로 공격을 수행할 때에는 NIDS 모델에 입력했을 때 정상으로 분류될 수 있는 네트워크 패킷 데이터만을 이용해 쿼리하는 것이 현실적인 가정이다. 또한 기존의 AI 모델 종류 추론 공격 실험에서 공격자가 쿼리의 결과로 클래스별 confidence score가 높은 순서대로 정렬한 순위 벡터만 얻을 수 있는 블랙박스 상황[17]에서 본 연구의 실험을 진행한다. 현실적인 공격 시나리오를 시뮬레이션하기 위해 공격자가 훈련한 후보 모델군에 사용되지 않은 데이터셋을 쿼리한 출력값을 이용해 공격 모델을 훈련하고, 후보 모델군과 공격 모델 모두에 사용되지 않은 데이터셋을 후보 모델군에 쿼리한 데이터셋으로 공격 모델의 성능을 측정한다. 공격 모델을 훈련하고 모델 종류 추론 공격을 수행했을 때 공격 모델의 분류 정확도는 약 0.9155의 성능을 보였다. 본 논문의 기여는 다음과 같다.

- 최초로 NIDS 모델에 대한 모델 종류 추론 공격을 수행하였다.
- 한 클래스데이터만으로 쿼리해 공격했으며, 타겟 모델 훈련에 쓰인 데이터를 공격자가 알 수 없는 상황을 가정하기 위해 후보 모델군과 공격 모델의 훈련에 다른 데이터셋을 사용해 실험하였다.
- 정상 네트워크 데이터만을 쿼리에 사용해 공격이 탐지되지 않는 상황에서 NIDS 모델의 반응을 지속적으로 관찰할 수 있도록 실험했다.

본 논문에서는 기존의 연구보다 현실적인 가정을 적용하고, NIDS 모델에 최적화된 프레임워크로 모델 종류 추론 공격을 수행하며 타겟 NIDS 모델의 적대적 공격에 대한 취약점을 드러내고 이를 통해 네

트워크 침입 탐지 시스템의 보안 기술 강화의 필요성을 강조한다.

II. 배경지식

2.1 AI 기반 NIDS 모델

2.1.1 AlertNet

2019년 R. Vinayakumar et al.은 논문[18]에서 다중의 은닉층을 사용해 높은 차원의 데이터를 학습할 수 있는 DNN 구조의 NIDS 모델 AlertNet을 제안하였으며, 네트워크 기반(NIDS)과 호스트 기반의 침입탐지 시스템(HIDS)[19]을 결합한 하이브리드 접근 방식을 구현하였다. physical 네트워크 데이터를 딥 러닝에 적합한 데이터로 변환하기 위해 N-gram[20] 등의 기술을 적용하였으며, 실시간으로 대량의 네트워크 데이터를 처리할 수 있는 구조로 설계되었다. 또한 continual learning[21]을 통해 새로운 데이터를 지속적으로 학습하기 때문에 새로운 유형의 사이버 공격에 적응할 수 있는 프레임워크이다.

입력층, 특징추출 층, 은닉층, 배치정규화 및 드랍아웃, 출력층으로 이루어진 5층의 레이어 구조를 가지며 각 레이어가 층이 깊어질수록 적은 수의 뉴런을 가지도록 설계되었다. 복잡한 네트워크 환경에서의 활용을 목표로 하며, 논문에서 KDDCup99[22], KSL-KDD[23] 등의 여러 데이터셋에서 성능을 검증했다.

2.1.2 DeepNet

Minghui Gao 외는 2020년 DNN 기반 NIDS 모델 DeepNet을 제안하였다[24]. DNN 모델에 입력하기 적절한 형식으로 네트워크 데이터를 처리하기 위해 Apriori 알고리즘을 통해 연관분석해 데이터가 가진 규칙을 생성하였고, 해당 규칙을 기반으로 DNN이 정상 트래픽과 악의적 트래픽을 분류한다. 해당 모델은 AlertNet과 아주 유사한 구조로 입력층 및 은닉층 및 드랍아웃 레이어, 출력층으로 이루어진 4층의 레이어 구조를 가지며 모든 레이어가 같은 수의 뉴런을 가지고 있어 모든 레이어에서 동일한 특징 추출 및 변환을 거친다. NSL-KDD 데이터셋과 CIC-IDS2017[25] 데이터로 성능을 검증하였으며, 트래픽의 복잡한 패턴에서 유의미한 정보를 추출하고 훈련했다.

2.1.3 IdsNet

B.E. Zolbayar et al이 발표한 논문[26]에서 실험에 사용하기 위해 자체적으로 개발한 IdsNet 또한 DNN 기반의 NIDS 모델로, 네트워크에서 트래픽 데이터를 수집하고 피쳐를 추출해 모델이 입력받을 수 있는 형태로 변환한다. 데이터를 전처리 할 때에는 스케일링 및 정규화 등의 과정을 거친다. 해당 논문은 NIDS의 아키텍처를 최적화하는 과정에 대한 실험도 포함하고 있는데, 데이터셋과 task의 복잡성에 따라 DNN의 최적의 히든 레이어 수를 결정하는 방법론에 초점을 두고 있다. 실험 결과, 3개 이상의 레이어를 추가하는 것이 정확도를 높이지 않았으며, 계산 비용의 증가와 과적합을 유발했다는 결론을 내렸다. 따라서 IdsNet은 해당 논문에서 가장 효율이 높다고 결론 내린 바에 따라 3층의 레이어를 가진 모델로 설계되었다.

2.2 AI 모델에 대한 적대적 공격

2.2.1 Membership Inference Attack

Shokri et al. 이 2017년에 제안한 멤버십 추론 공격은 특정 데이터 샘플이 모델의 훈련 데이터셋에 포함되었는지 여부를 판단하는 공격이다. 해당 공격은 모델이 훈련하지 않은 데이터와, 훈련 데이터에 대해 다르게 반응한다는 점을 이용한다. 쿼리 액세스를 가진 공격자는 타겟 모델의 멤버십을 추론하기 위해 공격자가 훈련한 대체 모델에 특정 데이터 샘플을 입력해 예측한 confidence score vector를 기반으로 멤버십을 추론하는 공격 모델을 훈련한다. Salem et al. 은 2019년 멤버십 추론 공격을 다양한 타겟 모델과 시나리오에서 실험해, 일반화된 멤버십 추론 공격이 가능함을 입증하였다. 멤버십 추론 공격은 의료, 금융 등 민감한 정보를 다루는 분야에서 프라이버시에 심각한 위협이 될 수 있다.

2.2.2 Model Type Inference Attack

멤버십 추론 공격과 유사한 공격 기법으로 모델 종류 추론 공격이 제안되었다. 모델 종류 추론은 타

겟 모델이 어떤 구조의 네트워크로 훈련되었는지 유추하는 것이 목적이다. 타겟 모델이 널리 알려진 구조를 가진 인공지능 분류 모델일 경우, 공격자는 타겟 모델일 것이라고 예상되는 후보 모델군을 훈련하고 후보 모델군의 훈련에 쓰인 데이터셋과 같은 분포를 가진 모든 클래스에 해당되는 샘플 데이터셋을 후보 모델군에 쿼리한다. 해당 논문은 타겟 모델에 쿼리한 결과로 confidence score vector를 얻을 수 있는 상황을 그레이박스, confidence score가 높은 순서대로 얻은 클래스 순위 정보만 얻을 수 있는 상황을 블랙박스라고 정의하였다. 각 환경에서 얻을 수 있는 출력값을 이용해 가공한 피처와 출력값을 데이터 셋으로 사용해 공격 모델을 훈련하고 성능을 측정하였으며, 블랙박스 및 그레이박스 환경에서 모델의 종류를 추론할 수 있음을 보였다. 모델 종류 추론 공격은 공격자가 타겟 모델의 네트워크 구조를 알아내고 모델의 동작 방식을 알아낼 수 있게 한다. 공격자는 추론한 타겟 모델의 정보를 기반으로 타겟 모델에 더욱 최적화된 공격 방법을 개발할 수 있으므로, 모델 종류 추론 공격에 대한 방어가 중요하다.

2.2.3 Model Attribute Inference Attack

Property inference attack은 AI 모델의 훈련 데이터의 통계적 특성을 추론하는 공격으로, 유사한 데이터셋과 알고리즘으로 훈련한 모델이 유사한 기능을 할 것이라는 직관을 이용한다. Ateniese et al. (2015)은 데이터셋의 특정 속성을 추론하는 연구를 수행하였으며, 해당 공격 방법을 제안하였다. Melis et al. (2019)은 협력 학습(splitting learning) 환경은 모델의 파라미터를 업데이트 하는 동안 의도치 않은 정보 유출이 발생할 수 있다. 이때 공격자가 개입하는 방식에 따라 크게 수동 속성 추론, 능동 속성 추론의 두 가지 유형으로 나뉜다. 이러한 연구는 데이터 기밀성 유지의 중요성을 강조하며, 데이터 마스킹, 차분 프라이버시 등의 방어 기법이 필요함을 시사한다.

III. NIDS 모델 종류 추론 방법

NIDS 모델은 네트워크 트래픽을 모니터링하고 비정상적인 활동을 탐지한다. 타겟 NIDS 모델의 종류를 추론하기 위해 모든 클래스 데이터에 대한 쿼리 응답을 사용하는 기존의 모델 종류 추론 공격 방법을 그대로 사용할 시 한 사용자가 지속적으로 공격 클래스에 해당하는 네트워크 패킷을 전송하게 되므로 NIDS가 네트워크를 보호하기 위해 설정에 따라 해당 사용자의 트래픽을 차단할 가능성이 있다. 따라서 타겟 모델이 NIDS 모델일 경우, 탐지하기 어려운 공격을 수행하기 위해 본 논문에서 제안하는 프레임워크와 같이 공격자가 정상 클래스에 해당하는 네트워크 패킷을 사용해 NIDS를 쿼리하는 것이 바람직하다. 공격에 정상 클래스 네트워크 패킷만을 사용하는 경우 여러 이점이 있다. 공격자

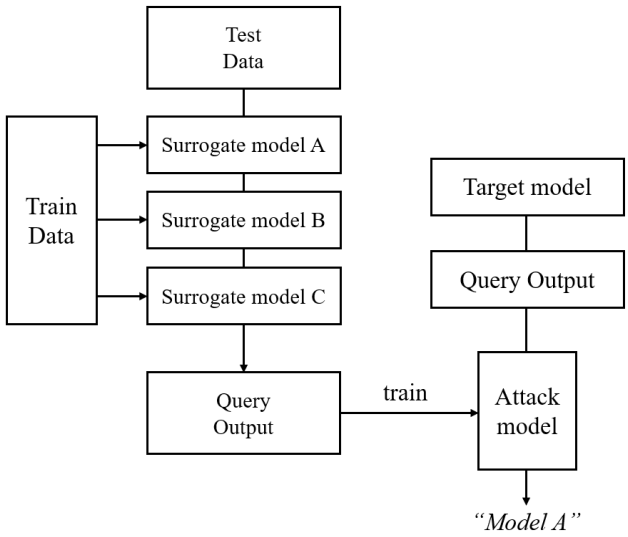

Fig. 1. Model type inference attack structure diagram against NIDS model

가 쿼리하는 동안 NIDS가 공격자의 활동을 악의적인 행위로 간주할 가능성을 낮추고, 더 긴 시간 동안 NIDS 모델에 쿼리한 데이터를 수집할 수 있게 된다. 따라서 공격자가 정상 네트워크 패킷만으로 쿼리해 정보를 추출하는 것이 탐지를 피하며 타겟 모델의 구조를 파악하는 현실적인 공격 시나리오가 된다. 본 논문은 공격자가 타겟 모델의 출력으로 분류 확률에 따른 클래스 순위만 얻을 수 있는 블랙박스 환경에서 정상 네트워크 패킷 데이터만으로 타겟 모델의 종류를 추론하는 NIDS 모델에 대한 공격 프레임워크를 제안한다. 그림 1은 NIDS 모델에 대한 모델 종류 추론의 구조도이다. 또한 기존의 모델 종류 추론 공격이 후보 모델군의 훈련에 사용된 데이터셋을 쿼리한 출력값을 사용해 공격 모델을 훈련시킨 것과 달리 후보 모델군의 훈련에 사용되지 않은 테스트 데이터로 쿼리한 출력값으로 공격 모델을 훈련해 기존 공격의 데이터 의존성을 해소하고자 했다.

3.1 후보 모델군 훈련

NIDS는 네트워크 트래픽 데이터에서 특징을 추출해서 정상/비정상 트래픽 간 피쳐들의 차이를 기반으로 분류하는 경우가 많다. 본 논문의 실험에서도 훈련의 효율성을 높이기 위해 후보 모델군을 전처리를 통해 피쳐화한 데이터를 활용해 훈련한다.

공격자는 타겟 모델의 종류를 추론하기 위해 타겟 모델이 될 수 있을 만한 후보 모델을 선정하고 각각 동일한 네트워크 데이터셋을 활용해 후보 모델군을 훈련한다. 해당 모델군에 같은 데이터를 쿼리한 클래스별 확률의 순위 정보 벡터를 얻는다. 해당 벡터는 모델의 종류별 출력값 간 차이의 패턴을 학습해 모델의 종류를 분류하는 공격 모델을 훈련하는 데 사용된다. 실제 환경에서 사용되는 NIDS 모델은 안정적이면서 높은 탐지 성능이 보장되어야 할 것이므로, 실험에서 실제 환경과 비슷한 상황을 시뮬레이션하기 위하여 후보 모델군들의 성능을 모두 높은 수준에 도달할 때까지 미세조정을 하며 훈련시켜야 한다.

3.2 피쳐 가공

모델 종류 추론 공격 논문에서 정의한 것과 같이 NIDS 모델에서 confidence score가 높은 순서대로 정렬된 클래스 순위를 얻을 수 있는 블랙박스 환경을 가정한다. NIDS 모델에 쿼리해 confidence score vector를 얻을 수 있는 그레이박스 상황에서의 실험은 제외한다.

후보 모델군에 쿼리한 confidence score가 높은 순서대로 정렬된 클래스의 순위 벡터에서 중간 순위보다 낮은 순위의 클래스에는 0, 높은 순위의 클래스에는 1을 대체한 벡터를 훈련 데이터에 포함될 피쳐로 사용한다. 해당 벡터는 전체 클래스의 confidence score가 어떤 클래스에서 상대적으로 크고 작은지를 강조한다. 그림 2는 피쳐 가공 과정을 나타낸 구조도이다.

3.3 공격모델 훈련

훈련한 후보 모델군에 쿼리한 클래스 순위 벡터와 가공한 피쳐를 훈련 데이터로, 후보 모델군의 모델 종류를 label로 MLP 모델을 훈련한다. 이때 훈련된

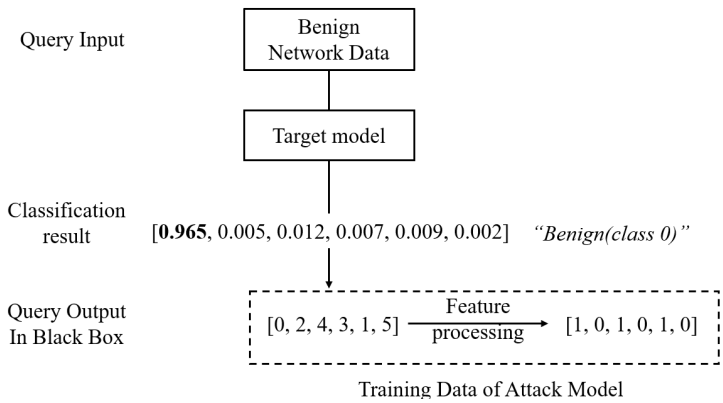

Fig. 2. Feature Processing Structure Chart

공격 모델은 타겟 NIDS 모델에 정상적인 네트워크 패킷 데이터로 쿼리해 얻은 응답을 입력하면 후보 모델군 중 타겟 NIDS 모델이 어떤 모델과 같은 종류인지를 추론할 수 있는 공격 모델로 사용된다. 공격 모델의 성능이 높을수록 정확하게 타겟 모델의 종류를 추론한다. 훈련된 모델이 테스트 데이터 셋에서도 높은 성능으로 작동하는지 확인한다. 테스트 데이터셋은 후보 모델군이 훈련하지 않은 정상 네트워크 패킷 데이터를 후보 모델군에 쿼리하고 피쳐를 가공해서 concatenate 한 데이터를 사용한다. 후보 모델군의 훈련에 사용한 데이터로 쿼리한 출력으로 공격모델을 훈련하고, 후보 모델군의 훈련에 사용되지 않은 데이터로 쿼리한 결과를 공격 모델에 입력해도 공격 모델이 모델의 종류를 분류할 수 있다면 타겟 모델을 훈련한 데이터의 분포를 공격자가 모르는 상황에서도 공격자가 모델의 종류를 분류할 가능성이 높아진다.

IV. NIDS 모델 종류 추론 실험

4.1 데이터 전처리 및 후보 모델군 훈련

NIDS 모델 종류 추론 실험에서 추론하고자 하는 타겟 모델이 될 만한 후보 모델군의 훈련에는 CIC-IDS2017 데이터셋을 사용한다. CIC-IDS2017 데이터셋은 실제 데이터(PCAP)와 유사한 최신 공격을 포함한 양성 네트워크 패킷 데이터와 머신러닝 학습을 위한 피쳐 데이터를 함께 제공한다. Sharafaldin, et al. 2016 이 제안한 B-Profile 시스템을 사용해 실제 인간 상호작용과 유사한 양성 트래픽을 생성했다. 공격 패킷은 DDos[27], Dos, Heartbleed[28], Brut Force SSH, Brut Force FTP, Web Attack, Infiltration, Botnet 공격을 구현한 데이터가 포함되어 있다.

본 논문의 실험에서는 실제와 유사한 상황을 시뮬레이션하고 실제 데이터를 다루는 기술을 구축하기 위해 실질 환경의 데이터인 PCAP 에서 CICFlowmeter로 labelling을 진행했고, botnet 공격은 다수의 봇으로 NIDS 모델을 공격하는 시나리오이기 때문에 훈련데이터에서 제외했다. 또한 사용자가 특정될 수 있는 식별정보 등의 네트워크 구성 정보가 포함된 column은 모델의 일반화 및 성능 향상을 위해 삭제한다. 위와 같은 방법으로 처리한 데이터는 모델 종류 추론의 시뮬레이션을 위한 후보 모델군의 훈련 데이터로 사용한다.

본 논문은 멀티 클래스 AI 기반 분류기인 AlertNet, IdsNet, DeepNet을 후보 모델군으로 설정하였으며, CIC-IDS2017 데이터셋 중 train data에 해당하는 862,648개의 데이터를 전처리한 뒤 모델의 훈련에 사용했다. 미세조정 및 충분한 훈련을 거친 후보 모델군의 성능은 표 1과 같다. 각 모델을 훈련한 CIC-IDS2017 데이터 셋을 라벨링 한 결과 2번 클래스에 해당하는 Heartbleed 공격데이터가 단 한 개의 샘플이며, 대다수의 데이터셋이 0 혹은 1에 편중된 클래스 불균형이 있었기 때문에 2번 클래스에 대한 예측 성능은 떨어지나, 세 가지 모델 모두 정확도 0.9 이상으로 전반적으로 공격을 높은 성능으로 탐지할 수 있다. CIC-IDS2017 데이터셋은 Benign에 해당하는 정상 클래스 데이터가 가장 많이 포함되어 있다. 따라서 같은 클래스에 대한 쿼리 일지라도 모델의 종류에 따라 다른 양상을 가진다면, 너무 적은 수의 샘플을 가지는 클래스의 데이터보다 충분히 훈련이 될 만큼의 샘플을 가지는 0 클래스 데이터로 쿼리를 하는 것이 NIDS의 탐지를 피할 수 있으며, 결과를 일반화하기 적절할 것이다.

Table 1. Performance of Candidate models

	AlertNet			IdsNet			DeepNet		
metric / class	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.99	1.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	1.00	0.50	0.67	1.00	0.07	0.13	0.83	0.36	0.50
4	1.00	0.99	1.00	1.00	0.99	0.99	1.00	0.99	1.00
5	0.98	0.08	0.16	0.95	0.97	0.96	0.96	0.99	0.97

4.2 공격 모델 성능

공격자가 타겟 모델이 훈련된 것과 동일한 데이터셋을 확보할 확률은 매우 낮기 때문에 후보 모델군을 훈련한 데이터와 공격 모델의 훈련에 사용될 쿼리 데이터는 겹치지 않을수록 모델 종류 추론 공격이 일반화될 수 있다. 따라서 공격 모델의 훈련에 CIC-IDS2017 데이터셋 중 test data에 해당하는 369,707개의 데이터 중 일부를 AlertNet, IdsNet, DeepNet에 각각 쿼리해 얻은 클래스별 확률 순위와 가공한 피처를 훈련 데이터를 이용해 공격 모델을 훈련한다. 이후 공격 모델을 테스트 할 때에는 후보 모델군 및 공격 모델 훈련에 쓰이지 않은 데이터셋을 사용해 후보 모델군에 쿼리하고, 해당 데이터로 모델의 종류를 추론한다. 실험에서는 CIC-IDS2017 데이터셋중 테스트 데이터셋 360,707개의 데이터 중 50,000개를 이용해 공격 모델을 훈련한 뒤 성능을 측정하였고, 15,000개의 데이터를 공격 모델의 테스트에 사용해 성능을 측정하였다.

후보 모델군에 쿼리해 얻은 클래스별 확률 순위 벡터와 가장 높은 순위부터 중간 순위의 클래스에는 1을, 그 외의 클래스에는 0으로 대체한 피쳐 벡터를 concatnenate 해서 훈련 데이터는 (50,000, 12) 형태를 가지며, 테스트 데이터는 (1000, 12) 형태를 가지는 데이터이다.

공격모델은 4개의 은닉층과 Dropout 층으로 이루어진 MLP 모델이다. 공격 모델을 100 에포크까지 훈련했을 시의 Loss와 Accuracy 추이는 그림 3의 좌측 및 중앙의 그래프와 같다. Loss가 줄어들도록 안정적으로 공격모델을 훈련시킬 수 있었으며, 100 에포크에서의 정확도는 약 0.9155이다. 그림 3의 우측의 그래프에 해당하는 테스트에서의 Precision-recall graph와 ROC 커브를 확인했을 시 0 클래스의 AUC는 1.00, 1 클래스의 AUC는 0.97, 2 클래스의 AUC는 0.98의 성능을 보인다. 테스트에서의 정밀도, 재현율 및 F1 score는 표 3과 같다. 실험 결과로 미루어 봤을 때, 공격자가 종류를 추론하고자 하는 타겟 모델에서 confidence score vector가 아닌 분류 확률별 클래스의 순위만 알 수 있는 블랙박스 상황이여도 쿼리를 통해 얻은 데이터로 타겟 모델의 종류를 추론할 가능성이 있다. 기존의 모델 종류 분류 공격과 달리 후보 모델군과 공격 모델의 훈련에 사용되는 쿼리를 다른 데이터셋을 사용하고 그중에서도 한 클래스의 데이터로만 쿼리를 생성했을 시에도 높은 성능의 공격 모델을 훈련시켰다는 점에서 더 현실적인 가정을 하고 제약이 있는 상황에서도 이러한 위험성이 여전히 존재한다는 것을 알 수 있다.

Table 2. Test Performance of the Attack Model

metric / class	precision	recall	f1 score
0	1.00	0.95	0.97
1	0.81	0.99	0.89
2	0.97	0.80	0.88

V. 결 론

전통적인 NIDS는 정의된 규칙 혹은 피쳐 기반으로 공격을 탐지하기 때문에 이미 알려진 형태의 공격 유형에는 효과적으로 대응할 수 있으나 새로운 형태의 공격에 취약하기 때문에 머신러닝 및 딥러닝과 같

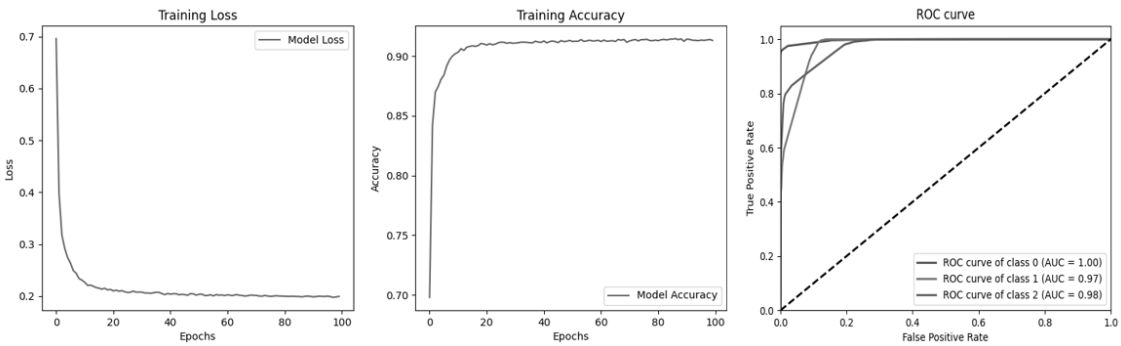

Fig. 3. Training Performance and ROC Curves in Attack Models

은 고급 인공지능 기술을 활용해 효과적으로 다양하고 복잡한 패턴의 공격을 탐지 및 분류할 수 있도록 하는 추세이다. 이는 전통적인 NIDS의 한계를 극복할 수 있게 하지만, AI 모델이 가지는 보안 취약점인 정보 추출 공격 등의 적대적 공격에 노출될 수 있다는 위험성을 가지게 된다.

AI에 대한 정보 추출 공격 중 하나는 모델 종류 추론 공격이다. 기존의 모델 종류 추론 공격은 이미지 분류 모델에 대해 실험했으며, 타겟 모델의 종류를 분류해 내는 공격 모델을 훈련시키기 위해서 타겟 모델이 될 수 있을 만한 여러 종류의 구조를 가지는 후보 모델군을 훈련하고 모든 클래스의 데이터를 쿼리해 응답을 얻는다. 이 때 쿼리하는 데이터는 공격 모델을 훈련한 데이터와 동일한 분포를 가져야 한다는 전제조건이 있다.

그러나 NIDS 모델에 적용할 때에는 NIDS 모델을 훈련할 때 사용되는 네트워크 데이터는 정상 행동 데이터 외에도 여러 가지 공격 클래스로 구성되어 있기 때문에, 모든 클래스의 데이터를 사용해서 NIDS 모델에 지속적으로 입력할 시 시스템 설정에 따라 해당 사용자를 차단하는 등 비정상 행위로 탐지되어 타겟 NIDS 모델의 반응을 지속적으로 관찰하기 어려워진다. 본 논문은 이러한 모델의 특징에 걸맞게 정상 네트워크 데이터만을 이용해 타겟 모델에 쿼리하는 상황을 가정하고 실험을 설계하였으며, 기존의 연구와 달리 후보 모델군을 훈련한 데이터와 공격 모델을 훈련하기 위한 쿼리 데이터 셋이 동일하지 않더라도 높은 성능을 보이는 공격 모델을 훈련시킬 수 있었다. 따라서 타겟 모델에 훈련되지 않은 데이터를 활용해 쿼리를 하더라도 모델의 종류를 분류할 수 있는 공격 모델을 훈련할 수 있는 현실적인 공격의 가능성을 제시한다. 본 논문이 제안하는 프레임워크는 공격자가 설정한 후보 모델군 안에 타겟 모델과 일치하는 구조를 가진 모델이 포함되어야 한다는 점과 공격자가 선정한 후보 모델들이 모두 높은 성능을 가지도록 훈련해 확보할 수 있는 상황이여야 한다는 제약이 있기 때문에 일정 수준의 한계를 가진다. 따라서 더욱 제한된 공격자의 권한으로 NIDS 모델의 구조 정보를 효율적으로 추출하는 후속 연구도 필요하다.

공격자가 NIDS 모델의 종류를 추론하게 되면 해당 모델을 복제하기도 용이하고, 더욱 최적화된 효과적인 적대적 공격을 통해 모델의 기능을 손상시켜 공격 패킷이 입력되어도 탐지를 하지 못하게 하거나 정상 패킷이 입력되어도 오탐하게 유도할 수 있다. 본 논문에서 제안한 대로 NIDS 모델을 추론할 때 정상 네트워크 패킷 데이터만을 이용해서 쿼리하면 타겟 모델의 반응을 지속적으로 관찰할 수 있으며 탐지도 피할 수 있기 때문에 NIDS에 대한 더욱 위협적인 보안 취약점이 될 수 있다. 따라서 이에 대한 방어 기술 개발이 필요하다.

References

[1] M.K. Asif et al, "Network Intrusion Detection and its strategic importance," IEEE Business Engineering and Industrial Applications Colloquium (BEIAC), Langkawi, Malaysia, pp. 140-144, Jul. 2013

[2] D. Joo, T. Hong, I. Han, "The neural network models for IDS based on the asymmetric costs of false negative errors and false positive errors," Expert Systems with Applications, vol. 25, no. 1, pp.69-75, Jul. 2003

[3] B. Subba, S. Biswas, S. Karmakar, "A Neural Network based system for Intrusion Detection and attack classification," Twenty Second National Conference on Communication (NCC), Guwahati, India, pp. 1-6, Mar. 2016

[4] M.S. Habeeb, T.R. Babu , " Network intrusion detection system: a survey on artificial intelligence based techniques," Expert Systems, vol. 39, no. 9, Jul. 2022

[5] K. He, D.D. Kim, M.R. Asghar, "Adversarial Machine Learning for Network Intrusion Detection Systems: A Comprehensive Survey," in IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 538-566, Jan. 2023

[6] K. Yang, et al, "Adversarial Examples Against the Deep Learning Based Network Intrusion Detection Systems," MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM),

Los Angeles, CA, USA, pp. 559-564, Oct. 2018

[7] I.J. Goodfellow, J. Shlens, C. Szegedy, "Explaining and Harnessing Adversarial Examples," ArXiv, 2015, Available at: https://arxiv.org/abs/1412.6572

[8] Z. Tian et al, "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning." ACM Computing Surveys, vol. 55, no. 8,Dec.2022

[9] S. Alahmed et al, "Impacting Robustness in Deep Learning-Based NIDS through Poisoning Attacks," MDPI Algorithms 2024, vol. 17, no. 4, pp.155, Apr.2024

[10] X. Zhang, C. Fang, J. Shi, "Thief, Beware of What Get You There: Towards Understanding Model Extraction Attack," arXiv, Apr. 2021, Available at : https://arxiv.org/abs/2104.05921

[11] R. Shokri et al,"Membership Inference Attacks Against Machine Learning Models," 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, pp. 3-18, May 2017

[12] H. Qiu et al, "Adversarial Attacks Against Network Intrusion Detection in IoT Systems," in IEEE Internet of Things Journal, vol. 8, no. 13,pp. 10327-10335, Jul.2021

[13] C. Zhang, X. Costa-Pérez and P. Patras, "Adversarial Attacks Against Deep Learning-Based Network Intrusion Detection Systems and Defense Mechanisms," in IEEE/ACM Transactions on Networking, vol. 30, no. 3, pp. 1294-1311, June 2022,

[14] Y. An, D. Choi, "Model Type Inference Attack Using Output of Black-Box AI Model," Journal of the Korea Institute of Information Security & Cryptology, 10(5), pp.817-826, Oct. 2022

[15] K. Roshan, A. Zafar, S.B.U. Haque, "Untargeted white-box adversarial attack with heuristic defence methods in real-time deep learning based network intrusion detection system," Computer Communications, vol.218, pp. 97-113, Mar. 2024

[16] D. Oliynyk, R. Mayer, A. Rauber, "I Know What You Trained Last Summer: A Survey on Stealing Machine Learning Models and Defences," ACM Computing Surveys, vol.55, no. 14, Dec. 2023

[17] A. Ilyas et al, "Black-box Adversarial Attacks with Limited Queries and Information," Proceedings of Machine Learning Research(PMLR), Stockholm, Sweden, vol. 80, pp. 2137-2146, Jul. 2018

[18] R. Vinayakumar et al,"Deep Learning Approach for Intelligent Intrusion Detection System," in IEEE Access, vol. 7, pp. 41525-41550, Apr. 2019

[19] M. Liu et al, "Host-Based Intrusion Detection System with System Calls: Review and Future Trends," ACM Computing Surveys, vol. 51, no. 5,pp. 1-36, Nov. 2018

[20] G.Kondrak, "N-Gram Similarity and Distance," String Processing and Information Retrieval. SPIRE 2005. Lecture Notes in Computer Science, vol. 3772. Springer, Berlin, Heidelberg, 2005

[21] F. Zenke, B. Poole, S. Ganguli, "Continual Learning Through Synaptic Intelligence," Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp.3987-3995, Aug. 2017

[22] M. Tavallaee et al, "A detailed analysis of the KDD CUP 99 data set," 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON,

Canada, pp.1-6, Dec. 2009
〔23〕 L. Dhanabal, S.P. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," International Journal of Advanced Research in Computer and Communication Engineering, vol.4, no. 6, Jun. 2015
〔24〕 M. Gao et al, "Malicious Network Traffic Detection Based on Deep Neural Networks and Association Analysis," MDPI Sensors, vol. 20, no. 5, March. 2020
〔25〕 A. Rosay et al, "Network Intrusion Detection: A Comprehensive Analysis of CIC-IDS2017," 8th International Conference on Information Systems Security and Privacy,Online Streaming, France. pp.25-36, Feb.2022
〔26〕 B.E. Zolbayar et al, "Generating Practical Adversarial Network Traffic Flows Using NIDSGAN," arXiv e-prints, Mar. 2022, Available at: https://arxiv.org/abs/2203.06694
〔27〕 J. Mirkovic, P. Reiher, "A taxonomy of DDoS attack and DDoS defense mechanisms," ACM SIGCOMM Computer Communication Review, vol. 34, no. 2,pp 39 - 53, Apr. 2004
〔28〕 Z. Durumeric et al, "The Matter of Heartbleed," IMC '14: Proceedings of the 2014 Conference on Internet Measurement, pp.475-488, Nov. 2014

〈저자소개〉

안 윤 수 (Yoonsoo An) 정회원
2019년 2월: 공주대학교 응용수학과 학사
2023년 8월: 숭실대학교 소프트웨어학과 석사
2023년 8월~현재: 숭실대학교 AI 안전성 연구센터 연구원
〈관심분야〉 AI 보안, 금융 데이터, 강화학습, 로보틱스

김 도 완 (Dowan Kim) 정회원
2019년 2월: 공주대학교 의료정보학과 학사
2022년 2월: 숭실대학교 소프트웨어학과 석사
2022년 2월~현재: 숭실대학교 AI 안전성 연구센터 연구원
〈관심분야〉 AI 보안, 머신러닝, 통계 분석

최 대 선 (Dae-seon Choi) 종신회원
1995년 2월: 동국대학교 컴퓨터공학과 학사
1997년 2월: 포항공과대학교 컴퓨터공학과 석사
2009년 1월: 한국과학기술원 전산학과 박사
1997년 1월~1999년 6월: 현대정보기술 선임
1999년 7월~2015년 8월: 한국전자통신연구원 인증기술연구실실장/책임연구원
2015년 9월~2020년 8월: 공주대학교 의료정보학과 부교수
2020년 9월~현재: 숭실대학교 소프트웨어학부 교수
2016년~현재: 정보보호학회 이사
〈관심분야〉 인증, 개인정보보호, AI 보안