

A Study on the Robustness of Insider Threat Detection via SHAP-Guided Adversarial Attacks and Retraining

Yu-Jin An*, Hyun-Dong Kim*, Sang-Hoon Han**, Geon-Yun Shin**,***

*Student, Dept. of Information Security, Hankyong National University, Gyeonggi, Korea

**Professor, Dept. of Information Security, Hankyong National University, Gyeonggi, Korea

***Professor, Research Institute of AI Security Convergence, Hankyong National University, Gyeonggi, Korea

[Abstract]

This study investigates the effects of SHapley Additive exPlanations (SHAP)-based feature selection on evasion attacks and defense through adversarial retraining in an insider threat detection setting using the CERT r4.2 dataset. Global and local SHAP values from a CatBoost model are combined to generate adversarial samples under a constrained feature-modification budget, and attack generation methods and retraining strategies are compared. The results show that SHAP-based feature selection achieves higher attack success rates (ASRs) than non-SHAP approaches, while searching both perturbation direction and magnitude generates stronger evasion attacks than using a fixed perturbation magnitude. In addition, the robustness gained through adversarial retraining transfers to attack types not used during training. In particular, retraining with a single strong attack improves robustness while limiting degradation in clean detection performance and increases in the false-positive rate for benign samples. With iterative retraining, additional performance gains are limited after the first round.

▶ **Key words:** SHAP, Insider Threat Detection, Evasion Attack, Adversarial Retraining, Model Robustness

[요 약]

본 연구는 CERT r4.2 내부자 위협 탐지 환경에서 SHapley Additive exPlanations(SHAP) 기반 피쳐 선택이 회피공격과 적대적 재학습의 방어 효과에 미치는 영향을 분석한다. CatBoost 모델의 전역·지역 SHAP을 결합해 제한된 변경 예산 내에서 적대적 샘플을 생성하고, 공격 생성 방식과 재학습 전략을 비교하였다. 실험 결과, SHAP 기반 피쳐 선택은 비-SHAP 방식보다 높은 공격 성공률(ASR)을 보였고, 방향·크기 탐색은 고정 변화 폭보다 강한 회피공격을 생성하였다. 또한 적대적 재학습은 학습에 사용되지 않은 공격에도 전이 방어 효과를 보였으며, 강한 단일 공격 기반 재학습은 원본 탐지 성능 저하와 정상 오탐률 증가를 비교적 제한하면서 강건성을 향상시켰다. 반복 재학습에서는 첫 번째 재학습 이후 추가적인 성능 개선이 제한적으로 나타났다.

▶ **주제어:** SHAP, 내부자 위협 탐지, 회피 공격, 적대적 재학습, 모델 강건성

• First Author: Yu-Jin An, Corresponding Author: Geon-Yun Shin

*Yu-Jin An (anyujin@hknu.ac.kr), Dept. of Information Security, Hankyong National University

*Hyun-Dong Kim (kimhdong6828@hknu.ac.kr), Dept. of Information Security, Hankyong National University

**Sang-Hoon Han (hansh0903@hknu.ac.kr), Dept. of Information Security, Hankyong National University

,*Geon-Yun Shin (gunyoon.shin@hknu.ac.kr), Dept. of Information Security, Hankyong National University, Research Institute of AI Security Convergence, Hankyong National University

• Received: 2026. 07. 15, Revised: 2026. 08. 14, Accepted: 2026. 09. 12.

I. Introduction

조직 내부 구성원이 합법적 접근 권한을 악용하여 시스템을 손상시키거나 기밀 정보를 유출하는 내부자 위협(Insider Threat)은 현대 사이버 보안 환경에서 심각한 위협 중 하나로 부상하고 있다. Ponemon Institute의 2026년 보고서에 따르면, 내부자 보안 사고로 인한 조직의 연평균 피해 비용은 2025년 기준 약 1,950만 달러에 달하며, 이는 전년도 보고서에서 제시된 1,740만 달러보다 증가한 수치이다. 또한 해당 보고서는 354개 조직에서 발생한 총 7,490건의 내부자 사고를 분석하였으며, 조직당 평균 25건의 내부자 사고가 발생한 것으로 나타났다. 사고 수습에는 평균 67일이 소요되었으며, 이는 2024년 81일, 2023년 86일보다 감소한 수치이다[1]. 이러한 위협의 심각성에도 불구하고 내부자는 정상 사용자와 동일한 접근 권한과 행위 경로를 공유하기 때문에, 외부 침입자와 달리 명확한 공격 흔적을 남기지 않으며 악성 행위는 로그온, 장치 사용, 이메일, 파일 접근, 웹 접속 로그 속에 희소하게 포함된다. 이로 인해 정상과 악성 행위의 경계가 모호하고 클래스 불균형이 심한 환경에서 오답과 미탐 문제가 지속되고 있다[2].

이러한 문제를 해결하기 위해 머신러닝 기반 내부자 위협 탐지 연구가 활발히 진행되고 있다. 사용자 행동 분석(User and Entity Behavior Analytics, UEBA) 기반 이상 탐지, 앙상블 모델, 시계열 딥러닝 등 다양한 접근 방식이 제안되었으며, CMU CERT 데이터셋을 활용한 벤치마크 연구들이 탐지 성능을 꾸준히 향상시켜 왔다[2]. 특히 최근 연구들은 특정 위협 시나리오를 구분하거나 장기 행동 의존성을 반영하는 방식으로 기존 탐지 모델의 한계를 보완하고 있다.

그러나 기존 CERT 기반 내부자 위협 탐지 연구의 상당수는 주어진 벤치마크 데이터에서 정상·악성 행위 또는 위협 시나리오를 정확히 분류하는 데 초점을 두고 있으며, 탐지 모델이 실제 운용 과정에서 마주할 수 있는 적응적 회피공격 상황은 상대적으로 충분히 다루지 못하고 있다[2-3]. 즉, 정적 데이터셋에서 높은 탐지 성능을 보이는 모델이라 하더라도, 공격자가 모델의 취약한 입력 피처를 파악하고 일부 행위 특성만 제한적으로 조정할 경우, 동일한 수준의 탐지 성능이 유지된다고 단정하기 어렵다. 따라서 내부자 위협 탐지 모델은 단순한 분류 성능뿐 아니라, 제한된 피처 변경을 통해 탐지를 회피하려는 공격에 대한 강건성 관점에서도 평가될 필요가 있다.

한편, 설명 가능한 인공지능(XAI) 기법, 특히

SHAP(Shapley Additive exPlanations)은 Shapley value에 기반하여 개별 예측과 전체 모델 수준에서 피처 기여도를 정량화할 수 있어[4], 탐지 모델이 민감하게 반응하는 취약 피처 축을 식별하는 데 활용될 수 있다. 기존 연구에서는 SHAP이 주로 모델 해석과 설명 목적으로 활용되었으나, 최근에는 이를 적대적 공격 생성에 응용하는 연구들이 등장하고 있다[5]. 그러나 내부자 위협 정형 피처 환경에서 SHAP 기반 회피공격의 효과와 이에 대한 적대적 재학습 방어 전략을 체계적으로 분석한 연구는 아직 부족한 실정이다.

이에 본 연구는 CERT r4.2 데이터셋을 대상으로 세션 단위 정형 피처 기반 CatBoost 탐지 모델을 학습하고, 전역·지역 SHAP 중요도를 활용하여 취약 피처를 선별한 뒤 제한된 변경 예산 내에서 다양한 SHAP 기반 회피공격을 생성하여 비교 분석한다. 또한 생성된 적대적 샘플을 원본 학습 데이터에 주입하여 재학습함으로써 탐지 강건성 향상 효과를 실험적으로 검증한다. 특히 단일·혼합·단계형 재학습 전략을 비교하고, 학습에 사용하지 않은 공격 유형에 대한 전이 방어 효과를 분석함으로써 SHAP이 단순한 설명 도구를 넘어 내부자 위협 탐지 모델의 취약 축 식별, 회피공격 생성, 방어 학습에 활용될 수 있음을 보인다.

본 논문의 구성은 다음과 같다. 2장에서는 정형 데이터에 대한 적대적 공격, XAI 기반 회피공격, 적대적 재학습에 관한 관련 연구를 검토한다. 3장에서는 특성 구성, CatBoost 모델 학습, SHAP 기반 취약성 분석, SHAP 기반 회피공격 생성, 적대적 재학습으로 이어지는 제안 프레임워크를 설명한다. 4장에서는 제안 방법의 회피공격 성능과 적대적 재학습을 통한 강건성 향상 효과를 실험적으로 분석한다. 마지막으로 5장에서는 연구 결과를 요약하고 한계점 및 향후 연구 방향을 제시한다.

II. Preliminaries

1. Related works

1.1 Adversarial Attacks on Tabular Data with Feature Constraints

정형 데이터에 대한 적대적 공격은 이미지나 텍스트와 달리 카테고리형 특성, 불변 특성, 특성 간 연동 관계와 같은 도메인 고유의 제약을 만족해야 한다는 점에서 차별화된다. 정형 데이터에서는 전문가의 시점에 인지되지 않는(imperceptible) 적대적 예제를 생성하기 위해 중요도가 낮은 특성에는 큰 교란을, 중요한 특성에는 작은 교란을

부여하는 방식이 제안되어 금융 분야에서의 회피 가능성이 확인되었다[6]. 이후 단순한 인지 불가능성을 넘어 공격에 소요되는 비용과 공격자가 얻는 효용을 함께 고려한 위협 모델이 제안되어 보다 현실적인 공격 조건이 정식화되었다[7]. 또한 기존 그래디언트 기반 공격이 특성 제약을 위반하는 문제를 해결하기 위해, 그래디언트 기반 공격인 CAPGD와 탐색 기반 공격인 MOEVA를 결합한 제약 적응형 공격(CAA)이 제안되어 신용평가·피싱 탐지·봇넷 탐지 등 다양한 사용 사례에서 특성 제약을 만족하면서도 높은 성능 저하를 유도할 수 있음이 확인되었다[8]. 본 연구는 동일하게 정형 데이터의 현실성 제약을 고려하지만, 트리 기반 CatBoost 모델에 적합한 SHAP 기반 특성 선택과 방향·크기 탐색을 결합한다.

1.2 XAI-Guided Adversarial Evasion

설명 가능한 인공지능(XAI)은 본래 모델의 예측 근거를 해석하기 위한 도구이나, SHAP과 같은 피쳐 기여도 기반 설명 방법은 모델이 민감하게 반응하는 특성을 식별할 수 있다는 점에서 적대적 공격 생성에도 활용될 수 있다[4]. Windows PE 악성코드 탐지 모델인 MalConv에 SHAP DeepExplainer를 적용하여 파일의 바이트 및 영역별 기여도를 분석하고, PE 구조의 수정 가능성과 파일 무결성을 고려하면서 SHAP 값이 큰 구간에 교란을 삽입하는 회피공격이 제안되었다[5]. 무작위 위치 기반 교란과 달리 이러한 SHAP 기반 선택은 모델이 민감하게 반응하는 영역을 우선 공략하여 적은 수정만으로 높은 회피 성공률을 달성하였다[5]. 유사하게 SHAP을 활용하여 IoT 봇넷 탐지 모델에서 예측에 가장 크게 기여하는 특성을 식별하고 해당 특성 값을 조정함으로써, 블랙박스 환경에서도 단일 특성 변경만으로 탐지를 회피할 수 있음이 확인되었다[9]. 이러한 연구들은 XAI가 단순한 모델 해석 도구를 넘어 적대적 공격 생성에도 효과적으로 활용될 수 있음을 실증한다. 본 연구는 이러한 XAI 기반 회피공격 개념을 확장하여 전역 SHAP과 지역 SHAP을 결합해 취약 특성을 선정하고, 이를 회피공격 생성의 기준으로 활용한다.

1.3 Adversarial Retraining for Intrusion Detection

적대적 재학습(adversarial training)은 생성된 적대적 예제를 학습 데이터에 포함해 모델을 다시 학습함으로써 회피공격에 대한 강건성을 높이는 대표적 방어 기법이다. FGSM, BIM, PGD와 같은 그래디언트 기반 공격을 침입탐지시스템(IDS)에 적용하여 교란 강도에 따른 탐지 성능 저하를 분석하고, 적대적 예제를 학습 데이터에 포함해 재학

습할 경우 공격 상황에서의 강건성이 유의미하게 향상됨이 확인되었다[10]. 다만 단순히 생성된 교란을 그대로 학습에 사용할 경우 비현실적인 예제가 포함될 수 있어, 클래스별 특성에 맞게 교란 패턴을 적응적으로 조정하는 A2PM 기법이 제안되어 현실성을 만족하는 적대적 예제 기반 재학습이 탐지 강건성을 높임이 확인되었다[11]. 또한 봇넷 탐지 환경에서 현실적 제약을 반영한 적대적 봇넷 공격(Adv-Bot)을 생성하고 이를 활용한 재학습의 효과가 검증되었다[12]. 이를 바탕으로 본 연구는 기존 연구들이 주로 딥러닝 기반 모델을 대상으로 한 것과 달리 트리 기반 CatBoost 모델을 대상으로 하며, 단일·혼합·단계형 재학습 전략과 공격 유형 간 전이 방어 효과를 비교한다.

III. Proposed Framework

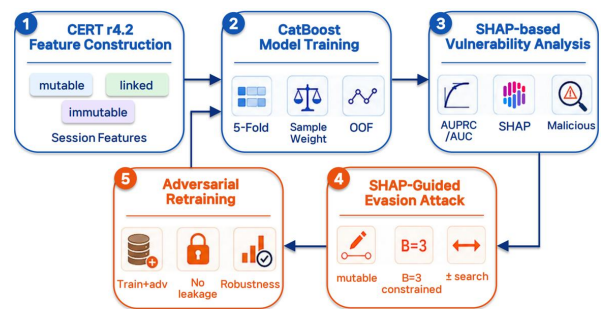


Fig. 1. Proposed Framework

본 논문에서는 다섯 단계로 구성된 프레임워크를 제안한다: (1) CERT r4.2 Feature Construction, (2) CatBoost Model Training, (3) SHAP-based Vulnerability Analysis, (4) SHAP-Guided Evasion Attack, (5) Adversarial Retraining.

먼저, (1)에서는 CERT 원시 로그를 탐지 모델이 학습할 수 있는 세션 단위 피쳐로 변환한다. (2)에서는 변환된 피쳐를 기반으로 CatBoost 단일 모델을 학습한다. (3)에서는 학습된 모델의 예측 결과에 대한 SHAP 분석을 통해 탐지 판단에 크게 기여하는 주요 피쳐와 취약한 행동 축을 식별한다. (4)에서는 식별된 취약 피쳐를 활용하여 제한된 피쳐 변경 예산 내에서 회피 적대적 샘플을 생성한다. 마지막으로, (5)에서는 생성된 적대적 샘플을 재학습에 활용하여 회피공격에 대한 탐지 강건성을 향상시킨다.

제안 프레임워크의 핵심은 탐지 모델의 학습과 성능 평가에 머무르지 않고, 모델의 판단 근거를 회피공격 생성과 방어 학습 과정에 순환적으로 연결하는 데 있다. 이 과정

에서 SHAP 분석은 단순한 사후 해석 도구가 아닌, 모델의 예측에 영향을 미치는 행동 축을 정량화하고 이를 공격 표적 선정과 방어 학습으로 연결하는 매개로 작용한다.

본 연구는 실제 공격자의 제한된 지식 수준을 직접 모사하기보다 탐지 모델이 가질 수 있는 취약성을 강한 공격 조건에서 평가하기 위한 white-box 위협 모델을 가정한다. 공격자는 학습된 CatBoost 탐지 모델과 입력 피처의 정의 및 전처리 방식에 접근할 수 있으며, 전역 SHAP 중요도와 개별 세션의 지역 SHAP 값에 접근하고 후보 샘플의 악성 예측 점수를 반복적으로 평가할 수 있다고 가정한다. 반면 탐지 모델 자체나 학습 데이터를 변경할 수 없으며, 현재 세션의 모든 피처를 자유롭게 조작할 수 없다. 공격은 3.1절에서 정의하는 수정 가능 피처 집합으로 제한되며, 각 피처의 허용 범위와 변경 예산을 만족해야 한다. 따라서 본 위협 모델은 공격자에게 충분한 모델 지식이 주어진 강한 조건에서 제한된 피처 변경만으로 탐지 모델의 취약한 결정 경계를 식별하고 그 과정에서 확인된 취약성을 적대적 재학습에 반영하여 궁극적으로 회피공격에 대한 탐지 모델의 강건성을 향상하기 위한 설정이다.

3.1 CERT r4.2 Feature Construction

CERT r4.2 원시 데이터는 이벤트 단위로 기록되므로, 본 연구에서는 이를 세션 단위의 정형 피처로 변환한다. 세션은 사용자의 logon 이벤트를 기준으로 구성하고, device, email, file, http 행위를 해당 세션에 집계한다. 세션별 피처는 로그 유형별 행위량, 사용 비율, 시간대 정보, 이벤트 간 전이 특성, 이전 행위 대비 변화량, 요일과 주말 여부와 같은 메타 정보를 포함하며, 총 83개의 정형 피처로 구성하였다.

CERT 기반 내부자 위협 탐지에서는 user-day와 user-session 등 서로 다른 시간 단위의 행위 집계 방식이 사용되어왔다. User-day는 일정 기간의 사용자 행위를 통합적으로 반영할 수 있는 반면, 악성 행위가 특정 세션에서만 발생하고 동일 날짜의 다른 세션은 정상일 수 있어 세션 단위 분석에 비해 세밀한 행위 구분에는 한계가 있다 [13]. 한편 CERT r4.2를 활용한 기존 연구에서는 하루 중 하나 이상의 악성 행위가 발생한 경우, 해당 user-day를 악성으로 정의하는 방식이 사용되었다[14].

본 연구의 CERT 정답에는 전처리 과정에서 생성한 세션 ID와 직접 대응되는 정보가 없으며, 원시 로그를 기반으로 구성한 세션 경계 또한 모든 경우에 완전한 logon-logoff 구간을 보장하지 않는다. 따라서 악성 이벤트의 발생 시각을 특정 세션에 직접 대응할 경우, 라벨이 누락되거나 다른

세션에 잘못 귀속될 가능성이 있다. 이에 본 연구에서는 하나 이상의 악성 행위가 존재하는 user-day를 악성으로 정의하고[14], 해당 사용자와 날짜에 속하는 세션에 동일한 라벨을 부여하였다. 따라서 본 연구의 세션 라벨은 개별 악성 이벤트를 정확히 나타내는 event-level ground truth 라기보다, user-day 수준의 정답을 세션 단위에 적용한 coarse-grained label로 해석한다.

회피공격의 현실성을 확보하기 위해 피처는 수정 가능 피처 43개, 수정 불가 피처 14개, 연동 제약 피처 26개로 구분된다. 수정 가능 피처는 공격 과정에서 독립적인 후보로 선택할 수 있는 행동 피처이며, 특정 세션에서 발생한 파일 접근 횟수, HTTP 요청 횟수, 이메일 전송 횟수와 같은 변수가 이에 해당한다. 이는 정상 데이터에서 산출한 허용 범위 안에서만 값을 변경한다. 수정 불가 피처는 요일, 주말 여부, 시간 변화량처럼 임의의 조작이 어렵거나 의미가 훼손되는 값이다. 또한 개수형 피처와 비율형 피처처럼 서로 함께 변해야 하는 연동 피처는 하나의 값만 독립적으로 수정될 경우, 실제 로그 구조와 맞지 않는 비현실적 샘플을 생성할 수 있으므로 단일 피처 공격 대상에서 제외한다. 이러한 분류는 이후 SHAP 기반 회피공격에서 공격 후보를 제한하는 제약 조건으로 사용되며, 본 연구에서는 수정 가능 피처 집합을 공격 대상 피처로 한정한다.

3.2 CatBoost Model Training

탐지 모델로는 정형 데이터 처리에 적합한 CatBoostClassifier를 사용한다. 내부자 위협 데이터는 악성 세션의 비율이 매우 낮기에, 단순 정확도 중심 학습은 정상 클래스에 편향될 가능성이 있다. 이를 완화하기 위해 악성 클래스에 더 높은 샘플 가중치를 부여하여 희소한 악성 행위 패턴이 학습 과정에서 충분히 반영되도록 한다.

또한 동일 사용자의 세션이 학습 데이터와 검증 데이터에 동시에 포함될 경우, 정보 누수가 발생할 수 있으므로 사용자 단위 5-fold GroupKFold 교차검증을 적용한다. 각 검증 fold는 서로 중복되지 않는 200명의 사용자로 구성되며, fold별 전체 세션 수는 각각 94,119, 94,122, 94,115, 94,114, 94,121개이다. 악성 라벨 세션을 하나 이상 포함한 사용자 수는 각 fold에서 17, 15, 17, 12, 9명이며, 악성 라벨 세션 수는 각각 338, 322, 278, 348, 147개이다. 사용자 단위의 독립성을 유지함에 따라 악성 사용자와 악성 세션의 분포에는 fold 간 차이가 존재하지만, 동일 사용자의 반복 패턴에 의한 정보 누수를 방지하고 독립적인 사용자에 대한 일반화 성능을 평가하기 위해 사용자 단위 분할을 유지하였다.

3.3 SHAP-based Vulnerability Analysis

본 단계에서는 학습된 모델이 어떤 피처에 의존하여 악성 세션을 탐지하는지 분석한다. SHAP 값은 각 피처가 모델의 예측에 기여한 정도를 나타내므로, 값이 큰 피처는 모델의 판단에 큰 영향을 미치는 피처로 해석할 수 있다.

본 연구에서는 전역 SHAP과 지역 SHAP을 함께 사용한다. 전역 SHAP은 전체 데이터에서 모델이 반복적으로 중요하게 사용하는 피처를 보여주고, 지역 SHAP은 개별 악성 세션에서 탐지 판단에 직접 기여한 피처를 보여준다. 두 정보를 결합하면 전체적으로 중요한 피처이면서 동시에 특정 샘플의 탐지 결과를 바꾸는 데 영향을 줄 수 있는 취약 피처를 찾을 수 있다.

$$S_{i,j} = 0.5|\phi_{i,j}| + 0.5 \cdot \frac{1}{1+r_j} \quad \text{식 (1)}$$

SHAP 기반 공격에서 제한된 변경 예산 내에서 우선 탐색할 피처를 선정하기 위해 지역 SHAP과 전역 SHAP 순위를 결합한 우선순위 점수는 식(1)과 같다. $|\phi_{i,j}|$ 는 현재 공격 대상 샘플 (i)에서 피처 (j)가 모델의 악성 예측에 기여한 정도를 나타내는 지역 SHAP 절댓값이다. r_j 는 전역 SHAP 중요도 기준 피처 (j)의 순위이며, 전역 중요도는 $1/(1+r_j)$ 의 reciprocal-rank 형태로 변환하여 전역적으로 반복해서 중요한 피처에 보조적인 우선순위를 부여하였다.

따라서 식(1)의 두 항은 서로 다른 척도를 가지며, 각각 정확히 50%의 정량적 기여도를 갖는다는 의미가 아니다. 지역 SHAP 절댓값을 통해 개별 샘플에 대한 직접적인 민감도를 우선 반영하면서, 전역 SHAP 순위를 보조 정보로 함께 반영하기 위한 구현상의 가중치이다. 본 연구에서는 이를 취약성의 정량적 척도로 해석하지 않고, 공격 후보 피처의 상대적 탐색 순서를 결정하는 용도로 사용한다.

3.4 SHAP-Guided Evasion Attack

회피공격의 목표는 기존 모델이 악성으로 탐지한 세션의 일부 피처를 제한적으로 변경하여 악성 예측 점수를 낮추는 것이다. 본 연구의 회피공격은 모델의 강건성을 평가하기 위한 제한적 위협 모델을 가정하며, 공격자는 모든 로그를 자유롭게 조작하는 것이 아니라 세션 내 일부 수정 가능 행동 피처만 제한적으로 변경할 수 있다고 본다. 이는 내부자가 실제 운영 환경에서 전체 행위 패턴을 급격히 바꾸기보다 탐지에 민감한 일부 행동량을 조정해 탐지를 회피하려는 상황을 모사하기 위한 설정이다. 이에 한 샘플에서 수정 가능한 피처 수는 변경 예산 (B)로 제한된다. 또한 각 피처 값은 정상 데이터 분포를 바탕으로 설정한 허용 범위 안에서만 변경된다.

$$\begin{aligned} \min_{x'} f(x') \\ \text{s.t. } \|x' - x\|_0 \leq B, \\ x'_j \in [l_j, u_j] \end{aligned} \quad \text{식 (2)}$$

이를 최적화 문제로 표현하면 식(2)와 같다. 이는 원본 샘플 x 로부터 공격 후 샘플 x' 를 생성할 때, 제한된 수의 피처만 변경하면서 모델의 악성 예측 점수 $f(x')$ 를 최소화하는 과정을 나타낸다. 여기서 (B)는 변경 가능한 피처 수를 의미하며, $[l_j, u_j]$ 는 피처 (j)의 허용 범위이다.

공격은 greedy 방식으로 수행된다. 먼저 SHAP 기반 우선순위가 높은 수정 가능 피처를 후보로 정렬한 뒤, 매 단계에서 모델의 악성 예측 점수를 가장 크게 낮추는 피처 변경을 선택한다. 특히 본 연구에서는 고정된 변화량만 적용하는 대신, 각 후보 피처에 대해 증가/감소 방향과 여러 변화 폭을 함께 탐색하여 악성 예측 점수를 가장 크게 낮추는 조합을 선택하는 방식인 선 탐색(line search)을 적용한다. 정상 데이터에서 산출한 피처의 허용 범위를 기준으로 10%, 20%, 30%, 50%에 해당하는 변화 폭을 적용하였으며, 이는 해당 피처의 정상 범위 크기에 비례하여 결정된다. 예를 들어 정상 범위의 10% 변화는 해당 피처의 하한과 상한 사이 간격의 10%만큼 값을 증가 또는 감소시키는 것을 의미한다.

이를 통해 단순히 중요한 피처를 선택하는 데 그치지 않고, 실제로 탐지 점수를 낮추는 방향과 크기를 함께 찾는다. 일반적으로 피처를 증가시키면 악성 기여가 클수록 모델 예측 점수가 오르고, 감소시키면 낮아지지만, 일부 피처는 정상 패턴과의 유사성에 따라 반대로 작용할 수도 있다. 따라서 피처 변화의 영향은 모델의 결정 경계와 데이터 분포에 따라 달라지므로, 단순한 증가/감소가 아니라 방향성과 변화량을 함께 고려한 탐색이 필요하다.

3.5 Adversarial Retraining

적대적 재학습의 목적은 특정 공격 샘플을 단순히 외우는 것이 아닌, 공격 과정에서 드러난 취약한 결정 경계를 보완하는 것이다. 이를 위해 원본 학습 데이터를 유지한 상태에서, 회피공격 절차를 적용한 악성 적대적 샘플을 원본 학습 데이터 대비 5% 규모로 추가하여 모델을 재학습하였다. 추가된 샘플은 원본 악성 라벨을 유지하며, 악성 클래스 평균 가중치를 부여하였다. 공격 샘플은 정상 데이터의 1-99 percentile에서 산출한 피처별 허용 범위와 피처 유형별 제약을 만족하도록 제한하였다. 또한 검증 fold에서 생성된 공격 샘플이 동일 fold의 재학습에 포함되지 않도록 하여 평가 누수를 방지한다.

본 연구의 목적은 적대적 샘플 주입 비율 자체를 최적화하는 것이 아니라, 동일한 적대적 학습 예산에서 재학습 전략 간 강건성을 비교하는 데 있다. 따라서 모든 비교 전략을 동일한 데이터 추가량에서 비교하기 위한 고정된 적대적 데이터 증강 예산으로 5%를 사용하여 주입량에 따른 영향을 통제하였다. 다양한 적대적 샘플 주입 비율에 따른 성능 변화는 향후 연구에서 추가로 분석할 필요가 있다.

IV. Experiments

1. Experimental Setup

본 연구에서는 Carnegie Mellon University Software Engineering Institute (CMU SEI)에서 공개한 CERT Insider Threat Dataset r4.2를 사용하였다. 전체 데이터는 470,591개의 세션으로 구성되며, 이 중 악성 세션은 1,433개로 전체의 0.305%에 불과하다. 따라서 본 실험 환경은 정상 클래스가 압도적으로 많은 극심한 클래스 불균형 문제를 포함한다.

내부자 위협 탐지에서는 악성 클래스 비율이 매우 낮아 AUROC만으로는 실제 탐지 성능을 충분히 설명하기 어렵다. 따라서 본 연구에서 공격 성능은 공격 성공률(Attack Success Rate, ASR)을 중심으로 평가하고, 재학습 이후의 방어 성능은 clean AUPRC, 강건 재현율(robust recall), 정상 오탐률(benign FPR)을 함께 사용하였다. ASR은 공격 전 악성으로 탐지되던 샘플이 공격 후 정상으로 오분류 되는 비율로 정의하였다.

실험은 Python 3.11 환경에서 수행하였으며, 주요 모델 학습과 적대적 재학습에는 NVIDIA GeForce RTX 5090 GPU를 활용하였다. 초기 CatBoost 모델의 탐지 성능을 기준으로 설정하였으며, 전체 fold의 out-of-fold (OOF) 검증 예측을 통합하여 산출한 pooled OOF AUROC와 AUPRC는 각각 0.9775와 0.5905로 나타났다. 또한 fold별 AUPRC의 산술평균은 0.5990으로 계산되었다.

회피공격은 검증 fold의 악성 세션 전체에 대해 적용되, 공격 성공률은 공격 전 $\tau=0.5$ 기준으로 악성으로 탐지된 786개 샘플을 기준으로 계산하였다. 여기서 τ 는 모델의 악성 예측 점수를 이진 분류 결과로 변환하기 위한 판정 임계값이다. 이진 분류에서는 일반적으로 0.5가 기본 판정 임계값으로 사용되며[15], 본 연구에서는 이를 운영 환경에서의 최적 임계값으로 가정하지 않고 공격 전후 및 재학습 전후의 변화를 동일한 기준에서 비교하기 위한 고정 평가 임계값으로 사용하였다. 이에 따라 악성 예측 점수가

0.5 이상인 경우, 해당 세션이 악성으로 탐지된 것으로 간주하였다. 본 연구에서 공격자가 수정 가능한 피쳐 수로 $B \in \{1, 2, 3, 5\}$ 를 비교하였으며, $B=3$ 이후 ASR 증가 폭이 둔화되고 실제 내부자 행위 조정 가능성을 고려할 때 과도한 피쳐 변경은 현실성이 낮다고 판단하여 $B=3$ 을 주요 비교 기준으로 설정하였다.

2. Evasion Attack Results

그림 2는 지역 SHAP 값과 전역 SHAP 순위를 결합하여 산출한 공격 후보 피쳐의 우선순위 점수를 나타낸다. 상위 피쳐에는 장치 사용, 이메일 전송, 파일 접근, HTTP 활동 및 시간대 관련 피쳐가 포함되었으며, 이는 제한한 후보 선정 기준이 다양한 로그 유형의 피쳐를 공격 후보로 포함하고 있음을 보여준다. 해당 시각화는 이후 회피공격 성능 비교에 앞서, SHAP 기반 공격에서 어떤 피쳐들이 우선 탐색 대상으로 설정되었는지 확인할 수 있다.

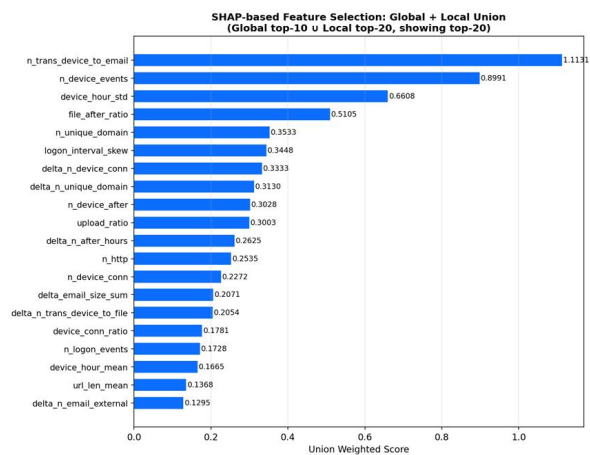


Fig. 2. Priority of candidate features for SHAP-guided evasion attacks

표 1은 회피공격 방식별 성능을 비교한 결과이다. 공격 방법은 피쳐 선택 방식과 변화 방향 탐색 방식에 따라 구분하였다. 비-SHAP 계열 공격은 SHAP 정보를 사용하지 않고 수정 가능 피쳐를 대상으로 무작위로 탐색하며, SHAP 기반 공격은 전역 SHAP과 지역 SHAP을 활용하여 모델의 악성 판단에 크게 기여한 피쳐를 우선하여 선택한다. 추가적인 비교를 위해 진화 탐색 기반의 MOEVA를 적용하였으며, 모든 공격 방식은 동일한 피쳐 수정 제약조건 하에서 실험되었다.

실험 결과, SHAP 기반 피쳐 선택을 사용한 공격이 비-SHAP 공격보다 더 높은 회피 성공률을 보였다. MOEVA는 pure_random보다 높은 ASR을 보였으나 SHAP 기반 공격들보다는 낮은 성능이 나타났다. 또한 방향 탐색에 있어, 선 탐색이 가장 높은 성능을 보였다. 이는 SHAP이 취

약 피처 식별에 유효하며, 고정된 변화 폭을 사용하는 방식보다 방향과 크기를 함께 탐색하는 방식이 회피공격 성능에 더 효과적임을 보여준다.

Table 1. Evasion Attack Performance

Category	Method	Feature Selection	Search Strategy	B=3 ASR
Non-SHAP attack	moeva	Random mutable features	Evolutionary search (crossover + mutation)	0.5636
	pure_random	Random mutable features	Random $\pm$ direction	0.4656
SHAP-based attack	shap_random	SHAP-ranked features	Random $\pm$ direction	0.7672
	shap_guided	SHAP-ranked features	SHAP-decreasing	0.5967
	shap_multistart	SHAP-ranked features	5-trial random $\pm$ search	0.8270
	shap_linesearch	SHAP-ranked features	$\pm$ direction with step-size search	0.8626

추가적으로, 가장 높은 공격 성능을 보인 선 탐색 방식에 대해 피처 선택 기준에 따른 성능 차이를 비교하였다. 지역 SHAP만을 사용한 경우의 ASR은 0.8410, 전역 SHAP만을 사용한 경우는 0.8613, 전역 및 지역 SHAP을 결합한 경우는 0.8626으로 나타났다. 전역·지역 SHAP 결합 방식이 가장 높은 ASR을 기록하였으나, 전역 SHAP 단독 방식과의 차이는 매우 작았다. 이는 지역 또는 전역 SHAP을 개별적으로 사용하더라도 효과적인 취약 피처 선정이 가능하며, 두 정보를 결합할 경우 가장 높은 공격 성능을 얻을 수 있으나 그 추가적인 성능 향상은 제한적일 수 있음을 보여준다.

3. Adversarial Retraining Results

표 2는 적대적 재학습 전략별 성능 비교 결과를 나타낸다. 재학습 전략은 단일 공격 기반 재학습, 혼합 공격 기반 재학습, 단계형 재학습으로 구분하였다. 단일 공격 기반 재학습은 하나의 강한 공격 소스를 사용하여 적대적 샘플을 구성하는 방식이고, 혼합 공격 기반 재학습은 여러 공격 유형에서 생성된 샘플을 함께 사용하는 방식이다. 단계형 재학습은 상대적으로 약한 공격에서 강한 공격으로 단계적으로 학습을 진행하는 전략이다. 또한 적대적 샘플 추가에 따른 재학습 효과를 동일한 학습 조건에서 비교하기 위해, 해당 재학습 실험에서 별도로 학습한 재학습 비교용 clean baseline을 기준으로 사용하였다.

혼합 공격 기반 재학습과 단계형 재학습 방식은 일부 방어 지표를 개선했으나, 동시에 원본 탐지 성능 저하나 정상 오탐률 증가가 동반되었다. 내부자 위협 탐지 시스템은 실제 운영 환경에서 오탐 비용이 크기 때문에, 단순 ASR만 낮추는 전략보다 원본 탐지 성능과 정상 오탐률을 함께 고려한

균형이 중요하다. 이런 관점에서 shap_linesearch B=3 기준 단일 재학습이 가장 안정적인 상충관계를 보였다.

Table 2. Performance Comparison of Adversarial Retraining Strategies

Strategy	Attack Source	Clean AUPRC	Δ AUPRC	def_ASR	bFPR
No retraining	None	0.5902	-	-	0.0013
Single attack retraining	shap_random	0.5624	-0.028	0.4429	0.0016
Single attack retraining	shap_linesearch	0.5702	-0.020	0.4341	0.0015
Mix-based retraining	shap + global linesearch mix	0.5563	-0.034	0.4298	0.0016
Curriculum retraining	shap_linesearch B=1 $\rightarrow$ 2 $\rightarrow$ 3	0.5507	-0.039	0.3973	0.0037

4. Cross-Attack Defense Generalization

표 3은 공격 유형 간 전이 방어 효과를 확인하기 위한 transfer matrix 결과이다. 이 실험의 목적은 특정 공격으로 재학습한 모델이 동일한 공격에만 강해지는지, 아니면 재학습에 사용하지 않은 다른 공격 유형에 대해서도 방어 효과를 보이는지 확인하는 것이다. 표의 Attack Source는 각 모델의 재학습에 사용된 공격 유형을 의미한다. 이후 각 재학습 모델에 여러 공격 유형을 다시 적용하여 ASR을 측정하였으며, Attack Source와 다른 공격 유형에 대한 ASR은 재학습에 사용되지 않은 공격에 대한 전이 방어 성능을 나타낸다. 공격 적용 전 기준 성능과의 비교에는 초기 학습 모델을 baseline으로 사용하였다.

기준 모델의 전체 ASR은 0.822였으나, 주요 재학습 모델에서는 약 0.43~0.44로 감소하였다. 이는 재학습 모델이 특정 공격 샘플에만 국한된 반응을 학습했다기보다, 공격 과정에서 노출된 취약한 결정 경계를 일부 보완했을 가능성을 시사한다. 특히 주목할 점은 SHAP 기반 공격으로 재학습한 모델이 비-SHAP 공격에 대해서도 방어 효과가 나타났다는 것이다. 이는 SHAP 기반 공격 샘플이 모델의 핵심 취약 축을 효과적으로 자극하여 해당 샘플을 이용한 재학습이 다른 유형의 피처 변경 공격에 대해서도 일반화된 방어 효과를 제공할 수 있는 것으로 해석될 수 있다.

Table 3. Transfer Matrix of Adversarial Retraining

Training Setting	Attack Source	Clean AUPRC	TPR@ FPR1%	Fixed-step SHAP	Search-based SHAP	non-SHAP	Overall
Baseline	None	0.599	0.6948	0.7964	0.8998	0.6661	0.8221
Adv. Retraining	shap_random	0.5676	0.6825	0.4342	0.4358	0.4459	0.4366
	Top2mix	0.5534	0.6779	0.4324	0.4333	0.4494	0.4353
	Top3mix	0.4911	0.6571	0.4238	0.4225	0.4474	0.4266
	Hardness-weighted	0.5529	0.6836	0.4299	0.4307	0.4491	0.4330
	shap_linesearch	0.5702	0.6842	0.4393	0.4373	0.4542	0.4406

5. Effect of Iterative Adversarial Retraining

표 4는 shap_linesearch B=3 공격을 반복적으로 생성하고 재학습하는 loop 실험 결과를 나타낸다. 이 실험은 각 재학습 단계의 변화를 비교하기 위해 초기 학습 모델을 기준으로 사용하여 1회 재학습 이후에도 새로운 취약점이 반복적으로 발견되는지, 추가 재학습이 실질적인 성능 개선으로 이어지는지 확인하기 위해 수행하였다.

1회 재학습 이후 강건 재현율이 크게 향상되었으나, 이후 라운드에서는 clean AUPRC와 강건 재현율이 일관되게 개선되지 않았다. 이는 반복적인 공격-재학습 과정에서 새로운 취약점이 일부 드러날 수는 있으나, 본 실험 환경에서는 첫 번째 재학습에서 주요 강건성 개선이 나타났으며 이후 반복에 따른 추가적인 성능 이득은 제한적이었음을 보여준다.

Table 4. Performance of the SHAP-LineSearch Retraining Loop

Round	Clean AUPRC	Robust recall
0	0.5990	0.1484
1	0.5696	0.5676
2	0.5665	0.5538
3	0.5705	0.5609
4	0.5705	0.5639
5	0.5563	0.5676

V. Conclusions

본 연구는 CERT r4.2 내부자 위협 데이터셋을 대상으로 SHAP 기반 회피공격 생성과 적대적 재학습을 결합한 탐지 강건화 프레임워크를 제안하였다. 제안 방법은 SHAP을 학습된 모델의 판단 근거를 설명하는 사후 해석 도구가 아닌, 탐지 모델이 어떤 행동 피처에 취약하게 의존하는지를 분석하고, 이를 회피공격의 표적 선정과 방어 학습을 위한 적대적 샘플 생성 과정으로 연결하는 점에서 의미가 있다.

실험 결과, SHAP 기반 피처 선택은 무작위 피처 선택보다 제한된 변경 예산 내에서 더 높은 회피 성능을 보였다. 특히 변화 방향과 크기를 함께 탐색하는 선 탐색 방식이 가장 강한 회피공격을 생성하였다. 이는 내부자 위협 탐지 모델이 일부 핵심 피처의 제한적 변화에도 탐지 결과가 크게 흔들릴 수 있음을 나타낸다. 또한 SHAP 기반 강한 단일 공격으로 생성한 샘플을 활용한 재학습은 원본 탐지 성능 저하와 정상 오탐률 증가를 비교적 제한하면서도, 학습에 사용하지 않은 다른 공격 유형에 대해서도 방어 효

과를 보였다. 이는 강한 단일 공격 기반 재학습이 모델의 핵심 취약 축을 효과적으로 자극하면서, 원본 탐지 성능 유지와 방어 효과 사이에서 안정적인 균형을 제공할 수 있음을 시사한다. 또한 반복 재학습에서는 첫 번째 재학습에서 대부분의 강건성 개선이 나타났으며, 이후 추가적인 성능 개선은 제한적이었다.

따라서 본 연구의 기여는 다음과 같다. 첫째, 극심한 클래스 불균형과 피처 변경 제약이 존재하는 내부자 위협 탐지 환경에서 SHAP 기반 회피공격의 유효성을 확인하였다. 둘째, SHAP 정보를 활용하여 모델의 취약 피처를 선별하고, 방향·크기 탐색을 결합한 회피공격 절차를 제시하였다. 셋째, 단일·혼합·단계형 재학습 전략과 반복 재학습의 라운드별 성능 변화를 비교함으로써, 방어 효과, clean 성능 유지, 정상 오탐 억제 간의 상충관계를 분석하였다. 넷째, 학습에 사용하지 않은 공격 유형에 대해서도 방어 효과가 나타남을 확인함으로써, SHAP 기반 적대적 샘플이 특정 공격 방식에 국한되지 않고 모델의 일반적인 취약성을 보완하는 학습 신호로 활용될 수 있음을 검토하였다.

다만 본 연구는 사용자별 직무, 근무 시간, 접근 권한, 장치 사용 습관과 같은 개인화된 행동 특성을 충분히 반영하지 못했다는 한계가 있다. 또한 서로 연동되는 피처를 개별적으로 조정하는 데에는 제약이 있으므로, 향후에는 사용자별 행동 이력과 접근 권한을 반영한 개인화 제약 및 연동 피처의 그룹 단위 변경을 적용하고, 세션 경계와 이벤트 발생 시각을 고려하여 악성 이벤트와 세션 간 대응 관계를 보다 정밀하게 구성할 필요가 있다. 또한 적대적 샘플 주입 비율에 따른 재학습 효과를 분석하고, 다른 내부자 위협 데이터셋 및 실제 운영 로그를 통해 제안 방법의 일반화 가능성과 계산 비용을 포함한 실제 적용 가능성을 검증하고자 한다.

ACKNOWLEDGEMENT

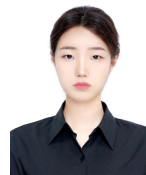
This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-25122978).

REFERENCES

- [1] Ponemon Institute, "2026 Cost of Insider Risks Global Report," DTEX Systems, URL: <https://ponemon.dtex.ai/>, 2026.

- [2] T. Tian, C. Zhang, B. Jiang, H. Feng, and Z. Lu, "Insider Threat Detection for Specific Threat Scenarios," *Cybersecurity*, Vol. 8, Article 17, March 2025. DOI: 10.1186/s42400-024-00321-w
- [3] A. Alotaibi and M. A. Rassam, "Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense," *Future Internet*, Vol. 15, No. 2, Article 62, Jan. 2023. DOI: 10.3390/fi15020062
- [4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, Vol. 30, pp. 4765-4774, Long Beach, CA, USA, Dec. 2017. DOI: 10.5555/3295222.3295230
- [5] K. Aryal, M. Gupta, M. Abdelsalam, and M. Saleh, "Explainability Guided Adversarial Evasion Attacks on Malware Detectors," *Proceedings of the 2024 IEEE International Conference on Computer Communications and Networks (ICCCN)*, pp. 1-9, Big Island, HI, USA, July 2024. DOI: 10.1109/ICCCN61486.2024.10637577
- [6] V. Ballet, X. Renard, J. Aigrain, T. Laugel, P. Frossard, and M. Detyniecki, "Imperceptible Adversarial Attacks on Tabular Data," *arXiv preprint arXiv:1911.03274*, Nov. 2019. DOI: 10.48550/arXiv.1911.03274
- [7] K. Kireev, B. Kulynych, and C. Troncoso, "Adversarial Robustness for Tabular Data through Cost and Utility Awareness," *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, USA, Feb. 2023. DOI: 10.14722/ndss.2023.23142
- [8] T. Simonetto, S. Ghamizi, and M. Cordy, "Constrained Adaptive Attack: Effective Adversarial Attack Against Deep Neural Networks for Tabular Data," *Advances in Neural Information Processing Systems*, Vol. 37, pp. 27817-27849, Vancouver, Canada, Dec. 2024. DOI: 10.52202/079017-0873
- [9] M. M. Alani, A. Mashatan, and A. Miri, "Adversarial Explainability: Utilizing Explainable Machine Learning in Bypassing IoT Botnet Detection Systems," *arXiv preprint arXiv:2310.00070*, Sept. 2023. DOI: 10.48550/arXiv.2310.00070
- [10] I. Debicha, T. Debatty, J.-M. Dricot, and W. Mees, "Adversarial Training for Deep Learning-based Intrusion Detection Systems," *The Sixteenth International Conference on Systems (ICONS 2021)*, pp. 45-49, Porto, Portugal, April 2021. DOI: 10.48550/arXiv.2104.09852
- [11] J. Vitorino, N. Oliveira, and I. Praça, "Adaptive Perturbation Patterns: Realistic Adversarial Learning for Robust Intrusion Detection," *Future Internet*, Vol. 14, No. 4, Article 108, March 2022. DOI: 10.3390/fi14040108
- [12] I. Debicha, B. Cochez, T. Kenaza, T. Debatty, J.-M. Dricot, and W. Mees, "Adv-Bot: Realistic Adversarial Botnet Attacks against Network Intrusion Detection Systems," *Computers & Security*, Vol. 129, Article 103176, June 2023. DOI: 10.1016/j.cose.2023.103176
- [13] D. C. Le, N. Zincir-Heywood, and M. I. Heywood, "Analyzing Data Granularity Levels for Insider Threat Detection Using Machine Learning," *IEEE Transactions on Network and Service Management*, Vol. 17, No. 1, pp. 30-44, March 2020. DOI: 10.1109/TNSM.2020.2967721
- [14] P. Manoharan, W. Hong, J. Yin, H. Wang, Y. Zhang, and W. Ye, "Optimising Insider Threat Prediction: Exploring BiLSTM Networks and Sequential Features," *Data Science and Engineering*, Vol. 9, No. 4, pp. 393-408, December 2024. DOI: 10.1007/s41019-024-00260-z
- [15] C. Elkan, "The Foundations of Cost-Sensitive Learning," *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 973-978, 2001.

Authors



Yu-Jin An is currently pursuing the B.S degree in Information Security at Hankyong National University, Gyeonggi, Korea. Her research interests include AI Security, Explainable AI (XAI) and Trustworthy AI.



Hyun-Dong Kim is currently pursuing the B.S degree in Information Security at Hankyong National University, Gyeonggi, Korea. His research interests include Malware Analysis, AI Security and Digital Forensics.



Sang-Hoon Han received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Dongguk University, Korea, in 1990, 1995 and 2002, respectively. Dr. Han joined the faculty of the Department of

Computer Information Security at Korea National University of Welfare, Pyeongtaek, Korea, in 2003. Since 2023, he has been a professor at the School of Computer Engineering & Applied Mathematics, Hankyong National University. He is interested in Information Security, Internet of Thing(IoT) and Computer Vision, and multimedia computing.



Geon-Yun Shin received MS and PhD degrees in computer engineering from Gachon University, Republic of Korea, in 2018 and 2023, respectively. He is a professor in the major of information security, Hankyong

National University, Korea. His research interests include AI security, LLM Security, unknown attack detection, dark web, open set recognition, and reinforcement learning.