JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 11, pp. 89-101, November 2023
https://doi.org/10.9708/jksci.2023.28.11.089

Techniques for Improving Host-based Anomaly Detection Performance using Attack Event Types and Occurrence Frequencies

Juyeon Lee*, Daeseon Choi**, Seung-Hyun Kim*

*Student, Dept. of Computer Education, Korea National University of Education, Cheongju, Korea
**Professor, Dept. of Software, Soongsil University, Seoul, Korea
*Assistant Professor, Dept. of Computer Education, Korea National University of Education, Cheongju, Korea

[Abstract]

In order to prevent damages caused by cyber-attacks on nations, businesses, and other entities, anomaly detection techniques for early detection of attackers have been consistently researched. Real-time reduction and false positive reduction are essential to promptly prevent external or internal intrusion attacks. In this study, we hypothesized that the type and frequency of attack events would influence the improvement of anomaly detection true positive rates and reduction of false positive rates. To validate this hypothesis, we utilized the 2015 login log dataset from the Los Alamos National Laboratory. Applying the preprocessed data to representative anomaly detection algorithms, we confirmed that using characteristics that simultaneously consider the type and frequency of attack events is highly effective in reducing false positives and execution time for anomaly detection.

▸ **Key words**: Anomaly Detection, LANL2015, HBOS, Feature extraction, Logon Type

[요 약]

사이버 공격으로 인한 국가, 기업 등의 피해를 막기 위해 공격자의 접근을 사전에 감지하는 이상 탐지 기술이 꾸준히 연구되어왔다. 외부 혹은 내부에서 침입하는 공격들을 즉각적으로 막기 위해 실행시간의 감축과 오탐지 감소는 필수불가결하다. 본 연구에서는 공격 이벤트의 유형과 빈도가 이상 탐지 정탐률 향상 및 오탐률 감소에 영향을 미칠 것으로 가설을 세우고, 검증을 위해 Los Alamos National Laboratory의 2015년 로그인 로그 데이터셋을 사용하였다. 전처리 된 데이터를 대표적인 이상행위 탐지 알고리즘에 적용한 결과, 공격 이벤트 유형과 빈도를 동시에 적용한 특성을 사용하는 것이 이상행위 탐지의 오탐률과 수행시간을 절감하는데 매우 효과적임을 확인하였다.

▸ **주제어**: 이상행위탐지, LANL2015, HBOS, 특성 추출, Logon Type

• First Author: Juyeon Lee, Corresponding Author: Seung-Hyun Kim
*Juyeon Lee (juyeon5741@knue.ac.kr), Dept. of Computer Education, Korea National University of Education
**Daeseon Choi (sunchoi@ssu.ac.kr), Dept. of Software, Soongsil University
*Seung-Hyun Kim (kimsh@knue.ac.kr), Dept. of Computer Education, Korea National University of Education
• Received: 2023. 09. 13, Revised: 2023. 11. 15, Accepted: 2023. 11. 15.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

모바일 정보 통신 기술의 발달과 네트워크 기술이 발달하면서 언제 어디서나 인터넷에 접속할 수 있는 '초연결시대'가 도래하였다[1]. 하지만, 기술의 발달은 인간의 편의뿐만 아니라 여러 가지 문제를 초래하였다. 특히 사이버 공간에서의 보안 취약점으로 인해 정보 유출은 빈번하다.

중소벤처기업부(2022)는 '중소기업 기술보호 수준 실태조사'에서 국내 중소기업은 기업당 평균 1.1건의 기술 또는 경영상의 정보 침해 피해를 본 것으로 보고했다[2]. 또한 기술 또는 경영상의 정보가 침해당했을 때, 중소기업들이 대부분 피해 사실을 발생 직후 인지하지 못하는 것으로 보고했다[2]. 지난 2023년 4월 10일 가상 화폐 거래소 '지닥(GDAC)'이 해킹 범죄로 가상화폐 보관자산 23%에 해당하는 금액에 피해를 보았다[3]. 이와 같은 악의적인 침입 시도가 전 세계적으로 2021년 대비 2022년에는 19% 증가하였다[4]. 악의적인 침입 시도가 증가하는 만큼 피해를 막기 위해 사전에 악의적인 침입을 침입 탐지 시스템(Intrusion Detection System)으로 탐지하여 차단한다. 하지만 침입 탐지 시스템에서는 정상적인 접근을 공격으로 잘못 판단하는 오경보가 자주 발생하는데, 모든 오경보 알람을 관리자가 확인하는 것은 불가능하다[5]. 이를 해결하기 위한 방안으로 정상 접근과 다른 빈도를 보이는 접근을 공격으로 탐지하는 이상 탐지 기법이 꾸준히 연구되고 있다.

본 연구는 새로운 이상 탐지 모델을 제안하기보다, 이상 탐지에 활용할 새로운 특징을 생성하는 방안을 제안한다. 데이터셋이 제공하는 기존 속성의 특징만으로는 공격 이벤트를 식별하기 어려운 상황에서, 정상행위에서의 빈도와 큰 차이를 보이는 이벤트를 이상 행위로 인식하기 위해 기존 속성을 활용한 새로운 특징을 생성한다. 이상 탐지 분야에서 사용되는 Los Alamos National Laboratory의 데이터셋[6]을 대상으로, 다양한 이상 탐지 알고리즘에 적용 가능한 새로운 특징과 방법을 제안하고 성능을 검증한다.

기존 연구 중에서, 임선영 외[7]의 연구는 대상 데이터셋의 선정과 연구 방법론 측면에서 본 연구와 가장 유사한 특성을 보인다. 따라서 위 연구를 비교 대상으로 선정하되, 본 연구는 기존 연구와는 다른 새로운 특징을 제시한다. 그리고 기존 연구에서 활용한 다양한 머신러닝 기법에 본 연구의 특징을 적용하여, 해당 연구 결과의 정확도와 오탐률을 개선하는 방향으로 본 연구의 유효성을 검증한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서 인공지능 알고리즘을 적용한 이상 탐지 분야의 선행 연구를 소개하고, 이상 탐지를 위해 사용한 데이터셋의 구성을 소개한다. Ⅲ장에서는 가설을 기반으로 새로운 특성을 제시하고 타당성 확인을 위한 실험 방법을 소개한다. Ⅳ장에서는 제시한 연구 방법의 결과를 분석하고, Ⅴ장에서는 제시하는 연구의 결론을 맺는다.

II. Preliminaries

2. Related works

이상 탐지는 국내외에서 꾸준하게 연구되는 주제이고, 크게 모델과 특징 추출 측면에서 연구되어 이상 탐지의 성능을 향상시키고 있다. 모델 측면의 연구는 이상 탐지 기법에서 사용되는 모델들의 개선과 적용을 연구하는 분야로, 딥러닝을 이용한 모델링이 큰 비중을 차지한다. 이러한 모델들은 정상적인 데이터와 비정상적인 데이터를 분류하는 데에 사용되며, 높은 정확도와 낮은 오진율을 달성하기 위한 방법들을 연구한다. 특징 추출 측면의 연구는 이상 탐지 기법에서 데이터의 특징을 추출하는 방법을 연구하는 분야로, 이상적인 특징을 추출하고 이를 기반으로 더욱 정확하고 효율적인 이상 탐지 모델의 구현하는 데에 사용된다.

2.1 Model

모델 측면의 연구는 규칙 기반, 머신러닝 기반, 딥러닝, 기반, 앙상블 기반으로 분류할 수 있다. 첫째, 규칙 기반 이상 탐지는 미리 정의된 규칙 집합에서 벗어난 테스트를 이상으로 인식하는 룰 기반 연구[8]로, Tandon은[9]은 기존 규칙 알고리즘을 수정하여 시스템 호출 시퀀스와 인수에 대한 규칙을 생성하여 이상 행위를 구분했다. Qing 외[10]는 정상적인 시스템 호출 시퀀스에서 최소화된 규칙 집합을 추출하는 집합 이론에 기반한 이상 감지 접근 방식을 제안했다.

둘째, 머신러닝 기반의 모델 연구는 병렬로 실행 가능하며 분산 기계 학습 알고리즘에 맞게 대규모 데이터 세트를 적절하게 분할한다. k-최근접 이웃 알고리즘(kNN)[11], k-평균 클러스터링 알고리즘(KMC)[12], 결정 트리[13], 서포트 벡터 머신(SVM)[14], HMM[15] 및 베이지안 네트워크[16]가 HIDS(Host-based Intrusion Detection System)에서 구현되었다. 윤지영 외[17]는 설명 가능한 로그 이상 탐지를 위해 폐쇄 순차패턴 마이닝, 베이지안 규칙 리스트을 사용했다.

셋째, 빅 데이터 내에서 근본적인 패턴을 발견하는 딥

러닝 기술은 빅데이터 환경에서 HIDS에 적용되고 있다[18]. CNN[19], GRU[20], LSTM[21]을 적용한 연구가 존재한다. 오상현 외[22]는 LSTM을 활용하여 적대적 생성 신경망의 구조를 구성하고, 새로운 이상 탐지 방법을 제시하여 오토인코더에 비해 정밀도와 정확도를 개선하였다.

마지막으로 넷째, 앙상블(Ensemble) 기법은 여러 개별 분류 모델을 조합하여 높은 예측 성능을 달성하는 머신러닝 기법으로, 침입탐지 시스템에서도 개별 모델의 한계점을 보완하고 높은 탐지율과 낮은 오진율을 동시에 달성한다. Kaiafas 외[23]는 세 가지 모델(Random Forest, LogitBoost, Logistic Regression)로 인증 로그들을 분류하고, 각 모델의 예측 결과를 활용하여 다수결 원칙(Majority Voting)으로 악의적인 인증 이벤트를 탐지하였다. Aghaei[24]는 앙상블 분류를 이용한 빈도 기반 오용 탐지 방식을 개발하였다. 분류 모델은 나이브 베이즈, SVM, PART, decision tree, random forest에서 예측을 결합하여 최종 예측을 하는 다수결 원칙(Majority Voting) 앙상블 기법을 기반으로 개발되었고, 제안된 오용 탐지 시스템은 공격 탐지에서 높은 성능을 보였다.

2.2 Feature engineering

전처리 등 특징 추출과 관련된 연구는 특정 속성을 선별하는 연구와, 기존 속성을 조합/변형하여 추가 속성을 생성하는 연구로 분류할 수 있다. 첫째, 이상행위를 효과적으로 식별하기 위해 특정 속성을 선별하는 연구로, Siadati 와 Memon[25]은 사용자, 출발지 컴퓨터, 도착지 컴퓨터의 속성으로 로그인 패턴을 추출하여 정상적인 로그인을 모델링하여 학습한 후, 정상 범위 밖의 이상 로그인을 식별하여 측면이동을 탐지하였다. Liu 외[26]는 원격 로그인 정보를 활용하여 이상 행위를 탐지하고, 오탐지를 줄이기 위해 로그온 이후 이루어지는 이벤트 정보로 이상 행위를 탐지하였다. Meijerink[27]는 이벤트 로그에서부터 특성들을 선정하고 HDBSCAN (Hierarchical Density-Based Spatial Clustering for Applications with Noise) 클러스터링 알고리즘을 적용하여 클러스터에서 멀리 떨어진 이상치(outlier)를 탐지하여 이상 행위 탐지를 시도하였다. Besharat 외[28]는 데이터셋을 로지스틱 회귀 분석한 뒤, 사용 중요도가 높은 특성을 선택하여 이상 탐지를 수행했다.

주성분분석을 통해 차원을 축소하여 특성을 선별하는 연구로, Meijerink[27]는 고차원의 데이터를 PCA 차원축소를 통해 변환한 후 표준편차의 3배의 차이가 나는 데이터를 이상치로 판별하여 악의적인 데이터로 분류하였다. Powell[29]은 악의적인 로그인 활동을 탐지하기 위해 로그인 활동 이력을 로그인 그래프로 나타내는데 그래프 토폴로지로 특성(feature)들을 식별하여 NMF (비음수 행렬 분해)와 PCA(주성분분석) 알고리즘을 활용해 데이터를 저차원으로 변환하고 이를 다시 재변환할 때의 오류값을 정량화하여 비정상 행위를 식별하기 위한 지표로 사용했다.

둘째, 추가 특징을 생성하는 연구로, N-gram을 이용하여 연속 시퀀스를 표현한 결과, 이것이 HIDS에서 효과적인 것으로 입증되었다[11]. Creech 외[30]는 전체 시스템 호출 추적을 스캔하기 위해 다중 길이 슬라이딩 윈도우를 사용했으며 다중 길이 N-gram을 생성하였다. Tan 과 Maxion[31]은 다양한 시퀀스 길이를 선택하는 이론적이고 실험적인 조사를 수행했다. Aghaei[24]는 N-gram을 이용해 전처리 후 특징들을 추출하여 패턴을 생성하였다. 특정 특징의 빈도를 활용한 연구로, Xie 와 Hu[32]는 호스트 기반 비정상행위 탐지 방식의 개발을 위해 데이터셋의 패턴과 빈도를 이용한 분석을 이용했다. 임선영 외[7]는 데이터셋에서 정상 이벤트, 공격 이벤트의 발생 빈도(frequency) 특성을 사용 이상 탐지를 실시하였다.

2.3 Other research

Maske 와 Parvat[33]은 시스템 호출 번호가 포함된 데이터 셋에서 사용자가 단어의 길이를 결정하는 단어 사전을 구성하고, 이 단어 사전을 통해 구문 사전을 생성하여 특성 벡터를 추출하고, 콜 트레이싱 추론 방법을 사용하는 의미론적 해석을 통해 이상 행위를 탐지하는 방법을 제안했다. Kaiafas 외[23]는 인증 이벤트로부터 기본적인 피처 외에 합성 피처들을 추출하고, 인증 로그를 분류하는 지도학습 방식을 제시하였다. 박경선과 김강석[34]은 NGIDS-DS(Next Generation IDS Dataset)을 슬라이딩 윈도우, 패딩을 적용하여 이상 탐지를 실시하였다. 유승태와 김강석[35]은 슬라이싱과 제로 패딩, Doc2Vec를 적용하여 전처리한 데이터로 이상 탐지를 실시하였다.

III. The Proposed Scheme

3.1 Hypothesis

이상 탐지 연구 분야에서 실험을 위해 공개적으로 사용할 수 있는 데이터셋은 다른 분야의 데이터셋에 비해 충분하지 않다[36]. IT 분야 데이터 소스가 사이버 보안을 위해 수집된 것이 아닌, 서비스 운영을 위한 모니터링을 목표로 만들어진 데이터들이 대부분이기 때문이다[37]. 또한 기존

의 침입탐지 데이터셋이 로우 레벨의 성능메트릭을 수집하는데 비해 이상 탐지 분야의 데이터셋은 하이 레벨의 보안 이벤트를 다룬다. 본 연구에서는 이상 탐지 연구에서 널리 사용되는 데이터셋(UNM, DARPA, Firefox-DS, ADFA-LD, NGIDS-DS)[36]의 특징을 고려하여 가설을 설정하였다. 5가지 데이터셋 중 4개에서 10개 미만의 공격 유형이 포함되어있다는 점에서 가설 1을 설정하였다.

그리고 이상 데이터의 특성상 정상 데이터의 범주에서 비교하였을 때, 비교적 그 수가 적고 매우 불균형하므로 특정 유형의 공격이 매우 적은 빈도로 발생할 것으로 예상하였고 실제로도 Firefox-DS, ADFA-LD, NGIDS-DS 데이터 셋에서 공격 비율이 3% 미만으로 확인되어 가설 2를 설정하였다.

가설 1. 공격은 특정 인증 타입을 악용한 방식일 것이다.

가설 2. 가설 1을 만족하는 공격 데이터는 발생 빈도(frequency)가 매우 낮을 것이다.

3.2 LANL2015 dataset

Table 1을 보면 이상 탐지를 위한 대표적인 데이터셋이 5가지 있지만, UNM, DARPA 데이터셋의 경우 1998년에 만들어져 현재의 공격을 표현하지 못한다는 한계가 있어 제외하였다. Firefox-DS의 경우 Firefox 환경에 최적화되어 제외하였고, ADFA-LD, NGIDS-DS는 현재 제공되지 않아 제외하였다. 대신, 본 연구에서는 최근 이상 행위 탐지 연구[7][43][44][45][46][47]에 사용된 Los Alamos National Laboratory의 데이터셋[6]을 사용하였다. 이 기관에서 제공하는 데이터셋 중에서 이벤트 타입이 다양하고, 공격 데이터가 제공되는 'Comprehensive, Multi-Source Cyber-Security Events'(이하 LANL2015)[6]를 사용하였다.

2014년에 만들어진 'User-Computer Authentication Associations in Time' 데이터셋은 'time, user, computer'의 구조를 가진다. 이벤트, 인증 로그와 같은 정보를 얻을 수 없으며, 공격데이터가 없어 이상 탐지 데이터셋으로 적절하지 않아 제외하였다.

2018년에 만들어진 'Unified Host and Network Data Set'은 공격데이터가 포함되어 있으나, 공격 데이터가 라벨링 되어있지 않아 이상 탐지의 성능을 평가할 수 없어 실험을 위한 데이터셋으로 적절하지 않다.

LANL2015 데이터셋은 5가지의 파일로 구성되어 있으며 공격 이벤트를 식별하기 위한 라벨링이 되어있지 않지만, 공격 로그를 별도 제공하기 때문에 정상 이벤트와 공격 이벤트를 구분할 수 있다. 따라서 이상 행위 기법의 성능 평가가 가능하다.

LANL2015 데이터셋의 정보는 Table 2와 같다. 12,425명의 사용자, 17,684개의 컴퓨터, 62,974개의 프로세스에 대해 58일간의 1,648,275,307개의 이벤트가 기록되어 있다.

Table 2. LANL2015 data information

Users	12,425
Computers	17,684
Processes	62,974
Events	1,648,275,307
Data Volume	12 Gigabytes

LANL2015 데이터셋은 총 5개의 파일(auth, proc, flows, dns, redteam)로 구성되며, 모두 txt파일을 gz형

Table 1. Comparison of HIDS datasets(Modified from Host-based intrusion detection system with system calls: Review and future trends[36].)

Dataset	Year	Number of attack type	Percentage of Attacks(%)	Disadvantages
UNM[38]	1998	7	-	System call arguments are eliminated
				The current range of diverse attack methods is not being accounted for
DARPA[39]	1998	32	-	A limited range of attacks with a small data scale
				System call traces are uncomplicated
Firefox-DS[40]	2013	5	2.71	Focuses exclusively on the Firefox web browser
				Referred to in only a handful of studies
ADFA-LD[41]	2013	6	1.15	Only system call numbers are included
				Attacks with a small data scale are included
NGIDS-DS[42]	2017	7	1.40	Referred to in only a handful of studies
				Attacks with a small data scale are included

식으로 압축된 형태로 제공된다. 본 실험에는 인증 로그(auth.txt)와 redteam의 공격 로그(redteam.txt) 두 개의 데이터를 사용하였다. 인증 로그는 정상 로그와 공격 로그로 구분되며, Windows 기반 환경에서 수집된 이벤트가 1,051,430,459개 존재한다. 이벤트는 “time, source user@domain, destination user@domain, source computer, destination computer, authentication type ,logon type, authentication orientation, success/failure”의 형식으로 구성된다. 공격 로그(redteam.txt)에는 총 749개의 이벤트가 기록되며 “time, user@domain, source computer, destination computer”의 형식으로 저장된다[37].

이 데이터를 인용한 2021년부터 2023년의 연구 20건 중 16건을 살펴본 결과, 실제로 이 데이터를 사용한 논문은 5건[43][44][45][46][47]이었다. 이 연구들은 새로운 프레임워크 하에서 개발한 모델의 효과성[43], Hidden Markov Model로 intrusion detection agility를 예측[44], AML(Adversarial Machine Learning)의 중요성[45], PageRank 특징으로 내부전파경로를 탐지[46], categorical GLM(Generalized linear Models) 모델로 각 사용자별 침입탐지 점수 측정[47] 등을 강조한 연구로, 본 연구처럼 인증 타입이나 로그온 타입을 고려한 이상 행위 탐지 연구가 아니다. 따라서 현재까지 시도되지 않았던 방법으로 이상 행위 탐지를 실시하였다.

3.3 Workload

연구 절차는 Fig. 1과 같다. 먼저, LANL2015 데이터셋에서 공격 데이터와 인증 데이터를 합쳐 50,000개를 랜덤 추출한다. 그 다음 7:3의 비율로 훈련, 테스트 셋을 분리한다. 분리한 각각의 데이터에서 이벤트 발생 빈도(frequency)를 계산하고, 가설 검증에 필요한 특징을 추가한다. 이 데이터를 레이블 인코딩, PCA를 실시한 후에 7가지 인공지능 모델을 훈련한다. 임선영 외[7]의 연구에서는 ABOD, HBOS, IForest, kNN, LOF, OCSVM 까지 총 6가지를 실시하였으나, 본 연구에서는 실험 환경 구축 과정에서 문제가 발생한 ABOD는 제외하고, 대신 Autoencoder와 SUOD를 추가하였다. 마지막으로 테스트 셋을 대상으로 이상 행위 탐지 성능을 측정한다.

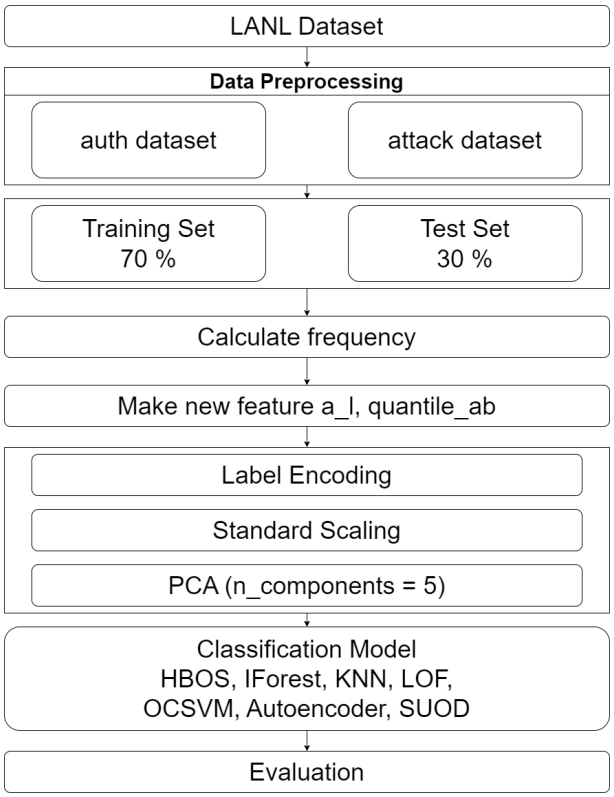

Fig. 1. Flowchart of the proposed anomaly detection methodology

3.3.1 Data preprocessing

인증 데이터에 포함된 레드팀 인증 로그를 추출하기 위해 인증 데이터와 레드팀 공격 이벤트 파일을 비교한 결과 실제로 인증 데이터에 포함되는 공격 데이터는 749개가 아닌 702개였다. 따라서 인증 데이터에서 공격 데이터를 702개 추출하였다. 그다음 positive 열을 추가하여 공격 데이터를 표현할 수 있도록 ‘1’ 값을 지정하였다. 인증 데이터에서는 150,000개의 데이터를 랜덤 추출하였고, 추출한 데이터에서 ‘?’로 표기된 결측치 데이터를 모두 삭제하고, 삭제한 데이터에서 49,298개를 추출하여 positive 열을 추가하고 정상 데이터를 표현하도록 ‘0’ 값을 지정하였다. 그 후 두 개의 데이터 프레임(인증/공격 데이터)을 합쳐 총 50,000개의 데이터를 사용하였다.

50,000개의 데이터에서 공격 데이터는 702개로 1.4%를 차지한다. 이상행위탐지를 위한 데이터셋의 경우에는 일반적인 데이터 분석용 데이터셋과 달리 비정상 데이터(공격 데이터)가 매우 적은 비율을 차지하는데, 현실에서 정상 데이터에 비해 공격 데이터의 빈도가 매우 낮은 특성을 반영한 것이다. 이러한 불균형적인 특성은 모델 학습을 어렵게 만들기 때문에, 여러 연구에서는 데이터셋의 비율을 조절하거나 오버샘플링을 사용하여 모델 학습을 용이하게

한다. 하지만 본 연구는 공격 데이터에 대한 명확한 사전 정보 없이 가설에 따라 모델을 학습하는 방법이므로 데이터셋 자체에 별도의 조치를 취하지 않았으며, 이처럼 불균형한 데이터 하에서 실시한 연구로는 임선영 외[7], 윤지영 외[17], 심기천과 김강석[48] 등에서도 확인할 수 있다.

Table 3과 같이 50,000개의 데이터를 훈련 데이터 35,000(공격 데이터 491개 포함)개와 테스트 데이터 15,000개(공격 데이터 211개 포함)로 7:3 비율로 분리하였다. 또한 본 연구에서는 도메인(source_domain, destination _domain) 특징과 user 정보를 함께 사용하므로 source_user@source_domain 정보를 '@'를 기준으로 분리하였다. 전처리 단계에서 분리한 특성들은 Table 4에서 확인할 수 있다.

Table 3. Number of redteam data in the train set and test set

	Train data(%)	Test data(%)
Total	35,000 (100%)	15,000 (100%)
Normal	34,509(98.6%)	14,789(98.6%)
Redteam	491(1.4%)	211(1.4%)

time은 인증 이벤트가 발생한 시간, source(destination) user@domain 인증 이벤트를 시작하는 사용자(또는 인증 이벤트가 매핑되는 사용자)를 의미한다[37]. source(destination) computer는 인증 이벤트를 발생(또는 종료)시킨 컴퓨터를 의미하며, authentication type(Negotiate, Kerberos, NTLM 등)은 인증이 발생하는 유형을 의미하고, logon_type은 (Network, Interactive keyboard session, Batch event, system service, screen saver Lock or Unlock 등을 통해 발생하는) 여러 인증 유형을 나타낸다[37]. authentication orientation은 인증 이벤트가 사용되는 방식을 의미하며, success/failure는 인증 이벤트가 성공하였는지 실패하였는지를 나타낸다[37].

Table 4. Original and preprocessed features of LANL2015 dataset

Raw data	Preprocess data
time	time
source user @domain	source_user
	source_domain
destination user @domain	destination_user
	destination_domain
source computer	source_computer
destination computer	destination_computer
authentication type	authentication_type
logon type	logon_type
authentication orientation	authentication_ orientation
success/failure	success/failure
-	positive

3.3.2 Feature engineering

가설 1(공격은 특정 인증 타입의 취약점을 악용했을 것이다)을 확인하기 위해 a_1 특성을 추가하였다. 훈련 데이터 중 공격 로그 491개의 authentication type과 logon type을 확인한 결과 모든 값이 NTLM, NETWORK로 구성되었다. 따라서 모든 train, test set에 authentication_type이 NTLM이고, logon_type이 NETWORK인 경우 1, 아닌 경우 0으로 저장하였다. 즉, 훈련 데이터셋의 모든 공격 데이터는 가설 1에 부합한다.

가설 2(특정 인증 타입의 취약점을 악용한 공격의 발생 빈도는 낮을 것이다.)를 확인하기 위해 frequency와 a_1_frequency를 추가하였다. frequency는 Table 5에 명시된 그룹 헤더를 기준으로 동일 이벤트가 발생하는 경우를 계산하였다. 훈련 데이터 중 491개의 공격 데이터에서 frequency를 확인한 결과 모두 7 이하의 값을 나타내었다. 따라서 a_1_frequency의 값을 1-7까지 조정하여 가설 1을 만족하면서 특정 빈도 n 이하인 경우 1, 아닌 경우 0의 값을 갖도록 하였다.

Table 5. Group headers for event frequency measurement

Group header
source_user
source_domain
destination_user
destination_domain
source_computer
destination_computer
authentication_type
logon_type
authentication_orientation
success/failure

3.3.3 Encoding, Standard Scaling, PCA

전처리한 데이터셋은 categorical data 형태를 가진다. 하지만 대부분의 기계학습 모델은 numerical data를 처리하기 때문에, 선택한 모델이 효율적으로 훈련하기 위해서 numerical data로 변환해야 한다[49]. 본 연구에서 categorical data를 numerical data로 변환하기 위해 레이블 인코딩(label encoding)을 사용하였다. 주로 사용되는 원-핫 인코딩(one-hot encoding)의 경우 한 열에 N개의 고유값이 있다면 N 차원의 벡터로 만들어 표현하는 방법이다. 따라서, 원-핫 인코딩의 과정을 거치면 훈련해야 할 벡터의 개수가 N만큼 증가하게 된다. 즉각적으로 이상 접근을 탐지해야 하는 이상 탐지에서 벡터가 증가하는 것은 시간적 효율을 낮추는 원인이 되기 때문에 원-핫 인코딩 대신에 레이블 인코딩을 채택하였다.

인코딩 방법으로 레이블 인코딩을 사용할 때, 레이블 인코딩의 최대값이 커지는 것을 방지하기 위해 source_user, destination_user의 U숫자, C숫자로 표현된 데이터들은 숫자 표기를 삭제하였다. 레이블 인코딩을 시행한 특징들은 Table 6과 같다. 레이블 인코딩 후 standard scaling과 PCA를 실시하였다.

Table 6. Encoding header

Encoding header
source_user
destination_user
authentication_type
logon_type
a_l
a_l_frequency

3.3.4 Performance metrics

제안하는 전처리 방법의 효과성을 확인하기 위해 측정 지표로 Table 7의 오차 행렬(Confusion Matrix), AUC(Area Under the Curve), 정확도(Accuracy), 정탐률(TPR), 오탐률(FPR), test에 소요된 실행 시간(초)을 사용하였다.

Table 7. Confusion matrix

		Predict	
		0	1
Actual	0	TN	FP
	1	FN	TP

TN은 정상 데이터를 정상으로 판별한 경우를 의미하고, TP는 이상 데이터를 이상 데이터로 판별한 경우를 의미한다.

Accuracy는 전체 결과에서 정상과 이상 값을 정확하게 예측한 비율을 의미하며, 수식(1)과 같다.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

TPR(정탐률, True Positive Rate)은 정상 데이터를 정상 데이터로 판별한 비율을 의미하고, 수식(2)와 같다.

$$TPR = \frac{TP}{TP + FN} \quad (2)$$

FPR(오탐률, False Positive Rate)은 정상 데이터를 이상 데이터로 판별한 비율을 의미하고, 수식(3)와 같다.

$$FPR = \frac{FP}{FP + TN} \quad (3)$$

AUC는 ROC Curve(Receiver Operating Characteristic curve)의 아래 면적을 의미하며, ROC Curve는 x축으로 FPR을 사용하고 y축으로 TPR을 사용하는 그래프이다.

IV. Experiments

4.1 Anomaly detection

본 연구의 실험 환경은 Table 8과 같으며, 성능 비교 대상인 임선영 외[7]의 실험 환경은 Table 9와 같다. 기존 연구[7]와 OS, CPU, Memory 등에서 동일한 실험 환경이라고는 할 수 없는 상황이다. 운영체제는 Ubuntu 20.04가 CentOS 7.9에 비해 최신 OS이므로, 이에 따른 성능 우위를 기대할 수 있다. CPU 벤치마크 점수[50]의 경우, 본 연구에서 사용하는 CPU 성능이 조금 더 우세(8.4%)하지만, single thread rating 점수에서 더 낮았다(-17.0%). 본 연구는 single thread로 시행되었기 때문에, 기존 연구[7]보다 오히려 불리한 환경에서 실시되었다고 볼 수 있다. 기존연구[7]와 동일하게 사용 언어는 Python이고, 이상 탐지 수행을 위해 PyOD 라이브러리[51]의 Autoencoder, HBOS, OCSVM, LOF, IForest, kNN, SUOD 알고리즘을 사용하였다. 성능 평가를 위해 sckit-learn 라이브러리[52]를 사용하였다.

Table 8. Experiment environment

OS	Ubuntu 20.04.4 LTS
Processor	Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz (CPU benchmark[50]:15204, single thread rating[50]:1806)
Memory	32GB
Programming Language	Python
Library	PyOD[51], Sckit-learn[52]

Table 9. Previous study[7]'s experiment environment

OS	CentOS 7.9
Processor	Intel(R) Xeon(R) CPU E5-2667 v4 (CPU benchmark[50]:13927, single thread rating[50]:2174)
Memory	20GB

4.1.1 Parameter

본 연구에서 적용한 대표적인 이상 탐지 알고리즘 7가지와 각 알고리즘에 적용한 파라미터는 Table 10과 같다. threshold를 조정하는 contamination은 모두 0.1로 설정하였다. 이는 기존 연구인 임선영 외[7]의 수치와 동일하다.

Table 10. Anomaly detection algorithm and parameters used in the experiment

Algorithm	Parameter
Auto encoder	hidden neurons = [4,2,2,4], epoch = 30
HBOS	n_bins = 3
OCSVM	default
LOF	n_neighbors = 300
IForest	max_features = 0.5, n_estimators = 10
kNN	n_neighbors = 100
SUOD	ocsvm, hbos(n_bins = 3)

4.2 Performance evaluation

HBOS 알고리즘의 경우 PCA 차원 축소를 했을 때와 안 했을 때의 TP와 FN 개수의 극단적 차이가 있어 Standard Scaling 후에 PCA 차원 축소를 한 경우와 하지 않은 경우로 나누어 결과를 확인하였다.

4.2.1 Anomaly Detection Results based on Authentication Type and Logon Type

가설 1을 확인하기 위해 source_user, destination_user, authentication_type, logon_type, a_l의 특성만을 사용하여 이상 탐지를 수행하였다. 이때 이상 탐지는 동일 이벤트의 발생 빈도인 frequency를 고려하지 않고 5회 반복하여 실행한 결과를 측정하였다. IForest의 경우 그 결과가 매번 달라지기 때문에 탐지 결과를 평균을 내었다. 모든 이상탐지 알고리즘의 평균 결과는 Table 11과 같다. 공격 데이터를 모두 탐지(TPR = 1)한 알고리즘은 HBOS, IForest, OCSVM, Autoencoder, SUOD이다. kNN 알고리즘의 경우 ACC가 가장 높았으나(0.974), 공격을 전혀 탐지하지 못하였다. 즉, TP가 0이고, FP가 가장 낮음(178)을 의미한다. 이것은 정상 접근을 공격으로 잘못 판단하는 오탐지 횟수가 가장 적음을 의미한다. 이 때문에 kNN 알고리즘을 사용하였을 때 AUC는 가장 낮은 수치(0.494)를 보였다. AUC는 OCSVM 알고리즘을 사용하였을 때 가장 점수(0.981)가 높았다. 처리 속도의 경우 HBOS가 가장 빠른 것(0.0033초)으로 확인하였다.

4.2.2 Anomaly detection results based on Frequency

가설 2를 확인하기 위해 source_user, destination_user, authentication_type, logon_type, a_l_frequency의 특성만을 사용하여 이상 탐지를 수행하였다. 이때 이상 탐지는 동일 이벤트의 발생 빈도인 frequency를 고려하여 1-7까지를 모두 측정하였다. 그 중 최적의 성능을 기록한 frequency에 따른 결과는 Table 12와 같다. 각 알고리즘의 최적의 frequency가 다른 것을 확인할 수 있다. 공격 데이터를 모두 탐지(TPR = 1)한 알고리즘은 HBOS, IForest, OCSVM, Autoencoder, SUOD이다. ACC가 가장 높은 알고리즘은 kNN(ACC=0.974)였고. AUC가 가장 높은 알고리즘은 OCSVM(AUC= 0.990), 처리 속도 면에서는 HBOS 알고리즘이 가장 빠른 것(0.0034, 0.0054초)을 확인하였다.

4.2.3 Anomaly detection results based on Authentication type, Logon type and Frequency

a_l, a_l_frequency의 특성을 모두 사용하여 이상 탐지를 실시한 결과 최적의 성능을 기록한 frequency에 따른 결과는 Table 13과 같다. 각 알고리즘의 최적 frequency가 다름을 확인할 수 있다. 각각의 특성 하나씩 사용했던 이전 결과와는 달리 가장 뛰어난 성능을 확인하였다. 공격을 모두 탐지(TPR = 1)한 알고리즘 HBOS, IForest, OCSVM, Autoencoder, SUOD이다. ACC, AUC, 실행시간은 HBOS에서 가장 높은 점수(ACC=0.985, AUC=0.992, SEC=0.0038)를 보였다.

Table 11. Results of performing an anomaly detection using feature a_l (Freq = frequency, SEC = second)

Algorithm	Freq	TP	FP	FN	TN	ACC	AUC	FPR	TPR	SEC
HBOS (scaling)	-	0	1387	211	13402	0.893	0.893	0.094	0.000	0.0034
HBOS (scaling, pca)	-	211	815	0	13974	0.946	0.960	0.055	1.000	0.0054
Iforest	-	126.6	983.8	84.4	13805.2	0.929	0.929	0.067	0.600	0.0486
kNN	-	0	178	211	14611	0.974	0.494	0.012	0.0000	8.1600
LOF	-	0	341	211	14448	0.963	0.489	0.023	0.000	7.2400
OCSVM	-	211	1223	0	13566	0.918	0.981	0.083	1.000	21.4000
Auto encoder	-	211	921	0	13868	0.939	0.957	0.062	1.000	1.6800
SUOD	-	211	1421	0	13368	0.905	0.972	0.096	1.000	3.0400

Table 12. The results of performing an anomaly detection using feature a_l_frequency (Freq = frequency, SEC = second)

Algorithm	Freq	TP	FP	FN	TN	ACC	AUC	FPR	TPR	SEC
HBOS (scaling)	1	0	1334	211	13455	0.897	0.868	0.090	0.000	0.0033
HBOS (scaling ,pca)	5	211	832	0	13957	0.945	0.955	0.056	1.000	0.0052
Iforest	5	211	923	0	13866	0.938	0.946	0.062	1.000	0.049
kNN	1	0	178	211	14611	0.974	0.494	0.012	0.000	8.1
LOF	5	0	341	211	14448	0.963	0.489	0.023	0.000	6.9
OCSVM	4	211	1048	0	13741	0.930	0.990	0.071	1.000	21
Auto encoder	3	211	846	0	13943	0.944	0.951	0.057	1.000	1.7
SUOD	1	211	878	0	13911	0.941	0.961	0.059	1.000	2.9

Table 13. Results of performing an anomaly detection using feature a_l and a_l_frequency (Freq = frequency, SEC = second)

Algorithm	Freq	TP	FP	FN	TN	ACC	AUC	FPR	TPR	SEC
HBOS (scaling)	4	211	230	0	14559	0.985	0.992	0.016	1.000	0.0038
HBOS (scaling, pca)	4	211	624	0	14165	0.958	0.966	0.042	1.000	0.0051
Iforest	4	211	909	0	13880	0.939	0.977	0.061	1.000	0.048
kNN	1	0	178	211	14611	0.974	0.494	0.012	0.000	8.2
LOF	5	0	341	211	14448	0.963	0.489	0.023	0.000	7
OCSVM	3	211	1114	0	13675	0.926	0.978	0.075	1.000	21
Auto encoder	5	211	832	0	13957	0.945	0.953	0.056	1.000	1.7
SUOD	3	211	740	0	14049	0.951	0.974	0.050	1.000	3

Table 14. Results of performance comparison with previous study(SEC = second)

	Algorithms	Proposed features	Previous work[7]	Difference(%, +-)
AUC	HBOS	0.992	0.96	+3.333
	Iforest	0.977	0.46	+112.391
	kNN	0.494	0.5	-1.200
	LOF	0.489	0.98	-50.102
	OCSVM	0.978	0.96	+1.875
TPR	HBOS	1	1	0.000
	Iforest	1	0.02	+4900.000
	kNN	0	0	-
	LOF	0	1	-100.000
	OCSVM	1	1	0.000
FPR	HBOS	0.016	0.08	-80.000
	Iforest	0.061	0.1	-39.000
	kNN	0.012	0	-
	LOF	0.023	0.05	-54.000
	OCSVM	0.075	0.08	-6.250
SEC	HBOS	0.0038	37.65	-99.990
	Iforest	0.048	142.08	-99.966
	kNN	8.2	4576.41	-99.821
	LOF	7	98.14	-92.867
	OCSVM	21	9675.56	-99.783

V. Discussions

본 연구를 통해 가설 1(공격은 특정 인증 타입의 취약점을 악용할 것이다)이 유효함을 확인하였다. 학습 결과, 모든 공격 이벤트는 'NTLM'의 인증 타입을 가지고, 'NETWORK'의 로그온 타입을 가진다. 따라서 이를 동시에 만족함을 의미하는 특징 'a_l'과 함께 source/destination_user, logon_type, authentication_type를 사용하여 이상 탐지를 수행한 결과, 7개 알고리즘 중 HBOS, IForest, OCSVM, Autoencoder, SUOD 5가지 알고리즘에서 공격 데이터 211개를 모두 탐지하여 높은 정확도(TPR 1.0)를 확인하였다.

가설 2(가설 1을 만족하는 공격 데이터는 발생 빈도(frequency)가 매우 낮을 것이다)를 확인하기 위해 빈도를 측정한 결과, 가설 1과 마찬가지로 5개의 알고리즘에서 공격 데이터 211개를 모두 탐지하였다. 하지만 오탐률의 경우, 가설 1과 가설 2는 큰 차이를 보이지 않았다.

마지막으로, 오탐률을 줄이기 위해 a_l, a_l_frequency 특성과 source/destination _user, logon_type, authentication_type까지 총 6개의 특성을 사용하여 훈련하고 이상 탐지를 수행했다. 가설 1과 가설 2의 결과와 비교하였을 때, kNN, LOF 알고리즘을 제외한 모든 알고리즘이 공격 데이터 211개를 모두 탐지(TPR 1.0)하였다.

Table 14의 결과는 가설 Table 13과 기존 연구[7]를 비교한 수치를 나타낸다. AUC의 경우 LOF, kNN 알고리즘을 제외한 나머지 알고리즘에서 최대 112% 상승하였고, 특히 FPR은 모든 알고리즘에서 최대 80%까지 감소하여 오탐율이 크게 줄어든 것을 확인하였다. 실행 시간의 경우 모든 알고리즘에서 99% 이상 감소하였다. 그 이유로는 원-핫 인코딩이 아닌 레이블 인코딩을 사용하여 특성의 개수를 줄였기 때문이다. 탐지하는 데에 시간이 줄어든 것은 공격이 발생했을 때 즉각적으로 대응할 수 있는 효과를 낼 수 있다.

또한 기존 연구[7]에서 최고 성능을 기록한 LOF(AUC 0.98)와 비교하여, HBOS의 경우 정탐지율을 100%로 유지하면서 오탐지율을 80% 감소시켜 기존 연구보다 우수한 성능(AUC 0.992)를 기록하였다. 또한 실행시간은 99.99%(37.65초에서 0.0038초) 감소시켰다. 오탐지율과 실행시간의 감소는 두 가지 이점이 있다. 첫째, 실제 환경에서도 빠른 시간에 급격히 누적되는 데이터를 빠른 시간에 학습하여 처리할 수 있다. 둘째, 실제 환경에서 오탐지로 인한 모니터링 부하를 줄임으로써 공격 시도에 좀 더 집중할 수 있다.

하지만 본 연구에서는 LOF의 경우 기존 연구[7]에서 최고 성능을 기록한 것에 비해, 공격 데이터를 전혀 탐지하지 못한 결과를 보였다. 또한 kNN 알고리즘도 이상행위를 전혀 탐지하지 못하였다. 이와 비슷한 사례로 Jiang과 Lin의 연구[53]에서 세 가지 데이터 셋에 대하여 동일 모델을 사용하여도 인코딩 방법에 따라서 예측 결과가 상이하게 바뀌는 결과를 확인 할 수 있었다. 따라서 데이터 전처리 및 모델 선택 시에 데이터 셋의 크기, 매개변수 등을 고려할 것을 제언하였다[53].

본 연구의 한계는 다음과 같다. 첫째, 가설의 한계가 있다. 다양한 공격 기법이 활용될 경우, 인증 타입과 로그인

타입 특성만으로는 공격 데이터 식별이 어려워져 이상 탐지 모델의 성능이 저하될 수 있다. 둘째, 공격 이벤트의 빈도가 잦은 경우, 이상 탐지의 빈도를 일정 threshold 이상으로 늘릴수록 정상 데이터와 공격 데이터를 구분하지 못해 오탐지율이 상승할 수 있다. 셋째, 본 연구는 특정 데이터셋(LANL2015)에 최적화 되어있어 다른 데이터셋으로 실시할 경우 매우 상이한 결과를 보일 가능성이 있다.

넷째, 본 연구에서 사용한 컴퓨팅 환경은 이전 연구와 다른 컴퓨팅 환경에서 실시되었으므로 공평한 환경에서 비교하였다고 할 수 없다. 하지만, 기존 연구의 제안 방법을 실행해 본 결과 HBOS 알고리즘의 경우 22초 소요되어 기존 연구[7]의 37.65초에 비해 약 42% 개선됨을 볼 수 있었다. 하지만 다른 알고리즘의 경우 기존 연구에서 충분한 정보가 제공되지 않아서 원핫인코딩 된 자료의 처리 시간이 몇 배나 오래 걸리거나 메모리 오버플로우가 발생하는 등 재현에 실패하여 동일한 환경에서의 검증이 어려웠다. 하지만 본 연구에서 제안한 방법에 HBOS 알고리즘 사용했을 때 소요시간이 99.99% 감소한 것은, 기존 feature를 융합한 핵심 feature가 성능 개선에 큰 효과를 준 것으로 판단된다. 또한 시간을 제외한 다른 측면(AUC, TPR, FPR) 에서의 성능 개선은 운영체제의 차이와 무관하다고 말할 수 있다.

VI. Conclusions

본 연구는 기존 이상 탐지 데이터셋에서 얻은 가설을 기반으로 새로운 특성을 추출하는 방법을 제시하고, LANL2015 데이터셋으로 이상 탐지를 실시하였다. PyOD 라이브러리에서 제공하는 대표적인 이상 탐지 알고리즘에 본 연구의 2가지 새로운 특성을 적용한 결과, TPR 측면에서는 kNN, LOF를 제외한 모든 알고리즘에서 100%의 성능을 확인하였다. 특히 기존연구[7]의 최고 성능(AUC 0.98)에 비해, 본 연구는 FPR(80%)와 실행 시간(99.99%)을 대폭 감소시킨 우수한 성능(AUC 0.992)를 기록했다. 이를 통해, 본 연구에서 제시한 특성인 공격 유형과 공격 발생 빈도가 이상 탐지에 있어 효과적임을 입증하였다.

향후 연구는 다음과 같다. 첫 번째, 본 연구는 LANL2015 데이터셋에서 효과적임을 보였지만, 다양한 공격 유형을 가진 데이터셋에서도 효과적임을 검증하기 위한 추가 연구가 필요하다. 두 번째, 데이터의 불균형을 감안하여 학습 및 테스트를 위한 데이터셋의 비율을 조절하여 구성하고, 구성한 데이터셋으로 실험하여 결과를 도출할 경우의 성능을 검증하기 위한 추가 연구가 필요하다. 세 번째, 실제 기업의 로그인 로그 데이터를 대상으로 본 연구에서 보여준 낮은 오탐지율과 실행시간이 효과적임을 검증해야 한다.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00211436).

REFERENCES

[1] S. Park and J. Lim, "Study On Identifying Cyber Attack Classification Through The Analysis of Cyber Attack Intention," Journal of the Korea Institute of Information Security and Cryptology, 27(1), pp. 103-113, Feb, 2017. DOI:10.13089/JKIISC.2017.27.1.103.

[2] Ministry of SMEs and Startups, "2022 Survey on the State of Technology Protection in Small and Medium-sized Enterprises, ", June 2022.

[3] Min Seonhee, "Virtual Currency Exchange 'GDAC' Hit by Hacking, 23% of Custodial Assets Stolen," Yonhap News, April, 2023.

[4] Sonicwall, 2023 Sonicwall Cyber Threat Report.

[5] F. Hachmi, K. Boujenfa and M. Limam, "Enhancing the Accuracy of Intrusion Detection Systems by Reducing the Rates of False Positives and False Negatives Through Multi-objective Optimization," J Netw Syst Manage, 27(1), pp. 93-120, Jan, 2019. DOI:10.1007/s10922-018-9459-y.

[6] A. D. Kent, "Comprehensive, multi-source cyber-security events data set," Los Alamos National Lab, 2015. DOI:10.17021/1179829.

[7] S. Im, S. Kim, S. Shim, S. Koo, B. Cho, K. Kim and T. Kim, "A Featurization Method to Improve Anomaly Detection Performance Using Login Logs," The Journal of Korean Institute of Communications and Information Sciences, 47(1), pp. 58-65, Jan, 2022. DOI:10.7840/kics.2022.47.1.58.

[8] X. Guan, W. Wang and X. Zhang, "Fast intrusion detection based on a non-negative matrix factorization model," Journal of Network and Computer Applications, 32(1), pp. 31-44, 2009. DOI:10.1016/j.jnca.2008.04.006.

[9] G. Tandon, "Machine learning for host-based anomaly detection," PhD Thesis. Florida Institute of Technology, 2008.

[10] Y. Qing, W. Xiaoping and Y. Bo, "An Intrusion Detection Approach Based on System Call Sequences and Rules Extraction," In 2010 2nd International Conference on E-business and Information System Security, pp. 1-4, May, 2010. DOI:10.1109/EBISS.2010.5473675.

[11] D. Yuxin, Y. Xuebing, Z. Di, D. Li and A. Zhanchao, "Feature

representation and selection in malicious code detection methods based on static system calls," Computers & security, 30(6), pp. 514-524, Sep, 2011. DOI:10.1016/j.cose.2011.05.007.

[12] M. Xie, J. Hu, X. Yu and E. Chang, "Evaluating Host-Based Anomaly Detection Systems: Application of the Frequency-Based Algorithms to ADFA-LD," Network and System Security, pp. 542-549, 2014. DOI:10.1007/978-3-319-11698-3_44.

[13] J. Arshad, P. Townend and J. Xu, "A novel intrusion severity analysis approach for Clouds," Future generation computer systems, 29(1), pp. 416-428, Jan, 2013. DOI:10.1016/j.future.2011.08.009.

[14] W. Khreich, S. S. Murtaza, A. Hamou-Lhadj and C. Talhi, "Combining heterogeneous anomaly detectors for improved software security," The Journal of systems and software, 137, pp. 415-429, Mar 2018. DOI:10.1016/j.jss.2017.02.050.

[15] D. Yeung and Y. Ding, "Host-based intrusion detection using dynamic and static behavioral models," Pattern Recognition, 36(1), pp. 229-243, 2003. DOI:10.1016/s0031-3203(02)00026-2.

[16] D. Mutz, F. Valeur, G. Vigna and C. Kruegel, "Anomalous system call detection," ACM transactions on information and system security, 9(1), pp. 61-93, Feb 01, 2006. DOI:10.1145/1127345.1127348.

[17] J. Yun, G. Shin, D. Kim, S. Kim and M. Han, "An Interpretable Log Anomaly System Using Bayesian Probability and Closed Sequence Pattern Mining," Journal of Internet Computing and Services, 22(2), 2021. DOI:10.7472/jksii.2021.22.2.77.

[18] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," Nature (London), 521(7553), pp. 436-444, May 28, 2015. DOI:10.1038/nature14539.

[19] D. Kwon, K. Natarajan, S. C. Suh, H. Kim and J. Kim, "An Empirical Study on Network Anomaly Detection Using Convolutional Neural Networks," ICDSC, pp. 1595-1598, Jul 2018. DOI:10.1109/ICDCS.2018.00178.

[20] S. Lv, J. Wang, Y. Yang and J. Liu, "Intrusion Prediction With System-Call Sequence-to-Sequence Model," IEEE Access, 6, pp. 71413-71421 2018. DOI:10.1109/ACCESS.2018.2881561.

[21] C. Kim, M. Jang, S. Seo, K. Park and P. Kang, "Intrusion Detection Based on Sequential Information Preserving Log Embedding Methods and Anomaly Detection Algorithms," IEEE Access, 9, pp. 58088-58101 2021. DOI:10.1109/ACCESS.2021.3071763.

[22] Sang-Hyun. Oh and Won-Suk Lee, "Anomaly Detection based on Clustering User's Behaviors," The transactions of the Korea Information Processing Society, 7(8), pp. 2411-2420. 2000.

[23] G. Kaiafas, G. Varisteas, S. Lagraa, R. State, C. D. Nguyen, T. Ries and M. Ourdane, "Detecting malicious authentication events trustfully," NOMS, pp. 1-6, Apr 2018. DOI:10.1109/NOMS.2018.8406295.

[24] E. Aghaei, "Machine Learning for Host-based Misuse and Anomaly Detection in UNIX Environment," PhD Thesis. University of Toledo, 2017.

[25] H. Siadati and N. Memon, "Detecting structurally anomalous logins within enterprise networks," Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1273-1284, 2017. DOI:10.1145/3133956.3134003.

[26] Q. Liu, J. W. Stokes, R. Mead, T. Burrell, I. Hellen, J. Lambert, A. Marochko and W. Cui, "Latte: Large-Scale Lateral Movement Detection," MILCOM, pp. 1-6, Oct, 2018. DOI:10.1109/MILCOM.2018.8599748.

[27] M. Meijerink, "Anomaly-based detection of lateral movement in a microsoft windows environment," Master's thesis, University of Twente, 2019.

[28] E. Besharati, M. Naderan and E. Namjoo, "LR-HIDS: logistic regression host-based intrusion detection system for cloud environments," J Ambient Intell Human Comput, 10(9), pp. 3669-3692, Sep, 2019. DOI:10.1007/s12652-018-1093-8.

[29] B. A. Powell, "Detecting malicious logins as graph anomalies," Journal of information security and applications, 54, pp. 102557, Oct 2020. DOI:10.1016/j.jisa.2020.102557.

[30] G. Creech and Jiankun Hu, "A Semantic Approach to Host-Based Intrusion Detection Systems Using Contiguousand Discontiguous System Call Patterns," IEEE Transactions on Computers, 63(4), pp. 807-819, Apr, 2014. DOI:10.1109/TC.2013.13.

[31] K. M. Tan and R. A. Maxion, "" Why 6?" Defining the operational limits of stide, an anomaly-based intrusion detector," In Proceedings 2002 IEEE Symposium on Security and Privacy, pp. 188-201, May, 2002. DOI:10.1109/SECPRI.2002.1004371.

[32] Miao Xie and Jiankun Hu, "Evaluating host-based anomaly detection systems: A preliminary analysis of ADFA-LD," CISP, 3, pp. 1711-1716, Dec, 2013. DOI:10.1109/CISP.2013.6743952.

[33] S. A. Maske and T. J. Parvat, "Advanced anomaly intrusion detection technique for host based system using system call patterns," International Conference on Inventive Computation Technologies, 2, pp. 1-4, Aug, 2016. DOI:10.1109/INVENTIVE.2016.7824846.

[34] K. Kim and K. Park, "Comparative Study of Anomaly Detection Accuracy of Intrusion Detection Systems Based on Various Data Preprocessing Techniques," KIPS Trans. Softw. and Data Eng, 10(11), pp. 449-456, 2021. DOI:10.3745/KTSDE.2021.10.11.449.

[35] S. Yoo and K. Kim, "Comparison of Anomaly Detection Performance Based on GRU Model Applying Various Data Preprocessing Techniques and Data Oversampling," Journal of The Korea Institute of Information Security & Cryptology, 32(2), April 2022.

[36] M. Liu, Z. Xue, X. Xu, C. Zhong and J. Chen, "Host-Based Intrusion Detection System with System Calls," ACM computing surveys, 51(5), pp. 1-36, Jan, 2019. DOI:10.1145/3214304.

[37] Alexander D. Kent, "Cyber security data sources for dynamic network research," Dynamic Networks and Cyber-Security, 1, pp. 37-65, 2016. DOI:10.1142/9781786340757_0002.

[38] S. A. Hofmeyr, S. Forrest and A. Somayaji, "Intrusion detection using sequences of system calls," Journal of computer security, 6(3), pp. 151-180, 1998. DOI:10.3233/JCS-980109.

[39] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. R. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod, R. K. Cunningham and M. A. Zissman, " Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection evaluation," In Proceedings DARPA Information Survivability Conference and Exposition, 2, pp. 12-26. 2000. DOI:10.1109/DISCEX.2000.821506.

[40] S. S. Murtaza, W. Khreich, A. Hamou-Lhadj and M. Couture, "A host-based anomaly detection approach by representing system calls as states of kernel modules," ISSRE, pp. 431-440, Nov, 2013. DOI:10.1109/ISSRE.2013.6698896.

[41] G. Creech and J. Hu, "Generation of a new IDS test dataset: Time to retire the KDD collection," 2013 IEEE Wireless Communications and Networking Conference (WCNC), Apr, 2013. DOI:10.1109/wcnc.2013.6555301.

[42] W. Haider, J. Hu, J. Slay, B. P. Turnbull and Y. Xie, "Generating realistic intrusion detection system dataset based on fuzzy qualitative modeling," Journal of Network and Computer Applications, 87, pp. 185-192, Jun, 2017. DOI:10.1016/j.jnca.2017.03.018.

[43] I. J. King and H. H. Huang, "Euler: Detecting Network Lateral Movement via Scalable Temporal Graph Link Prediction," Proceedings 2022 Network and Distributed System Security Symposium, 2022. DOI:https://doi.org/10.1145/3588771

[44] E. Muhati and D. B. Rawat, "Hidden-Markov-Model-Enabled Prediction and Visualization of Cyber Agility in IoT Era," JIoT, 9(12), pp. 9117-9127, Jun 15, 2022. DOI:10.1109/JIOT.2021.3056118

[45] E. Muhati and D. B. Rawat, "Adversarial Machine Learning for Inferring Augmented Cyber Agility Prediction, "INFOCOM WKSHPS, pp. 1-6, May 10, 2021. DOI:10.1109/INFOCOMWKSHPS51825.2021.9484471

[46] J. Yun, D. Kim, G. Shin, S. Kim and M. Han, "RDP-based Lateral Movement Detection using PageRank and Interpretable System using SHAP," Journal of Internet Computing and Services, 22(4), pp. 1-11, 2021, DOI : 10.7472/jksii.2021.22.4.1

[47] M. T. Wojnowicz, S. Aeron, E. L. Miller and M. Hughes, "Easy Variational Inference for Categorical Models via an Independent Binary Approximation," International Conference on Machine Learning, pp. 23857-23896, 2022.

[48] K. Sim and K. Kim, "Insider Anomaly Behavior Detection Method Using an Unsupervised Learning-Based Autoencoder," Journal of Digital Contents Society, 24(8), pp. 1929-1936, Aug 31, 2023. DOI:10.9728/dcs.2023.24.8.1929

[49] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," IEEE Access, 9, pp. 114381-114391 2021. DOI:10.1109/ACCESS.2021.3104357.

[50] PassMark, *https://www.cpubenchmark.net/compare/2830vs3752/Intel-Xeon-E5-2667-v4-vs-Intel-Xeon-Silver-4210R*, Sept 2023.

[51] Y. Zhao, Z. Nasrullah and Z. Li, "Pyod: A python toolbox for scalable outlier detection," Journal of Machine Learning Research, 20, pp. 1-7 2019.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, "Scikit-learn: Machine learning in Python," the Journal of machine Learning research, 12, pp. 2825-2830 2011.

[53] D. Jiang, W. Lin and N. Raghavan, "A Novel Framework for Semiconductor Manufacturing Final Test Yield Classification Using Machine Learning Techniques," IEEE Access, 8, pp. 197885-197895 2020. DOI:10.1109/ACCESS.2020.3034680.

Authors

Juyeon Lee received the B.S. degree in Mathematical Finance from Gachon University, Korea, in 2021, and received the M.Ed. degree in Computer Education from Korea National University of Education, Korea in 2023. She is interested in anomaly detection, and informatics education.

Daeseon Choi received the Ph.D. degree in computer science from Korea Advanced Institute of Science and Technology (KAIST), Korea. He is a currently professor in the Department of Software at Soongsil University, Korea. His main research directions include Privacy protection & Evaluation, AI security, Biometric/Beehavior Authentication, and Fraud Detection.

Seung-Hyun Kim received the Ph.D. degree in computer science from Korea Advanced Institute of Science and Technology (KAIST), Korea. He is currently an Associate Professor in the Department of Computer Education at Korea National University of Education, Korea. He is interested in computer education, information security, and so on.