J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

MITRE ATT&CK 기반 사이버 공격 목표 분류 : CIA 라벨링☆

Cyberattack Goal Classification Based on MITRE ATT&CK: CIA Labeling

신 찬 호[1] 최 창 희[1*]
Chan Ho Shin Chang-hee Choi

요 약

사이버 공격을 수행하는 주체와 그 목적이 점차 다양화되고 고도화되고 있다. 과거 사이버 공격은 개인 혹은 집단의 자신감 표출을 위해 수행되었지만, 최근에는 국가 단위의 후원을 받은 정치적, 경제적 목적의 공격도 활발히 이루어지고 있다. 이에 대응하고자 시그니처 기반의 악성코드 패밀리 분류, 공격 주체 분류 등이 이루어졌지만 공격 주체가 의도적으로 방어자를 속일 수 있다는 단점이 있다. 또한 공격의 주체, 방법, 목적과 목표가 다양해짐에 따라, 공격의 모든 과정을 분석하는 것은 비효율적이다. 따라서 방어자 관점에서 사이버 공격의 최종 목표를 식별해 유연하게 대응할 필요가 있다. 사이버 공격의 근본적인 목표는 대상의 정보보안을 훼손하는 것이다. 정보보안은 정보자산의 기밀성, 무결성, 가용성을 보존함으로써 달성된다. 이에 본 논문에서는 MITRE ATT&CKⓇ 매트릭스에 기반하여 공격자의 목표를 정보보안의 3요소 관점에서 재정의하고, 이를 머신러닝 모델과 딥러닝 모델을 통해 예측하였다. 실험 결과 최대 80%의 정확도로 예측하는 것을 확인할 수 있었다.

☞ 주제어 : 라벨링, 머신러닝, 딥러닝, MITRE ATT&CK, TTP, 정보보안 3요소

ABSTRACT

Various subjects are carrying out cyberattacks using a variety of tactics and techniques. Additionally, cyberattacks for political and economic purposes are also being carried out by groups which is sponsored by its nation. To deal with cyberattacks, researchers used to classify the malware family and the subjects of the attack based on malware signature. Unfortunately, attackers can easily masquerade as other group. Also, as the attack varies with subject, techniques, and purpose, it is more effective for defenders to identify the attacker's purpose and goal to respond appropriately. The essential goal of cyberattacks is to threaten the information security of the target assets. Information security is achieved by preserving the confidentiality, integrity, and availability of the assets. In this paper, we relabel the attacker's goal based on MITRE ATT&CKⓇ in the point of CIA triad as well as classifying cyber security reports to verify the labeling method. Experimental results show that the model classified the proposed CIA label with at most 80% probability.

☞ keyword : labeling, machine learning, deep learning, MITRE ATT&CK, TTP, CIA triad

1. 서 론

많은 시스템과 자산이 전산화되고, 다양한 형태의 소프트웨어가 등장하고 있다. 그리고 공격대상과 벡터가 많아진 만큼 보안 위협의 수도 나날이 증가하고 있다. 특히, 디지털 트윈(Digital Twin), 스카다 시스템(SCADA system), 임베디드 시스템(Embedded system) 등 서로 다른 목적을 가진 시스템이 등장하면서, 보안 위협은 다양한 형태로 나타나고 있다. 개인의 자신감 표출을 위해서 사이버 공격이 이루어진 과거와 다르게, 오늘날에는 경제적, 정치적 목적의 사이버 공격도 이루어지고 있다[1-6]. 초기에는 사이버 공격을 막기 위해 패턴 분석 및 시그니처 기반의 탐지 방법이 많이 연구되었다[7-12]. 하지만 공격 목적을 달성하기 위한 수단에는 여러 종류의 기술이 존재하기 때문에, 국가 단위의 공격에서는 쉽게 우회될 수 있다는 단점이 있다.

MITRE는 공격에 사용되는 여러 기술들을 TTP(Tactics, Techniques, and Procedures)로 정리하여 ATT&CK[13]를

1 Defense Cyber Technology Center, Agency for Defense Development, Seoul, 05661, Korea.
* Corresponding author (changhee84@add.re.kr)
[Received 25 August 2022, Reviewed 17 September 2022(R2 25 October 2022, Accepted 4 November 2022]
☆ 본 논문은 2022년 한국인터넷정보학회 춘계학술발표대회에서 발표한 논문인 "CIA 라벨 기반 사이버 공격 목표 분류 기술"을 확장한 것이다.
☆ 이 논문은 2022년 정부(방위사업청)의 재원으로 국방과학연구소의 지원을 받아 수행된 연구임(912880601)

발표하였다. ATT&CK는 현재 버전 11.3까지 나왔으며, 많은 업체와 기관에서 표준 수준으로 활용되고 있다. 사후 분석 측면에서 ATT&CK를 통해 사이버 공격을 분석하면 공격의 흐름을 체크하고 정보를 공유하는 데에 유용하다는 장점이 있다. 하지만 방어 및 예방 측면에서 ATT&CK 내의 모든 공격기술 자체에 일일이 대응하는 것은 비효율적일 수 있다. 따라서 공격자의 의도와 목표를 파악하는 것이 중요하다. 사이버 공격의 근본적인 목적은 공격대상의 정보보안을 훼손하는 것이다. 정보보안은 정보자산의 기밀성(Confidentiality,), 무결성(Integrity), 그리고 가용성(Availability)을 보존함으로써 달성된다 [14].

본 논문에서는 ATT&CK 매트릭스의 TTP를 기반으로 공격자의 목표를 정보보안의 3요소(CIA) 관점에서 재정의하고 이를 머신러닝 모델과 딥러닝 모델을 통해 예측하여 그 합리성을 보인다. 논문 구성은 다음과 같다. 2장에서는 공격 목표 식별 및 분류에 관한 과거 연구를 소개한다. 3장에서는 MITRE ATT&CK에 대한 분석과 정보보안의 3요소에 대한 상세한 설명을 기술하였다. 4장에서는 실험에 사용한 데이터셋과 샘플링 방법, 사용한 모델에 대해 설명하였다. 5장에서는 실험 결과와 그 결과 해석에 관하여 기술하였다. 마지막으로 6장에서는 결론과 앞으로의 연구 계획을 제시하였다.

2. 관련연구

과거 진행된 논문은 공격 목표보다는 악성코드 분류, 공격 그룹 분류에 관한 연구들이었다. 초창기에는 유사한 악성코드를 분류하기 위해 악성코드에 사용되는 API 등과 같은 시그니처 정보를 이용하였다[7-8]. 이후 APT공격을 비롯해 규모가 큰 사이버 공격이 나타나면서 악성코드 자체만의 시그니처를 이용한 분류가 어려워졌다. 이에 연구진들은 이메일 정보[9-10], 네트워크 트래픽 정보[11-12] 등 추가적인 특징점을 추출해내 공격 그룹을 추정하는 연구를 진행하였다. 하지만 이 역시 결국 시그니처를 이용한 분류 방법으로, 공격자가 의도적으로 분석가를 속일 수 있다[15-16].

시그니처를 이용한 공격그룹 분류가 가지는 단점을 해결하고자, 공격의 의도와 목표를 파악하고자 하는 연구가 진행되었다. Park 연구팀은 정치, 경제적 피해를 중심으로 사이버 공격의 의도를 분석하고, APT 공격을 분류하는 방법을 제안하였다. 연구팀은 피해자의 SNS 이미지 분석과 주식 정보를 이용하여, 사이버 공격의 의도를 사이버 전쟁, 테러, 범죄 등으로 분류하였다[17]. 하지만 이 연구는 3가지 케이스에 대해서만 공격유형을 검증하는데 그쳤다. 또한 MITRE ATT&CK 정보를 이용하여 공격 목표 분류를 시도한 연구가 진행되었다[18]. 이 연구에서는 수집한 사이버 공격 보고서에서 TTP를 자동으로 추출하여 데이터 셋으로 사용하였다. 추출한 TTP를 이용하여 보고서를 벡터화하고, 이를 딥러닝 모델을 이용하여 공격 목표를 예측하였다. 제안된 모델은 공격 목표에 해당하는 TTP를 예측하지만, 정답 라벨의 경우의 수가 너무 많아 정확한 예측이 힘들며 학습데이터 수가 적어 편향이 있다. 또한 공격자가 같은 목적 달성을 위해 다른 공격기술을 사용할 수 있으므로 쉽게 우회될 수 있다.

한편 APT 공격이 증가하면서 안랩(Ahnlab), 카스퍼스키(Kaspersky), 파이어아이(FireEye) 등 보안업체 등에서는 보고서를 내어 공격의 목표를 분석한 보고서를 발표하였다[4-5, 15, 19-21]. 하지만 이와 같은 전문가 분석 보고서는 사후분석이기 때문에 오래 걸린다는 단점이 있어 방어자 입장에서 대응이 늦어질 수 있다.

따라서 본 논문에서는 과거 연구들의 단점을 보완하고 사이버 공격에 효율적인 대응을 하기 위해 공격 목표를 재정의하고, 이를 머신러닝 및 딥러닝 모델을 통해 검증하고자 한다.

3. MITRE ATT&CK 기반 CIA 라벨링

3.1 MITRE ATT&CK

MITRE 사에서는 실제 공격사례를 기반으로 공격자의 전술과 공격기술, 절차를 정리한 ATT&CK를 발표, 관리하고 있다. 방어자는 이를 활용하여 보다 신속하고 효율적으로 공격을 정리하여 배포하거나, 배포된 공격예시를 참고할 수 있다. ATT&CK는 버전 11.0을 기준으로 14개의 전술(tactic), 191개의 공격기술(technique), 386개의 하위공격기술(sub-technique)로 이루어져 있다. 논문에서는 버전 10.0을 기준으로 라벨링과 분류작업을 수행하였다.

3.2 정보보안의 3요소와 CIA 라벨링

정보자산을 위협으로부터 보호하는 것을 정보보안이라 한다. 정보보안은 크게 기밀성, 무결성, 가용성 3가지 측면에서의 균형 있는 보호방법으로 달성된다. 기밀성이란 인가된 사용자만이 정보를 확인할 수 있어야 함을 말한다. 무결성이란 인가된 사람만이 정보를 수정, 변경할 수 있음을 말한다. 가용성은 인가된 사용자라면 언제나 그 정보에 접근하고자 할 때, 서비스가 원활하게 제공되어야 함을 말한다.

‘TA0010(Exfiltration)’은 정보유출에 관한 전술이다. 공격자는 여러 방법을 통해 수집한 데이터를 네트워크나

USB 등을 이용하여 데이터를 빼내간다. 따라서 방어자의 기밀성이 훼손되었다고 볼 수 있다. 'TA0040(Impact)'은 공격으로 발생한 데이터나 자원의 훼손, 파괴 등을 포함하고 있다. 따라서 이에 속한 공격기술들은 방어자 시스템의 무결성 훼손 혹은 가용성 침해 등에 사용될 수 있다. 특히 MITRE에서 'TA0040'의 공격기술들을 'impact type' 이라는 속성을 통해 무결성, 가용성 중 어느 목적으로 사용할 수 있는지 정의해두었다. 본 논문에서는 기본적으로 이를 참고하여 무결성, 가용성 라벨을 추가하였다. 추가로 OS 부팅 메커니즘, 펌웨어, 계정, 시스템 프로세스 등 시스템에 크게 영향을 끼치는 부분을 변조하는 공격 기술도 무결성 라벨로 정의하여 실험을 진행하였다. 기밀성, 무결성, 가용성 3가지 요소를 기준으로 ATT&CK를 정리하면 표 1과 같이 라벨링을 새로 정의할 수 있다.

(표 1) CIA 라벨링된 테크닉
(Table 1) CIA Labeled Techniques

Tactic	Technique	Proposed Label
TA0003 (Persistence)	T1542	Integrity
	T1098	Integrity
	T1554	Integrity
	T1543	Integrity
TA0010 (Exfiltration)	T1020	Confidentiality
	T1030	Confidentiality
	T1048	Confidentiality
	T1041	Confidentiality
	T1011	Confidentiality
	T1052	Confidentiality
	T1567	Confidentiality
	T1029	Confidentiality
	T1537	Confidentiality
TA0040 (Impact)	T1491	Integrity
	T1565	Integrity
	T1531	Availability
	T1485	Availability
	T1486	Availability
	T1561	Availability
	T1499	Availability
	T1495	Availability
	T1490	Availability
	T1498	Availability
	T1496	Availability
	T1489	Availability
	T1529	Availability

4. 데이터셋과 모델

4.1 데이터셋

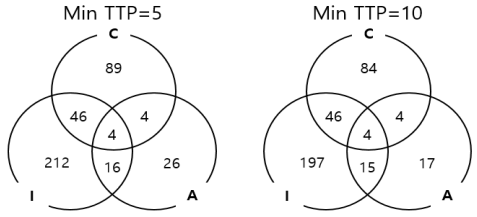
실험에는 APT&Cybercriminals Campaign Collection[22] (이하 CCC) 데이터셋을 사용하였다. CCC 데이터셋은 APT 및 사이버공격 캠페인들에 대한 블로그 게시물, 각종 보고서, 발표자료 등으로 이루어져있다. 연도와 날짜, 그리고 부분적으로 공격 그룹에 대한 설명을 포함하고 있으며 현재 2002년부터 2021년까지의 데이터를 수집하였다. 수집된 보고서는 총 1428개로, 이후 'rcATT'[23-24]를 이용하여 TTP를 태깅하였다.

'rcATT'는 보고서 내용을 입력으로 받아 머신러닝 모델을 이용하여 문서 내용으로부터 ATT&CK의 TTP를 태깅하는 도구이다. 태깅은 전술과 공격기술 레벨이 따로 이루어진다. 또한 기존 'rcATT'는 MITRE ATT&CK 버전 5를 기준으로 만들어졌기 때문에 버전 10.0 기준으로 일부 코드를 수정, 재학습하여 태깅에 사용하였다.

'rcATT' 태깅 후, 표 1에서 정의한 새 라벨의 분포는 표 2, 표 3과 같다. 보고서에서 추출된 최소 TTP 개수가 5개, 10개인 경우를 정리하였으며, 일부 보고서는 본문 내용이 추출되지 않아 제외되었다. 또한 한 보고서가 여러 개의 공격 기술을 담고 있으므로, 보고서는 여러 개의 라벨을 가질 수 있다. 그림 1은 CIA 라벨이 서로 얼마나 겹치는지 벤다이어그램으로 표현한 것이다. CIA 라벨링 예측은 멀티라벨 문제로, 이를 해결하기 위해, 제안하는 각 라벨에 대한 이진분류 문제를 수행한다.

(표 2) rcATT 데이터셋 CIA 라벨 분포
(Table 2) CIA label distribution of rcATT dataset

Label	Number of Reports Minimum TTP=5	Number of Reports Minimum TTP=10
Confidentiality	143/1042 (13.7%)	138/803 (17.2%)
Integrity	278/1042 (26.7%)	262/803 (32.6%)
Availability	50/1042 (4.8%)	40/803 (5.0%)



(그림 1) CIA 멀티라벨 분포
(Figure 1) CIA multi label distribution

(표 3) rcATT 데이터셋의 연도별 CIA라벨 분포
(Table 3) Yearly CIA label distribution of rcATT dataset

Year	Number of Reports minimum TTP=5						Number of Reports minimum TTP=10					
	Confidentiality		Integrity		Availability		Confidentiality		Integrity		Availability	
	True	False	True	False	True	False	True	False	True	False	True	False
2002	0	1	1	0	0	1	0	0	0	0	0	0
2007	1	3	2	2	1	3	1	2	1	2	0	3
2010	0	9	6	3	0	9	0	6	5	1	0	6
2011	0	7	2	5	0	7	0	3	1	2	0	3
2012	0	20	2	18	1	19	0	13	1	12	0	13
2013	3	49	15	37	1	51	2	39	15	26	0	41
2014	8	83	20	71	2	89	8	57	17	48	2	63
2015	5	97	30	72	2	100	5	67	29	43	1	71
2016	15	93	32	76	6	102	15	67	32	50	4	78
2017	7	86	28	65	9	84	7	68	27	48	9	66
2018	16	106	26	96	11	111	14	76	23	67	8	82
2019	27	133	40	120	4	156	25	101	37	89	4	122
2020	35	126	46	115	8	153	35	106	46	95	7	134
2021	24	68	21	71	4	88	24	46	21	49	4	66
2022	2	18	7	13	1	19	2	14	7	9	1	15
합계	143	899	278	764	50	992	138	665	262	541	40	763

(표 4) Bitdefender – StrongPity APT 보고서 rcATT 태깅 결과
(Table 4) rcATT tagging results of report 'Bitdefender – StrongPity APT'

Level	TTPs
Tactics	TA0042 TA0001 TA0002 TA0003 TA0004 TA0005 TA0009 TA0011 TA0010
Techniques	T1071 T1560 T1119 T1020 T1543 T1587 T1189 T1041 T1070 T1105 T1036 T1571 T1027 T1090 T1553 T1569 T1205 T1204 T1078

4.2 데이터셋 케이스 분석

표 4는 Bitdefender 사에서 작성한 'StrongPity APT – Revealing Trojanized Tools, Working Hours and Infrastructure'[19] 라는 보고서에 대한 'rcATT' 태깅 결과이다. 보고서는 정부 후원을 받는 것으로 추정되는 StrongPity 그룹의 공격 방식에 대한 내용을 담고 있다.

태깅 결과 총 9개의 전술과 19개의 공격기술이 태깅되었다. 또한 기밀성 라벨에 해당하는 공격기술로 'T1020'이 태깅되었음을 확인할 수 있다. 'T1020'은 'Automated Exfiltration' 공격기술로 공격자가 피해자의 정보를 취합하여 자동으로 바깥으로 빼가는 기술을 말한다. 보고서 내에 해당하는 문구는 다음과 같다.

"The Exfiltration component is responsible for running to the File Searcher component and for exfiltrating the files to the C&C server through a POST request. ~~ The .sft files are read by the Exfiltration component, sent to the C&C server and deleted from the disk."

4.3 샘플링 방법

논문에서 수집한 데이터셋의 분포가 매우 불균등하기 때문에 여러 가지 샘플링 방법을 사용하여 다양한 실험을 진행하였다. 샘플링 방법은 수집한 데이터를 모두 활용하는 경우와 일부만 선택하여 활용하는 경우로 나눌 수 있다. 또 일부를 선택하는 경우에도 최신 데이터만을 취할지, 과거 보고서부터 최신 보고서까지 골고루 취할지 등을 기준으로 선택하여 샘플링할 수 있다.

먼저 첫 번째 방법은 가지고 있는 보고서를 모두 사용하는 방법이다. 이 경우, 연도와 일자 상관없이 보고서를 무작위로 섞어 학습, 테스트 데이터로 나눈다. 논문에서는 이를 'random all' 샘플링이라고 정의한다.

두 번째 방법은 가지고 있는 보고서를 모두 사용하되, 연도를 기준으로 학습, 테스트 데이터로 나누는 방법이다. 논문에서는 이를 'standard all' 샘플링이라고 정의한다.

(표 5) 샘플링 후 데이터 분포
(Table 5) Distribution of sampled data

Label	Minimum TTP	Dataset Type	Random all		Standard all		Balanced random		Balanced standard	
			True	False	True	False	True	False	True	False
Confidentiality	5	Train	114	719	96	737	116	112	105	123
		Test	29	180	47	162	27	31	38	20
	10	Train	110	532	93	549	109	111	104	116
		Test	28	133	45	116	29	27	34	22
Integrity	5	Train	222	611	224	609	225	219	226	218
		Test	56	153	54	155	53	59	52	60
	10	Train	209	433	209	433	204	215	209	210
		Test	53	108	53	108	58	47	53	52
Availability	5	Train	40	793	40	793	41	39	40	40
		Test	10	199	10	199	9	11	10	10
	10	Train	32	610	33	609	32	32	32	32
		Test	8	153	7	154	8	8	8	8

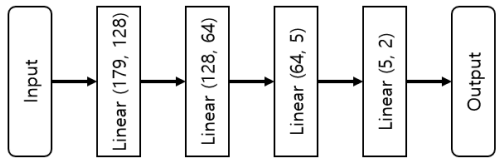
세 번째 방법은 수집한 보고서의 라벨 분포가 1:1이 되도록 샘플링하는 방법이다. True 라벨 데이터는 모두 사용하고 True 라벨 데이터와 같은 수가 되도록 False 라벨 데이터를 샘플링한다. 이 때, 연도와 일자 상관없이 보고서를 무작위로 섞어 학습, 테스트 데이터로 나눈다. 논문에서는 이를 'balanced random' 샘플링이라 정의한다.

네 번째 방법도 수집한 보고서의 라벨 분포가 1:1이 되도록 샘플링하는 방법이다. True 라벨 데이터는 모두 사용하되, False 라벨 데이터는 무작위로 선택한다. 그 후, 연도를 기준으로 과거 데이터는 학습 데이터로 사용하고 최신 데이터는 테스트 데이터로 사용한다. 'balanced random' 샘플링과 똑같이 True 라벨 데이터는 모두 사용한다. 논문에서는 이를 'balanced standard' 샘플링이라고 정의한다.

모든 샘플링 방법에서 학습과 테스트 데이터셋은 8:2 비율로 나누었다. 각 5회 실험을 돌린 후, 대표실험 하나씩만 뽑아 샘플링된 데이터의 분포를 보면 표 5와 같다.

4.4 실험 모델

실험 모델로는 머신러닝 모델 4종과 딥러닝 모델 1종을 사용하여 이진 분류를 수행하였다. 머신러닝 모델로는 sklearn 라이브러리[25]의 LogisticRegression, SVM, DecisionTreeClassifier 3종과 XGBoost[26] 모델을 사용하였다. 딥러닝 모델로는 간단한 FCN(Fully Connected Network) 모델을 사용하였다. 딥러닝 모델의 구조는 그림 2와 같다. 학습은 2500 epoch 동안 진행하였으며, 배치는 학습 데이터의 전체 사이즈, optimizer는 AdamW, learning rate는 10^{-3}를 사용하였다. 또한 모든 머신러닝 모델은 기본 파라미터로 초기화하여 실험에 사용하였다.



(그림 2) FCN 모델 구조
(Figure 2) FCN model structure

(표 6) 분류성능 평가지표
(Table 6) Metrics for classification performance

		Actual condition	
		Positive	Negative
Predicted condition	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

5. 실험 결과

모델의 성능을 평가하는 요소는 실제 정답과 모델의 예측 간의 관계로 정의할 수 있다. 표 6은 실제 정답과 모델 예측 간의 관계를 정의한 분류성능 평가지표이다. True Positive(TP)는 실제 Positive를 모델이 Positive로 예측한 경우이며, 정답에 해당한다. False Positive(FP)는 1종 오류(Type 1 Error)라고도 불리며, 실제 Negative를 모델이 Positive로 예측한 경우이다. False Negative(FN)는 2종 오류(Type 2 Error)라고도 불리며, 실제 Positive를 모델이 Negative로 예측한 경우이다. True Negative(TN)는 실제 Negative를 모델이 Negative로 예측한 경우로 정답에 해당한다.

모델의 성능은 정확도(Accuracy)와 F1 점수, ROC (Receiver Operating Characteristic) 곡선을 기준으로 평가하였다. 정확도는 일반적인 모델의 성능을 평가하는 가장 대중적인 평가지표다. F1 점수는 정밀도(Precision)와 재현율(Recall)의 조화평균으로, 라벨의 분포가 불균등할 때 주로 사용되는 평가지표이다. ROC 곡선은 재현율이 변함에 따라 TPR(True Positive Rate)이 어떻게 변하는지를 표현한 그래프이다. 재현율은 낮을수록, TPR은 높을수록 성능이 좋기 때문에 그래프가 직각에 가까울수록 성능이 좋음을 의미한다. 다만, 재현율에 따른 TPR을 그린 그래프이기 때문에 데이터셋이 균형적일 때 의미가 있는 성능 지표다. 실험은 5회 반복 수행하여 평균을 내었다. 다만, 딥러닝 모델인 FCN 모델의 경우, loss값이 떨어지지 않는 등 학습이 되지 않는 경우는 평균에서 제외하였다.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$$

$$Precision = \frac{TP}{TP+FP}$$

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

$$FPR = \frac{FP}{TN+FN}$$

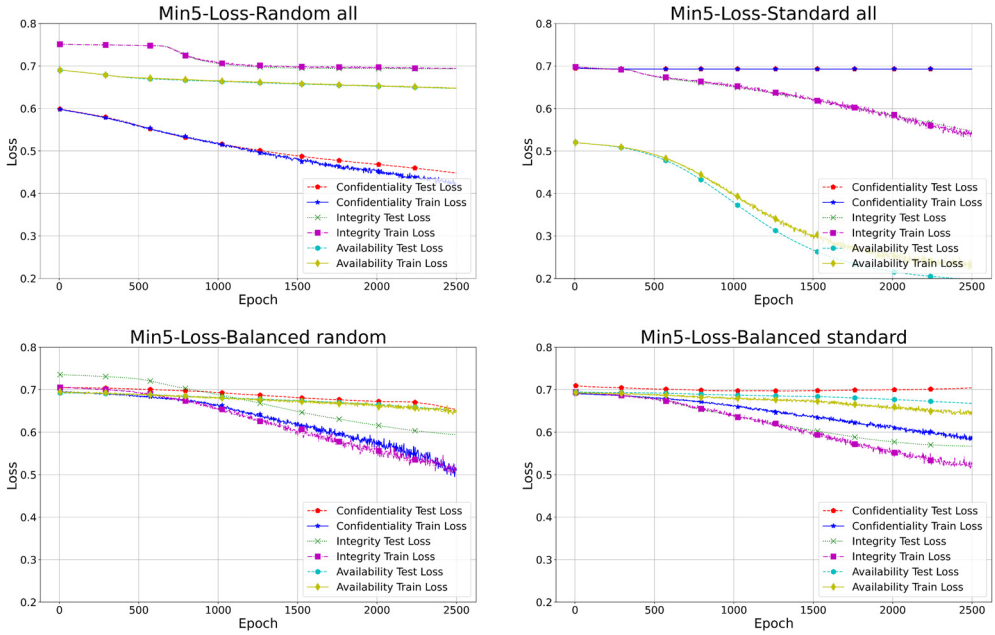
$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5.1 FCN 모델의 Loss 분석

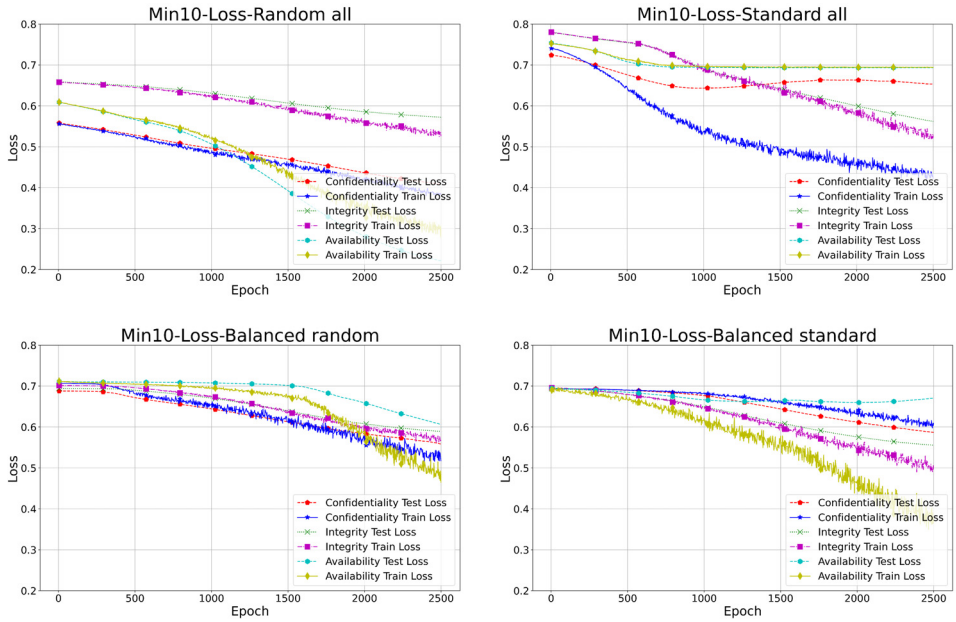
그림 3과 4는 epoch에 따른 FCN 딥러닝 모델의 loss 그래프이다. 5회의 실험 중 대표 실험 1건만 표시하였다. 그림 3은 TTP 개수가 최소 5개인 보고서를 대상으로 한 실험이며, 그림 4는 TTP 개수가 최소 10개인 보고서를 대상으로 한 실험이다.

'random all', 'balanced random', 'balanced standard' 샘플링의 경우 그림 3과 4에서 모두 2500 epoch까지 loss가 완만하게 떨어지고 있다.

'standard all' 샘플링은 loss가 초반에 소폭 감소하다가 loss가 다시 올라가는 경우가 있었다. 또한 그림 3과 같이 일부 실험에서는 학습이 되지 않아 loss가 유의미하게 떨어지지 않는 경우도 존재하였다.



(그림 3) Min5 Loss 그래프
(Figure 3) Min5 Loss Graph



(그림 4) Min10 Loss 그래프
(Figure 4) Min10 Loss Graph

결론적으로 4가지의 샘플링 모두 각 라벨에 대해 학습이 가능하나, 일부 샘플링의 경우 모델의 초기화가 적절하게 이루어지지 않으면 학습의 진행이 되지 않는 경우를 보였다.

5.2 정확도 및 F1 점수 결과분석

실험결과는 표 7에서 볼 수 있다. 먼저 'random all', 'standard all' 등 샘플링을 수행하지 않은 경우, 정확도는 0.67~0.97, F1 점수는 0.00~0.66로 계산되었다. 상대적으로 데이터가 많은 C라벨과 I라벨의 경우 F1 점수가 0.20~0.66으로 준수한 성능을 보였으나, A라벨은 데이터가 적고, 연도를 기준으로 학습 및 테스트 데이터셋을 나누면 라벨의 불균형이 심해져 F1 점수가 낮은 것을 볼 수 있다.

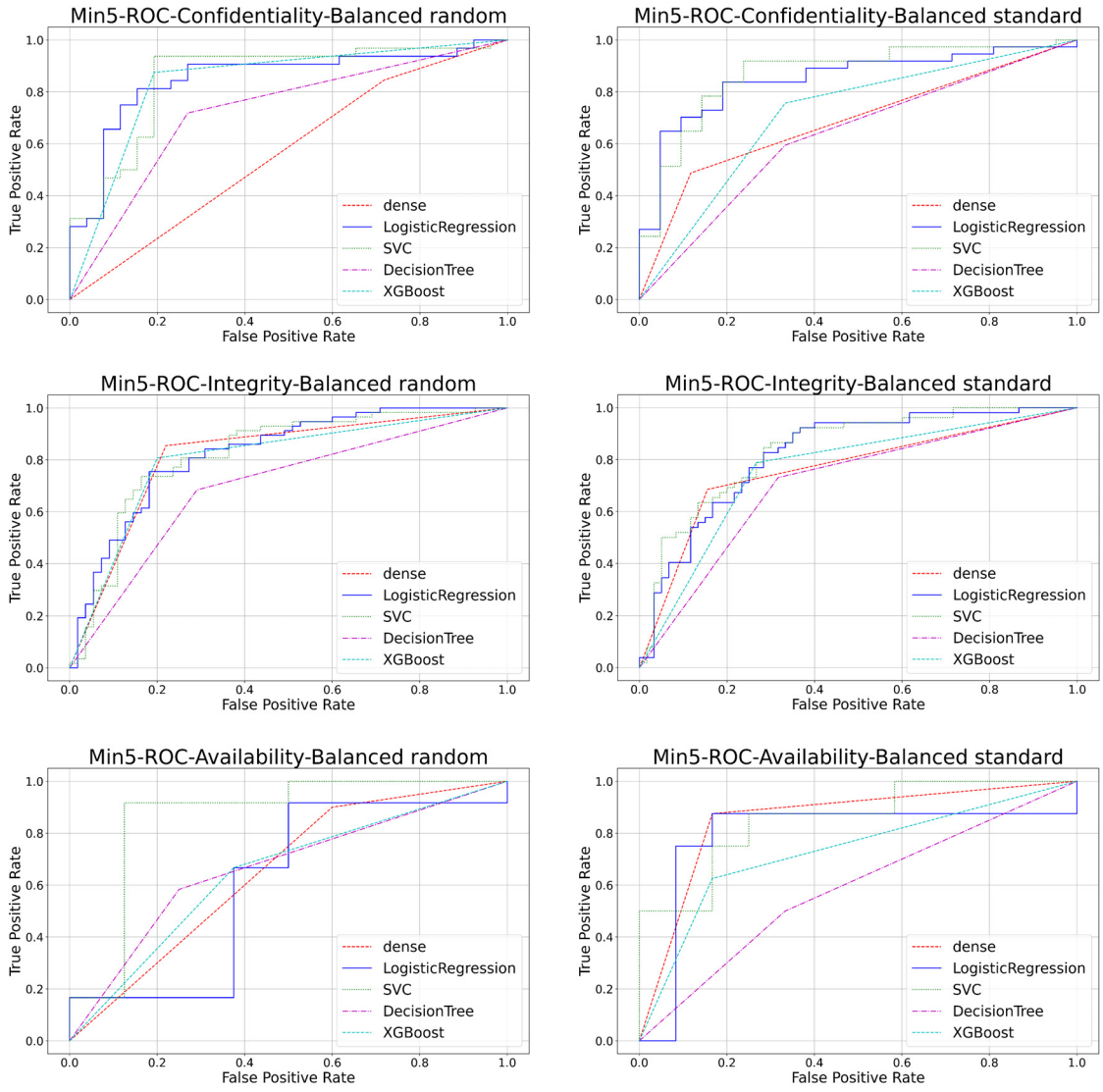
실험에 사용한 모델 중에서는 LogisticRegression, SVM, FCN 모델이 준수한 성능을 내었다. 특히 FCN의 경우 학습에 랜덤한 요소가 존재하기 때문에 loss가 떨어지지 않고 학습이 되지 않는 경우엔 성능이 낮았으나, 모델 초기화가 잘 되어 학습이 원활하게 수행될 경우, 가장 높은 성능을 보이기도 하였다.

'balanced random', 'balanced standard' 등의 샘플링 방법을 수행했을 때는, 라벨의 분포가 고르기 때문에 정확도와 F1 점수가 0.60~0.85로 비슷하게 나타난다. TTP 개수가 최소 10개 이상인 문서들을 튜닝 없는 간단한 모델로 세 라벨 모두 평균 70% 이상의 정답률을 보였다. 다만 'balanced random' 샘플링의 경우, 샘플링에 랜덤요소가 크게 관여하기 때문에 실험결과의 편차가 큰 것을 확인할 수 있었다.

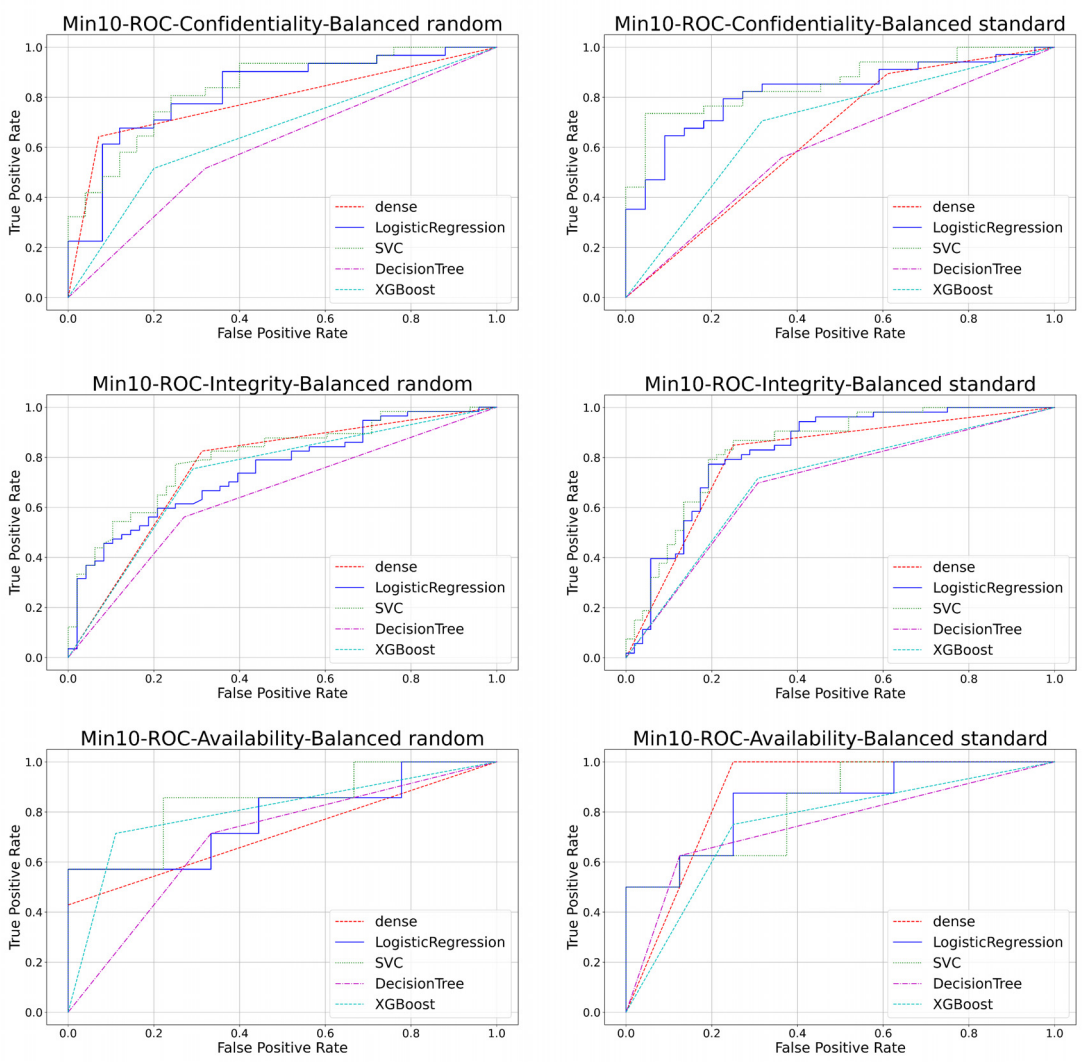
ROC 곡선은 불균형적인 데이터셋에는 적합하지 않으므로, 논문에서는 'balanced random', 'balanced standard' 샘플링만 살펴본다. 그림 5와 6은 각각 TTP 개수가 최소 5개, 10개인 보고서를 대상으로 한 실험의 ROC 곡선 그래프 중 하나이다. LogisticRegression, SVM이 준수한 성능을 보이며, 모델 초기화에 따라 FCN이 가장 좋은 성능을 보이는 경우도 있다. 이는 표 7의 결과와 일치한다.

(표 7) 모델별 정확도 및 F1 점수
(Table 7) Model Accuracy and F1 score

Label	Minimum TTP	Model	Random all		Standard all		Balanced random		Balanced standard	
			Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
Confidentiality	5	LogRes	0.8708	0.4173	0.8134	0.4658	0.7483	0.7387	0.7414	0.7867
		SVM	0.8794	0.3650	0.7962	0.2021	0.7552	0.7626	0.7966	0.8401
		DeciTree	0.8392	0.4028	0.7809	0.3880	0.6483	0.6458	0.6000	0.6427
		XGBoost	0.8679	0.4057	0.7895	0.2903	0.7069	0.7160	0.6828	0.7264
		FCN	0.8612	0.4627	0.7751	0.4367	0.6897	0.6801	0.6034	0.6005
	10	LogRes	0.8161	0.2862	0.7702	0.4638	0.7000	0.6736	0.7143	0.7600
		SVM	0.8360	0.2687	0.7391	0.1527	0.7179	0.7103	0.7429	0.7910
		DeciTree	0.7814	0.2872	0.7416	0.4125	0.6357	0.6050	0.6321	0.6781
		XGBoost	0.8224	0.3330	0.7391	0.3000	0.6786	0.6552	0.6321	0.6783
		FCN	0.8261	0.4524	0.7205	0.4188	0.7321	0.6510	0.6510	0.6250
Integrity	5	LogRes	0.8019	0.5902	0.8182	06275	0.7911	0.7857	0.7946	0.7892
		SVM	0.7971	0.5551	0.7943	0.5905	0.8071	0.8062	0.8054	0.8054
		DeciTree	0.7368	0.5043	0.7397	0.5533	0.6982	0.6893	0.7161	0.7152
		XGBoost	0.7971	0.5977	0.7799	0.5660	0.7839	0.7865	0.7804	0.7765
		FCN	0.7225	0.4194	0.7799	0.6301	0.8036	0.7984	0.7679	0.7602
	10	LogRes	0.7416	0.5907	0.7888	0.6600	0.7105	0.7175	0.7638	0.7535
		SVM	0.7714	0.6093	0.7665	0.6269	0.7429	0.7522	0.7752	0.7727
		DeciTree	0.7043	0.5617	0.6882	0.5358	0.6648	0.6726	0.6876	0.6807
		XGBoost	0.7429	0.5942	0.7764	0.6471	0.7200	0.7380	0.7429	0.7328
		FCN	0.6708	0.4015	0.7267	0.6652	0.7429	0.7311	0.7810	0.7809
Availability	5	LogRes	0.9493	0.1538	0.9426	0.0000	0.7300	0.6981	0.7600	0.7336
		SVM	0.9569	0.2009	0.9493	0.0000	0.8000	0.7817	0.7700	0.7434
		DeciTree	0.9129	0.2450	0.9196	0.2498	0.5800	0.5566	0.5500	0.5602
		XGBoost	0.9483	0.2674	0.9426	0.1429	0.7000	0.6825	0.6700	0.6278
		FCN	0.9522	0.4877	0.9522	0.4877	0.6500	0.6267	0.7000	0.7000
	10	LogRes	0.9429	0.1091	0.9689	0.4444	0.7250	0.7196	0.6750	0.6817
		SVM	0.9516	0.2238	0.9565	0.0000	0.6875	0.6809	0.7250	0.6996
		DeciTree	0.9230	0.3225	0.8969	0.1787	0.7375	0.7385	0.7250	0.7004
		XGBoost	0.9516	0.2984	0.9379	0.0000	0.6875	0.6797	0.6875	0.6726
		FCN	0.9503	0.4873	0.9565	0.4889	0.7500	0.7333	0.6875	0.6863



(그림 5) Min5 ROC 곡선 그래프
(Figure 5) Min5 ROC Curve Graph



(그림 6) Min10 ROC 곡선 그래프
(Figure 6) Min10 ROC Curve Graph

6. 결 론

논문에서는 공격자의 목표를 효율적으로 파악하기 위해 새로운 라벨링 방법을 정의하고, 머신러닝 모델과 딥러닝 모델을 통해 그 유효성을 검증하였다. 실험결과에 따르면, 샘플링 과정을 거쳐 평균 70%, 최고 80% 정답율의 성능을 보였다. 본 논문에서 제시된 라벨링 기법은 초기모델이므로 각 라벨에 대한 이진 분류 실험 정확도의 개선여지가 있으나, MITRE ATT&CK에 기반을 두어 CIA 관점에서 라벨링을 재정의한 첫 번째 시도임에 의의가 있다.

다만 논문에서는 'rcATT' 도구를 이용하여 문서 라벨링을 자동화하여 실험에 사용하였기 때문에 모델의 성능이 라벨링 도구에 의존적이게 된다. 따라서 후속 연구로는 라벨링 도구 성능 향상을 통한 데이터셋 신뢰도 보장과 추가적인 데이터셋 확보를 통한 모델 성능 향상 등에 집중할 계획이다.

참고문헌(Reference)

[1] Cybersecurity and Infrastructure Security Agency(CISA), “North Korean State-Sponsored Cyber Actors Use Maui Ransomware to Target the Healthcare and Public Health Sector”, 2022. https://cisa.gov/uscert/ncas/alerts/aa22-187a

[2] D. Kushner, “The real story of stuxnet”, IEEE Spectrum, Vol.50, No.3, pp.48-53, 2013. https://doi.org/10.1109/MSPEC.2013.6471059

[3] N. Falliere, L. Murchu, and E. Chien, “W32.Stuxnet Dossier”, 2011. https://docs.broadcom.com/doc/security-response-w32-stuxnet-dossier-11-en

[4] Mandiant, “APT41(Double dragon): A Dual Espionage and Cyber crime Operation”, 2022. https://www.mandiant.com/sites/default/files/2022-02/rt-apt41-dual-operation.pdf

[5] Faou, M., Tartare, M., Dupuy, T., “OPERATION GHOST. The Dukes aren’t back – they never left”, 2019. https://welivesecurity.com/wp-content/uploads/2019/10/ESET_Operation_Ghost_Dukes.pdf

[6] Federal Bureau of Investigation(FBI), “Update on Sony Investigation)”, https://www.fbi.gov/news/press-releases/update-on-sony-investigation

[7] C. Fan, H. Hsiao, C. Chou and Y. Tseng, “Malware Detection Systems Based on API Log Data Mining”, IEEE 9th Annual Computer Software and Applications Conference, pp. 255-260, 2015. https://doi.org/10.1109/COMPSAC.2015.241

[8] D. Uppal, R. Sinha, V. Mehra and V. Jain, “Exploring Behavioral Aspects of API Calls for Malware Identification and Categorization”, International Conference on Computational Intelligence and Communications Networks, pp.824-828, 2014. https://doi.org/10.1109/CICN.2014.176

[9] Choi, C. H., Lee, H. S., Jung, I. H., Yoo, C. G., and Yoon, H. S., “Statistical Analysis of EML Header for Cyber Attacker Tracing”, Proceedings of Korea Institute of Military Science and Technology annual conference, pp.1141-1142, 2017.

[10] Choi, C. H., Lee, H. S., Jung, I. H., Park, J. H., and Yoon, H. S.,“E-mail Clustering for Cyber Attack Attribution”, Proceedings of Korea Institute of Military Science and Technology annual conference, pp.1289-1290, 2018.

[11] N. Villeneuve, J. Bennett, “Detecting APT Activity with Network Traffic Analysis”, 2012. https://documents.trendmicro.com/assets/wp/wp-detecting-apt-activity-with-network-traffic-analysis.pdf

[12] G. Zhao, K. Xu, L. Xu and B. Wu, “Detecting APT Malware Infections Based on Malicious DNS and Traffic Analysis”, IEEE Access, Vol.3, pp.1132-1142, 2015. https://doi.org/10.1109/ACCESS.2015.2458581

[13] MITRE ATT&CK®, https://attack.mitre.org

[14] International Organization for Standardization(ISO), “Information technology – Security techniques – Information security management systems – Overview and vocabulary(ISO Standard No.27000:2009)”, 2009. https://www.iso.org/standard/41933.html

[15] Kaspersky, “LAZARUS UNDER THE HOOD”, 2018. https://media.kasperskycontenthub.com/wp-content/uploads/sites/43/2018/03/07180244/Lazarus_Under_The_Hood_PDF_final.pdf

[16] Securelist by Kaspersky, “OlympicDestroyer is here to trick the industry”, 2018. https://securelist.com/olympicdestroyer-is-here-to-trick-the-industry/84295/

[17] S. M. Park, J. I. Lim, “Study On Identifying cyberattack Classification Through The Analysis of cyberattack Intention”, Journal of the Korea Institute of Information Security & Cryptology, Vol.27, No.1, pp.103-113, 2017. https://doi.org/10.13089/JKIISC.2017.27.1.103

[18] Shin C. H., Shin S. U., Seo, S. Y., Lee, I. S., Choi, C. H., “Embedding and Training RNN for estimating the goal of cyberattack”, Proceedings of Korea Institute of Military Science and Technology annual conference, pp.1055-1056, 2021.

[19] Bitdefender, StrongPity APT – Revealing Trojanized Tools, Working hours and Infrastructure, https://bitdefender.com/files/News/CaseStudies/study/353/Bitdefender-Whitepaper-StrongPity-APT.pdf

[20] Ahnlab Security Emergency response Center(ASEC), “Analysis Report of Kimsuky Group’s APT Attacks

(AppleSeed, PebbleDash)”, http://download.ahnlab.com/global/brochure/Analysis%20Report%20of%20Kimsuky%20Group.pdf

[21] C. Beek, “Operation ‘Harvest’: A Deep Dive into a Long-term Campaign”, https://www.trellix.com/en-us/about/newsroom/stories/threa-labs/operation-harvest-a-deep-dive-into-a-long-term-campaign.html

[22] APT&CyberCriminal Campaign Collections, https://github.com/CyberMonitor/APT_CyberCriminal_Campaign_Collections

[23] Legoy, V., Caselli, M., Seifert, C., and Peter, A, “Automated retrieval of ATT&CK tactics and techniques for cyber threat reports.”, arXiv preprint arXiv:2004.14322, 2020.

[24] rcATT, https://github.com/vlegoy/rcATT

[25] Pedregosa et al., “Scikit-learn: Machine Learning in Python”, Journal of Machine Learning Research, Vol.12, pp. 2825-2830, 2011. https://dl.acm.org/doi/10.5555/1953048.2078195

[26] Chen, Tianqi and Guestrin, Carlos, “XGBoost: A Scalable Tree Boosting System”, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data mining, pp.785-794, 2016. https://doi.org/10.1145/2939672.2939785

◐ 저 자 소 개 ◑

신 찬 호(ChanHo Shin)
2018년 고려대학교 사이버국방학과(학사)
2018년~현재 국방과학연구소 현역연구원
관심분야 : 정보보호, 인공지능
E-mail : shinch2018@add.re.kr

최 창 희(Chang-hee Choi)
2008년 연세대학교 컴퓨터과학과(공학사)
2010년 한국과학기술원 대학원 전산학과(공학석사)
2013년 한국과학기술원 대학원 전산학과(공학박사)
2013년~현재 국방과학연구소 연구원
관심분야 : 머신러닝 기반 사이버 보안, AI, GAN, 디지털 포렌식
E-mail : changhee84@add.re.kr