

# 사이버보안 AI(악성코드/침해사고 분야) 데이터셋 소개

## 1. 악성코드 분야

### ① 백신 진단명 기반의 악성코드 AI 데이터셋

- (정의) 정상/악성 파일에 대한 백신 진단명 기반 위협유형(Threat Type)을 참고하여 AI 모델 학습용 라벨(labeling)을 부여한 데이터셋
- (파일유형) 윈도우, 리눅스, 모바일에서 동작하는 파일 유형 25종
  - ※ exe, dll, apk, elf, gzip, zip, rar, 7zip, cab, javascript, java, vbs php, powershell, html, pdf, doc(x), xls(x), jpg, ppt(x), ppsx, hwp
- (제공 데이터) 바이너리 정적·동적 및 가공 분석을 통해 추출되는 정보
  - 바이너리 특징정보\*가 포함된 메타데이터(JSON 포맷)
    - \* 파일 기본정보(Hash, Type, Size 등), 파일 헤더, APK Activity, 바이너리 n-gram 패턴, OP-code 카운트 및 API Call 시퀀스 등
  - 바이너리 이미지 변환 파일(BMP 포맷)

### ② 최신 위협 이슈 연관 악성코드 데이터셋

- (정의) 최신 국내외 사이버 위협 분석 보고서에 포함된 악성코드 대상 주요 이슈 키워드 기반 AI 모델 학습용 라벨(labeling)을 부여한 데이터셋
- (파일유형) 윈도우, 리눅스, 모바일에서 동작하는 파일 유형 25종
  - ※ exe, dll, apk, elf, gzip, zip, rar, 7zip, cab, javascript, java, vbs php, powershell, html, pdf, doc(x), xls(x), jpg, ppt(x), ppsx, hwp
- (제공 데이터) 바이너리 정적·동적 및 가공 분석을 통해 추출되는 정보
  - 바이너리 특징정보\*가 포함된 메타데이터(JSON 포맷)
    - \* 파일 기본정보(Hash, Type, Size 등), 파일 헤더, APK Activity, 바이너리 n-gram 패턴, OP-code 카운트 및 API Call 시퀀스 등

## 2. 침해사고 분야

### ① 사이버 공격 전술(MITRE ATT&CK Framework) 기반 AI 데이터셋

- (정의) MITRE ATT&CK Framework의 T-ID를 기반으로 AI 모델 학습용 라벨(labeling)을 부여한 데이터셋
- (카테고리) ATT&CK Framework 공격 전술 230종
- (제공 데이터) T-ID 탐지를 매칭 결과 및 악성코드 자동 분석 플랫폼을 통해 추출되는 행위분석 정보
  - T-ID 및 공격 행위 정보가 포함된 메타데이터(JSON 포맷)
    - \* TID 탐지 규칙 매칭 정보 등 T-ID 매칭 결과
  - Sandbox 행위 분석 정보(JSON 포맷)
    - \* EVTX, PCAP, MEMDUMP 등 포함

### ② 최신 사회 이슈 관련 사이버 침해사고 재현 데이터셋

- (정의) 최신 사이버 침해사고 시나리오 탐지 및 대응을 위해 AI 모델 학습용 라벨(labeling)을 부여한 데이터셋
- (파일유형) 보안장비 6종(FW, IPS, IDS, WAF, SYSLOG, WEB 등) 로그 데이터
- (제공 데이터) 침해사고 재현을 통해 수집된 보안 로그 대상 분석·가공된 정보 및 시나리오별 Playbook
  - 공격 기법(T-ID) 및 정상/악성 여부가 포함된 메타데이터(CSV, XLSX 포맷)
    - \* 출발지/목적지 IP 및 Port, 탐지명, 패킷 사이즈, http 정보(header, method, query, version 등), 페이로드, 프로토콜, 공격여부, T-ID, 시나리오명 등

#### □ [예시1] 백신 진단명 기반의 악성코드 AI 데이터셋 제공 리스트

#### □ [예시2] 최신 위협 이슈 연관 악성코드 데이터셋 제공 리스트

[illegible]

## □ [예시3] 이용 신청자 제공 데이터셋 세부 예시

Set-ID	AI-Label	Platform	Threat Type	File Type	데이터 개수	데이터 크기	신청 건수	조회수
A-24534	exe-ransomware	Windows	ransomware	exe	10,000	14.05 GB	1	1

#	악성 / 정상	분류 기준	Threat Type	File Type	File Size (bytes)	MD5	SHA1	SHA256
10000	악성	위험 유형	ransomware	exe_32bit	1,702,409	FC9981F4C5A...	7D66C476A0B402...	006220177268144D92017830F43...
9999	악성	위험 유형	ransomware	exe_32bit	1,716,224	A3306A4542...	3E90F09032...	2DA9FADF0E0E3132EF90A...
9998	악성	위험 유형	ransomware	exe_32bit	1,419,348	B98378036D...	F4322058BC...	000839E1568C6A182208CF6E19...
9997	악성	위험 유형	ransomware	exe_32bit	1,377,280	A0EF005C09...	3A48026D0A8C...	00089532B9FD663AC2D2DD85E...
9996	악성	위험 유형	ransomware	exe_32bit	3,723,264	B7F9ED107B...	C98824502FC...	00090345E93826EE2820C669F9E...
9995	악성	위험 유형	ransomware	exe_32bit	1,707,917	D6C918B6A5...	027FR35998D...	000FR11D5888F397A717781D6...
9994	악성	위험 유형	ransomware	exe_32bit	1,408,000	D71C59568A...	FAD82D24830...	0019EB3C845AD07A2BAD27AEB...
9993	악성	위험 유형	ransomware	exe_32bit	6,954,496	EAGE114C5D...	CTEFC037608...	00296908626406001019CD84F5C...
9992	악성	위험 유형	ransomware	exe_32bit	1,712,128	A7A4B8C187...	DAA11E67B93...	0029BCFA273DDA5CF089185B13...
9991	악성	위험 유형	ransomware	exe_32bit	5,373,440	B1367CE868...	E56D48325D0...	716D852B1678F13BF...

pe JSON 파일 313KB	pa_binary_grayscaleimage BMP 파일 82.0KB	pa_binary_wem BMP 파일 2.05KB	pa_binary_fingerprintcode BMP 파일 1.30KB
------------------------	--	-----------------------------------	---

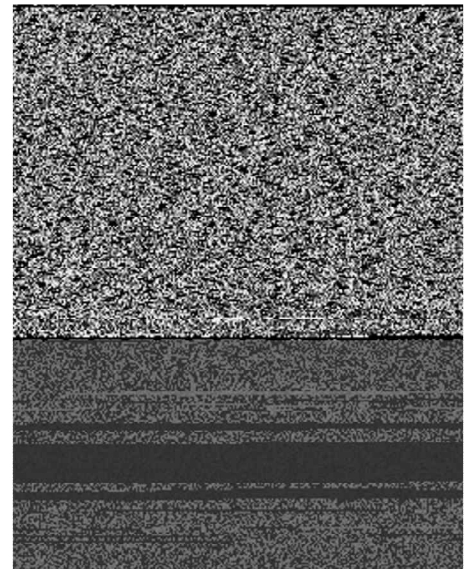
<메타데이터>

<바이너리 이미지 변환 데이터>

```

"md5": "80dbc86ce4131...-f6191061aec",
"sha1": "0786c50a59a93...7aeb7b04235a71718aa",
"sha256": "021e02f41de...b47cb168912e2298285c4be3bc49f8facffcc39af57994",
"file_name": [
  "WWW1we472k.exe",
  "021e02f41dec909d_1we472k.exe"
],
"file_type": "Win32 EXE",
"mime_type": "application/octet-stream",
"file_size": 82827,
"av_detection": {
  "A": "Gen:Variant.Ulise.42647",
  "B": "Trojan Horse",
  "C": "Gen:Variant.Ulise.42647",
  "D": "HEUR:Backdoor.Win32.Tiny.gen"
},
"pe_header_fileinfo_item_number": 0,
"pe_header_timestamp": "2021/10/08 04:56:27",
"pe_header_size": 512,
"pe_header_sectionsize_export": 0,
"pe_header_sectionsize_import": 60,
"pe_header_sectionsize_resource": 0,
"pe_header_section_number": 2,
"pe_header_baseofdata": 155648,
"pe_header_checksum": 0,
"pe_header_dll_importnumber": 2,
"pe_header_emaxalloc": 65535,
"pe_header_ecblp": 144,
"pe_header_ecp": 3,
"pe_header_ecparhdr": 4,
"pe_header_elfanew": 16,
"pe_header_esp": 17744,
"pe_header_entrypoint": 51664,
"pe_header_filealignment": 4096,
  
```

<메타데이터 상세>



<바이너리 이미지 변환 데이터 상세>

## 참고 2 침해사고 분야 데이터셋 제공 예시

### □ [예시1] 사이버 공격 전술 기반 AI 데이터셋 제공 리스트

전체선택									
<input type="checkbox"/>	[T1599.001]	Network Boundary Bridging: Network Address Translation Traversal	Sets ID: 601305	데이터 개수: 31	데이터 크기: 311.09 KB	신용도: 0	데이터 출처: 1	데이터 출처: 1	데이터 출처: 1
<input type="checkbox"/>	[T1590]	Gather Victim Network Information	Sets ID: 60154	데이터 개수: 15,062	데이터 크기: 124 MB	신용도: 0	데이터 출처: 1	데이터 출처: 1	데이터 출처: 1
<input type="checkbox"/>	[T1588.002]	Obtain Capabilities: Tool	Sets ID: 60153	데이터 개수: 3,188	데이터 크기: 9.75 MB	신용도: 0	데이터 출처: 1	데이터 출처: 1	데이터 출처: 1
<input type="checkbox"/>	[T1588]	Obtain Capabilities	Sets ID: 60162	데이터 개수: 36	데이터 크기: 272.73 KB	신용도: 0	데이터 출처: 1	데이터 출처: 1	데이터 출처: 1
<input type="checkbox"/>	[T1587]	Develop Capabilities	Sets ID: 60163	데이터 개수: 111	데이터 크기: 5.23 MB	신용도: 0	데이터 출처: 1	데이터 출처: 1	데이터 출처: 1
<input type="checkbox"/>	[T1574.011]	Hijack Execution Flow: Services Registry Permissions Weakness	Sets ID: 60160	데이터 개수: 11	데이터 크기: 191.83 KB	신용도: 0	데이터 출처: 1	데이터 출처: 1	데이터 출처: 1
<input type="checkbox"/>	[T1574.002]	Hijack Execution Flow: DLL Side-Loading	Sets ID: 60159	데이터 개수: 5,775	데이터 크기: 26.38 MB	신용도: 0	데이터 출처: 1	데이터 출처: 1	데이터 출처: 1
<input type="checkbox"/>	[T1574.001]	Hijack Execution Flow: DLL Search Order Hijacking	Sets ID: 60158	데이터 개수: 1,006	데이터 크기: 5.86 MB	신용도: 0	데이터 출처: 1	데이터 출처: 1	데이터 출처: 1
<input type="checkbox"/>	[T1574]	Hijack Execution Flow	Sets ID: 60157	데이터 개수: 2,333	데이터 크기: 28.3 MB	신용도: 0	데이터 출처: 1	데이터 출처: 1	데이터 출처: 1
<input type="checkbox"/>	[T1572]	Protocol Tunneling	Sets ID: 60156	데이터 개수: 201	데이터 크기: 1.07 MB	신용도: 0	데이터 출처: 1	데이터 출처: 1	데이터 출처: 1

### □ [예시2] 최신 사회 이슈 관련 사이버 침해사고 재현 데이터셋 제공 리스트

사라지오 #1	
사라지오 #2	
사라지오 #3	
사라지오 #4	
사라지오 #5	
사라지오 #6	
사라지오 #7	
사라지오 #8	
사라지오 #9	
사라지오 #10	
사라지오 #11	
사라지오 #12	
사라지오 #13	
사라지오 #14	
사라지오 #15	

1. CED5\_01\_플레이어를 통한 내부망 침략

1) 생성 결과 요약

구분	내용
사라지오 번호	CED5_01
사라지오 명	플레이어를 통한 내부망 침략
커맨드	Webshell, Scheduled Task, Data Exfiltration
사라지오 목적	웹 서비스를 제공하는 서버 관련 탈취 및 내부 시스템 침략 사나리오
사라지오 절차	1. 기존 수집된 정보로 탈취 2. 탈취 완료 3. 사나리오를 수행할 목적으로 내부망 4. 악성코드 실행 5. 공격 시작 6. 결과 수집

원시 데이터 (RAW)

2,206,183 건

원본 데이터 (SRC)

2,206,183 건

가장 데이터 (DRI)

2,206,183 건 (Attack: 121,363 건, Normal: 2,084,820 건)

학습 데이터 (ML)

Training: 1,764,542 건

Validation: 220,624 건

Testing: 220,617 건

AI 라벨: 침해사고 시나리오

T-ID

공격 여부

시나리오

사라지오 실행 도구세트

간혹

대용량 Playbook 다운로드

2) 세부 데이터 수

	구분	원시 데이터	원본 데이터	가장 데이터		학습 데이터		시험
				attack	normal	학습 (80%)	검증 (10%)	
<input type="checkbox"/>	FW	2,141,475	2,141,475	60,349	2,081,126	1,713,179	214,949	214,947
<input type="checkbox"/>	IPS	2,256	2,256	1,744	512	1,804	226	226
<input type="checkbox"/>	IDS	36,716	36,716	36,026	690	29,372	3,671	3,671
<input type="checkbox"/>	WAF	2,411	2,411	1,723	688	1,808	242	241
<input type="checkbox"/>	WEB	11,126	11,126	11,031	115	8,900	1,131	1,131
<input type="checkbox"/>	서버로그	11,199	11,199	10,410	1,789	9,759	1,231	1,219
<input type="checkbox"/>	합계	2,206,183	2,206,183	121,363	2,084,820	1,764,542	220,624	220,617

날 연재하기



전체검색					
<input type="checkbox"/>	[T1003.005] OS Credential Dumping: Cached Domain Credentials	Set ID : 60005	데이터 크기 : 533	데이터 크기 : 3.89 MB	신뢰수 : 0
<input type="checkbox"/>	[T1003.004] OS Credential Dumping: LSA Secrets	Set ID : 60004	데이터 크기 : 577	데이터 크기 : 4.04 MB	신뢰수 : 0
<input type="checkbox"/>	[T1003.003] OS Credential Dumping: NTDS	Set ID : 60003	데이터 크기 : 577	데이터 크기 : 1.04 MB	신뢰수 : 0
<input type="checkbox"/>	[T1003.002] OS Credential Dumping: Security Account Manager	Set ID : 60002	데이터 크기 : 577	데이터 크기 : 7.33 MB	신뢰수 : 0
<input type="checkbox"/>	[T1003.001] OS Credential Dumping: LSASS Memory	Set ID : 60001	데이터 크기 : 577	데이터 크기 : 15.34 MB	신뢰수 : 0
<input type="checkbox"/>	[T1003] OS Credential Dumping	Set ID : 60000	데이터 크기 : 577	데이터 크기 : 18.01 MB	신뢰수 : 0



### 〈행위 분석 데이터〉

### 〈메타데이터 상세〉

### 〈행위 분석 데이터 상세〉