사이버보안 AI(악성코드/침해사고 분야) 데이터셋 소개

1. 악성코드 분야

① 백신 진단명 기반의 악성코드 AI 데이터셋

o (정의) 정상/악성 파일에 대한 백신 진단명 기반 위협유형(Threat Type)을 참고하여 AI 모델 학습용 라벨(labeling)을 부여한 데이터셋

o (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

o (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

o (정의) 최신 국내외 사이버 위협 분석 보고서에 포함된 악성코드 대상 주요 이슈 키워드 기반 AI 모델 학습용 라벨(labeling)을 부여한 데이터셋

o (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

o (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

o (정의) MITRE ATT&CK Framework의 T-ID를 기반으로 AI 모델 학습용 라벨(labeling)을 부여한 데이터셋

o (카테고리) ATT&CK Framework 공격 전술 230종

o (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

o (정의) 최신 사이버 침해사고 시나리오 탐지 및 대응을 위해 AI 모델 학습용 라벨(labeling)을 부여한 데이터셋

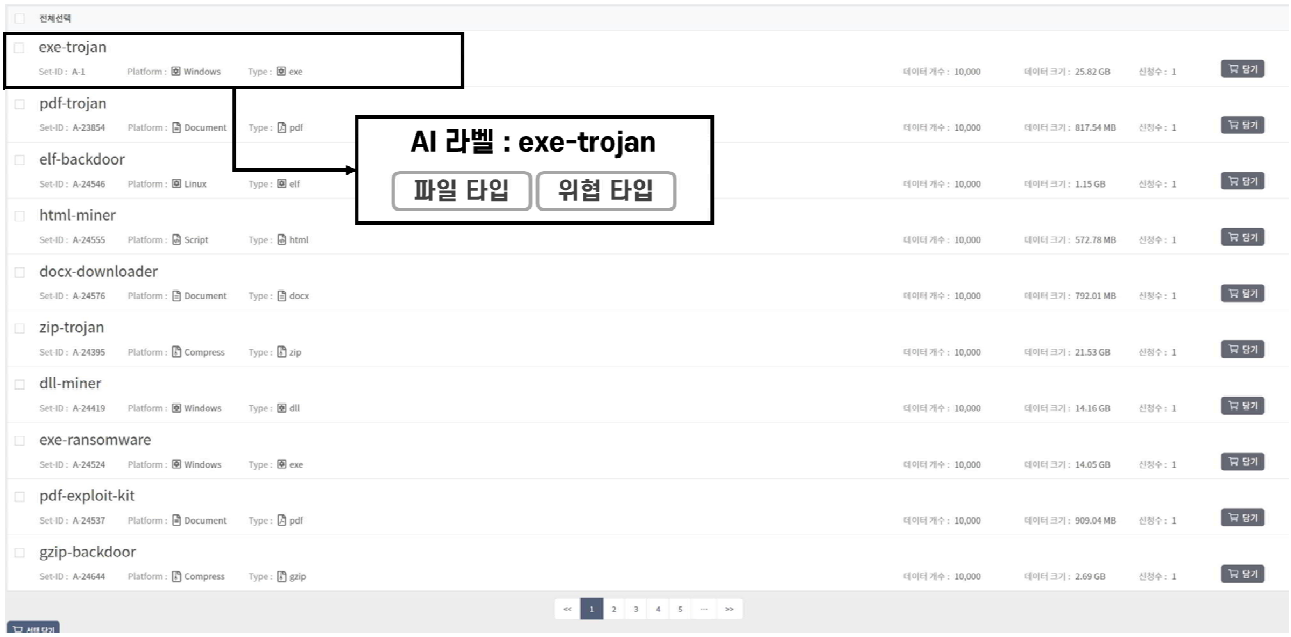
o (파일유형) 보안장비 6종(FW, IPS, IDS, WAF, SYSLOG, WEB 등) 로그 데이터

o (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

참고 1 악성코드 분야 데이터셋 제공 예시

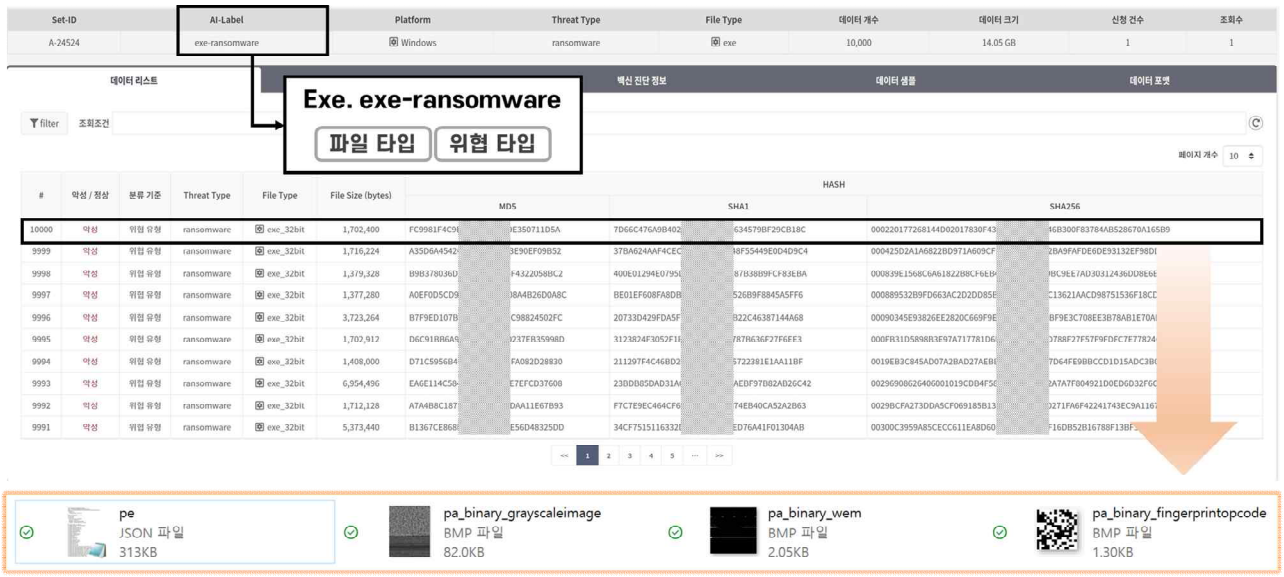
□ (예시1) 백신 진단명 기반의 악성코드 AI 데이터셋 제공 리스트



□ (예시2) 최신 위협 이슈 연관 악성코드 데이터셋 제공 리스트

※ multi-label 부여

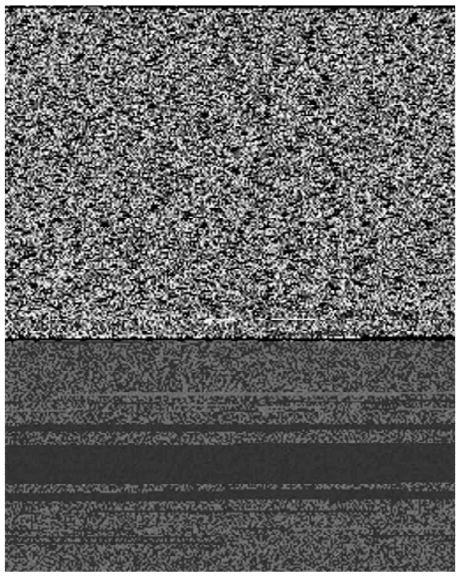
□ (예시3) 이용 신청자 제공 데이터셋 세부 예시



<메타데이터>

<바이너리 이미지 변환 데이터>


```
"md5": "80dbc86ce4131[redacted]cf6191061aec",
"sha1": "0786c50a59a93[redacted]7aebe7b04235a71718aa",
"sha256": "021e02f41de[redacted]b47cb168912e2298285c4be3bc49f8facfffc39af57994",
"file_name": [
    "₩₩1we472k.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,
```

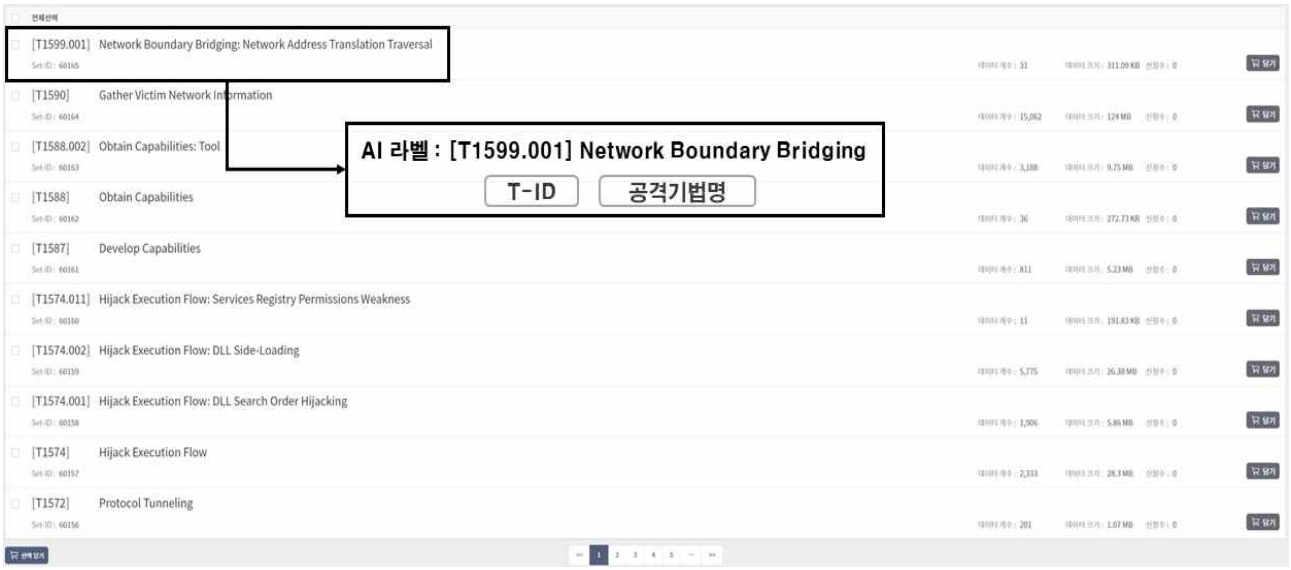



<메타데이터 상세>

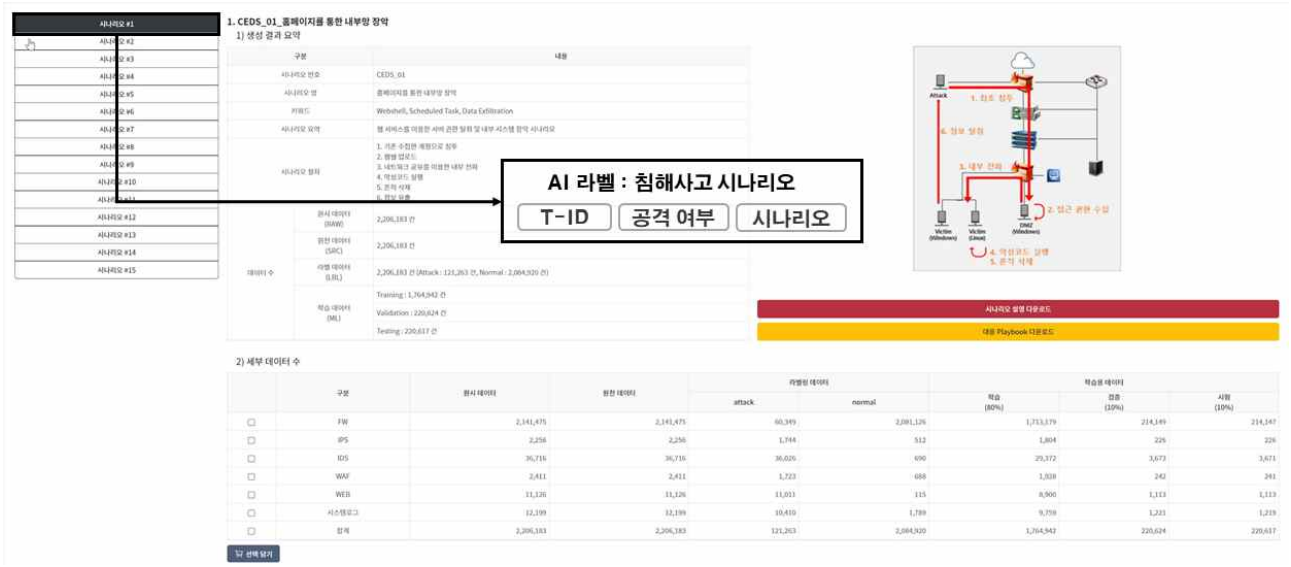
<바이너리 이미지 변환 데이터 상세>

참고 2 침해사고 분야 데이터셋 제공 예시

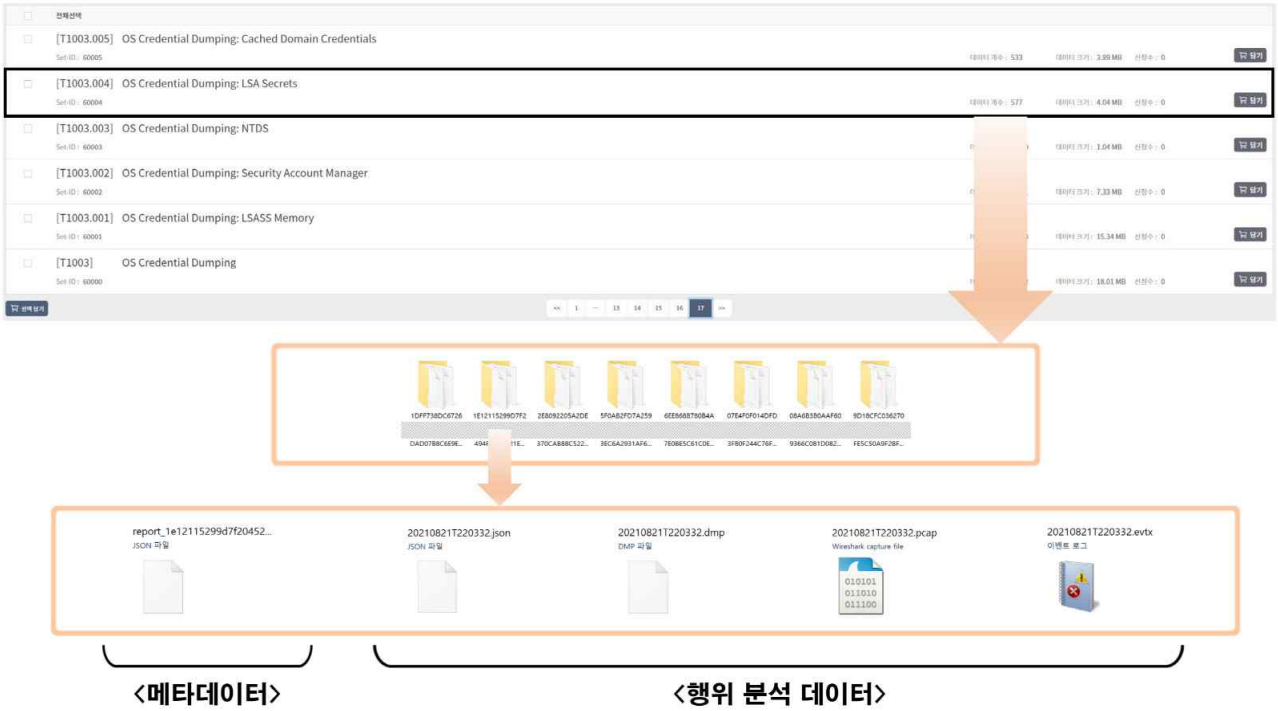
□ (예시1) 사이버 공격 전술 기반 AI 데이터셋 제공 리스트



□ (예시2) 최신 사회 이슈 관련 사이버 침해사고 재현 데이터셋 제공 리스트



□ (예시3) 이용 신청자 제공 데이터셋 세부 예시



<메타데이터 상세>

<행위 분석 데이터 상세>