



4 - 6 October, 2023 / London, United Kingdom

BUILDING A CYBER SECURITY AI DATASET FOR A SECURE DIGITAL SOCIETY

Bomin Choi, Juhyuk Kim & Hoseok Ryu

Korean Internet and Security Agency, Republic of Korea

bmchoi@kisa.or.kr

juhyukkim@kisa.or.kr

hsryu@kisa.or.kr

ABSTRACT

The Cyber Security Big Data Center in KISA (Korea Internet and Security Agency)¹ has been carrying out a large-scale cybersecurity AI dataset construction project based on cooperation with leading companies and institutions in Korea since 2021. AI modelling requires a large amount of high-quality learning data, but it is difficult for private companies to acquire it. Accordingly, we have been carrying out this project to bridge the technology gap due to a dearth of data by establishing and opening a cybersecurity AI dataset at the government level.

We conduct demand surveys and in-depth interviews with many experts in the field of cybersecurity every year, and through this, we have been selecting and expanding areas where the dataset will be built according to the urgency and utilization of demand-based dataset construction. By establishing and sharing various types of cybersecurity AI datasets that can be used throughout the entire lifecycle of breach response, we are striving to lay the groundwork for creating a safe digital society based on intelligent security technology. Our project consists of building an AI dataset by collecting, analysing, processing and labelling raw data such as malware, infringement incidents, and indicators of compromise (IoCs). We are not just building a dataset of two-dimensional labels that simply identify ‘normal or malicious’, but we are building a dataset through labelling based on keywords that are effective in responding to cyber threat issues that target our lives, e.g. social issues related to the latest infringements (the Covid-19 pandemic, Russia-Ukraine war, political or diplomatic conflicts, etc.), attack groups, campaigns, and TTP information.

The aim is to utilize the dataset to train an AI model that actively addresses various security threats faced by AI models in the real world. The dataset is verified to ensure it is effective through pilot application to domestic IT service companies or institutions, and various best practices are selected and disseminated to the private sector so that the dataset can be used more actively. In addition, it is expected that this project will contribute to creating a socially safe digital environment. Through this paper, we want to share the story of the trials and errors we have experienced and the know-how we have gained during the cybersecurity AI dataset construction project over the past two years, and create an opportunity to realize a safe digital society through cooperation with global related companies and institutions.

1. INTRODUCTION: WHAT IS THE AI DATASET FOR CYBERSECURITY?

The definition of an AI dataset may vary depending on the AI learning methodology (e.g. supervised learning, unsupervised learning, etc.). However, our definition is a set of labelled individual data with informational power for an AI model to learn to create knowledge for solving problems at hand. Here, ‘data, information and knowledge’ are related but different concepts. A classic distinction is made in McDonough’s *Information Economics* (1963). Among the concepts mentioned here, ‘data’ is defined as a message whose value is not assessed, ‘information’ as data evaluated in a specific situation, and ‘knowledge’ as a relationship between time and content, which is broader than information [1].

When applying these concepts to cybersecurity, we can define the term ‘AI dataset for cybersecurity’ as a collection of data designed to train AI models in resolving various cyber threats that occur in the digital environment. In other words, individual data lacks the ability to generate contextual meaning, thus we rely on datasets as the key ingredient when generating information. Therefore, within this concept, we can define AI dataset for cybersecurity as a curated collection of labelled data, aimed at training AI machines to tackle cyber threat issues at the individual data level.

2. BACKGROUND AND MOTIVATION

KISA opened the Cyber Security Big Data Center within KrCERT/CC in December 2018. Its purpose is to share various kinds of threat information, collected on the basis of cooperation with domestic and foreign companies and institutions, with the private sector, thereby strengthening the private companies’ infringement response capabilities and supporting the establishment of an intelligent infringement response system through the creation of a cycle of quality data.

When the Cyber Security Big Data Center was opened in 2018, the amount of data held was about 190 million pieces of IoC data with the purpose of simple detection/blocking. Through the cybersecurity AI dataset construction project that started in 2021, however, we have expanded and refined various types of threat information, e.g. infringement incidents, malicious files, and vulnerable source codes with not only IoCs but also various AI labels (attack groups, threat types, TTPs, CVE-IDs, etc.) identified to about 1.4 billion pieces, and we are opening an abundant level of data to the private sector. Through this, we intend not only to detect and block IoCs (such as malware, malicious IPs/URLs), but also to enable automated explanations of complex threat situations and facilitate the intelligence-driven enhancement of intrusion response systems, including timely prediction and prevention of future threats.

The following explains the purpose of our project from economic, technological and social perspectives.

- **Economic:** Resolving the issue of cost investment for AI dataset acquisition experienced by the private sector.

¹ KISA is a government agency of the Republic of Korea that performs a role similar to CISA in the United States, focusing on enhancing cybersecurity and ensuring a safe digital environment. KISA operates KrCERT/CC, which responds to hacking and virus attacks on the private sector’s ICT infrastructure in the Republic of Korea. It continuously monitors the internet network and, upon detecting any signs of abnormalities, disseminates ‘incident response alerts’ to the private sector for blocking and taking appropriate countermeasures.

- **Technical:** Bridging the gap in intelligent security technology caused by data imbalance.
- **Social:** Enabling intelligent cyber threat response capabilities in the complex (multi-dimensional) real world based on multi-labelled datasets.

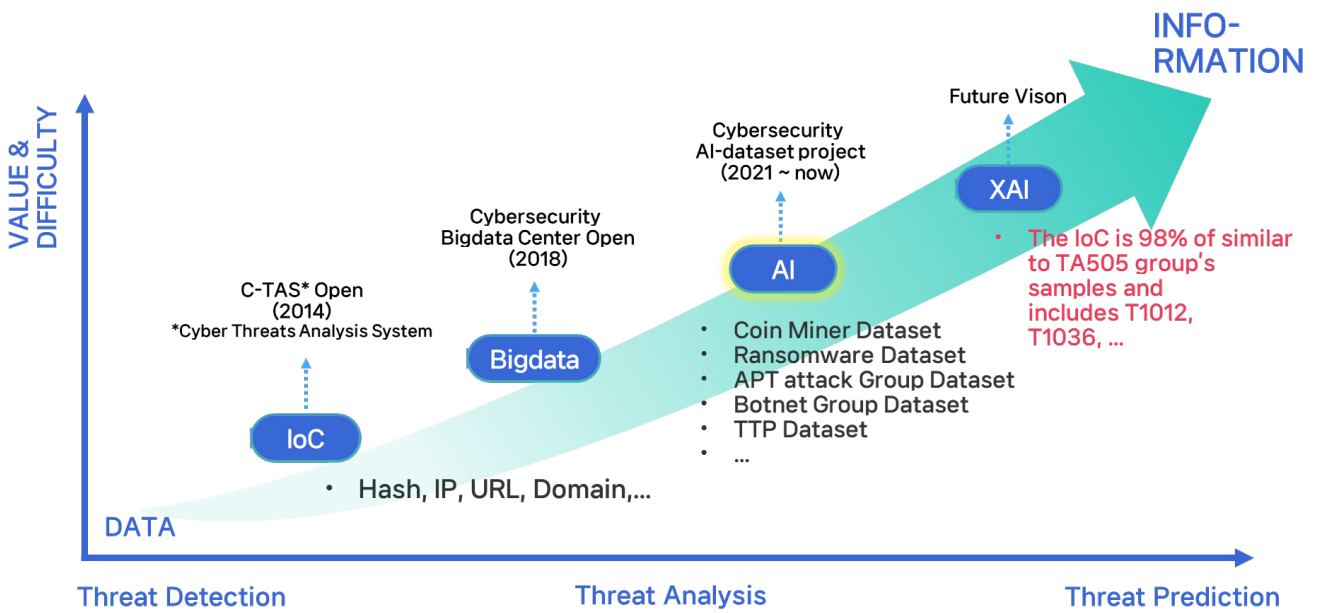


Figure 1: The strategy of KISA's dataset construction and sharing.

One of the most critical aspects in cyber incident response is prompt decision-making by security personnel, and to facilitate this, it is essential to have available timely and accurate information regarding the threat landscape. In particular, what we consider most important is the data labelling (annotation) work to realize intelligent cyber threat response capabilities in the complex (multi-dimensional) real world. Real cyber threats are no longer concerned solely with showing off their individual abilities or acquiring simple financial gains, as in the past. Recently, threats tend to be associated with multiple factors including political, social and economic issues and the pursuit of organizational profits [2]. Therefore, beyond simply detecting whether a file is legitimate or malicious, it is necessary to analyse the intention or purpose of files and the background of the attacks they are used in (attack groups, attacking countries, etc.), annotate the identified information, and use it in AI models to prevent further damages in advance.

Data 'labelling (annotation)' mostly requires analysis by experienced experts with extensive know-how. However, it is difficult for an individual analyst to accumulate knowledge and rich analysis experience in *all* cases of cyber threats, and there are limitations in terms of resources (time, cost, manpower, etc.) when a limited number of experts try to derive results by analysing the situations of the new types of cyber threats that are increasing day by day.

We intend to overcome the existing limitations by utilizing an AI model with strengths in complex operations. In other words, we intend to build a high-quality learning dataset (knowledge for AI machines) using the analysis know-how, knowledge and experience of individual experts who have rich experience in responding to incidents of each threat type, and make the AI model learn the dataset to simulate the human ability to analyse intelligence.

Related studies and dataset overview

There are a number of AI learning datasets that have already been released in the security field through previous research and projects. In particular, high-quality datasets such as Big-2015 (*Kaggle, Microsoft*) [3], EMBER (*Elastic*) [4] and SOREL-20 (*Sophos*) [5] are open to the public in the field of malware analysis where AI application is highly mature, and are therefore actively used in technology research and development in many schools, research institutes and security companies. However, since most of the existing datasets only identify normal/malicious or label threat types in a limited category, there are limits to generating intelligence information that can identify realistic attack intentions, purposes and attack groups in regard to the complex threats in the real world.

Against this backdrop, we have built a cybersecurity AI dataset containing more than 1.4 billion items for related malware, security logs, IoCs and CVE/CWE vulnerable source codes while tracking incidents that became social issues in the past. Table 1 shows the status and description of the AI dataset we have built. We are currently engaged in the construction of datasets in the fields of threat intelligence and threat hunting to enhance the explanatory and responsive capabilities of existing datasets.

Dataset	Year	Amount of data	Description	Format
Malware	2021	400 million records	The analysis of malicious and benign files (<i>Windows</i> , mobile, <i>Linux</i> ; e.g. EXE, DLL, PDF, APK, ZIP, JPEG, etc.) involves classification based on keywords such as family, threat type, and social issues	JSON, STIX
TTPs Simulation		400 million records	Various security appliance logs obtained through simulated attack scenarios such as supply chain attacks and spear phishing utilizing MITRE ATT&CK TTPs	JSON, CSV (IDS/IPS, WAF, FW, system logs)
CVE Vulnerable Code	2022	300 million records	Identified vulnerable source code and attack code with CVE/CWE for vulnerability assessment, detection, and analysis of applications developed with third-party software	Source code (C/C++, C#, Java, PHP, Python, C#)
Security Monitoring and Operation		200 million records	Security appliance logs obtained through reproduction of attack behaviour units Proactive attack response hunting rules and playbooks	JSON, CSV (IDS/IPS, WAF, FW, EDR, system logs)
Threat Profiling (APT/Botnet Group)		100 million records	Threat intelligence associated with APT attack groups and botnets, ransomware families operating domestically and internationally, including Kimsuky, Lazarus, APT29, etc.	JSON, STIX (IP, domain, URL, hash, files, etc.)

Table 1: Overview of KISA AI dataset construction.

3. FRAMEWORK AND METHODOLOGY

As mentioned earlier, we are trying to build adaptable AI datasets (flexible, versatile) that can be used in all stages of responding to threats (detection → analysis → response → prediction and prevention). Commonly, our datasets are built through the following stages: raw data collection, analysis/processing, labelling (annotation) and quality verification. However, since the methodology for collecting raw data and the detailed construction methods such as analysis/processing and labelling vary depending on the dataset type, the scope is too vast to describe all types of dataset construction methodologies. Therefore, this paper focuses on constructing a dataset in the field of threat profiling: APT Attack Group and Botnet Family Dataset.

Human × AI; hybrid dataset construction framework

First, we explain our dataset building framework. In order to maintain the reliability and quality of the dataset and increase the efficiency of the construction work, the work is carried out based on the ‘hybrid dataset construction framework’ in which the human work area and the AI work area are converged. Manual work has the advantage of providing high accuracy, but there are physical (time, manpower, etc.) limitations, and the standards for analysing and judging threats differ from person to person, and even the same analyst may make different judgments depending on situations or conditions. As a result, analyst dependence may be high and consistency may be lacking. In order to compensate for these problems, we analysed the features of each work step in each stage of dataset construction, as shown in Figure 2, and we derived a method of complementing the strengths and weaknesses of each work method and improving efficiency by identifying the manual work area and the automation area.

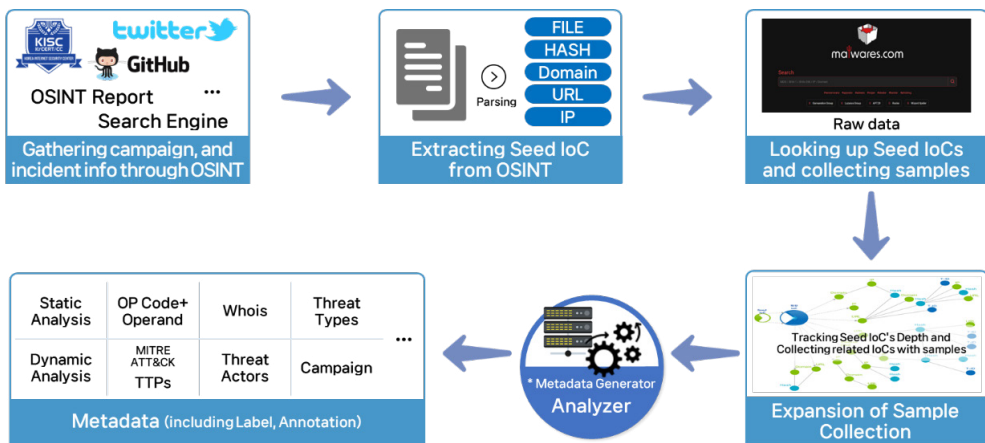


Figure 2: Our process of dataset construction.

In other words, we inject skilled professionals into work areas that require accuracy and decision making, e.g. review of reference reliability (analysis reports, SNS posts, etc.) to secure seed IoCs related to actual infringement incidents, attack groups and campaigns; establishment of the data labelling system (label keywords, annotation methods, thresholds, etc.) and seed data labelling, and creation of seed data and preparation of the foundation through manual work. Based on this, AI with an excellent ability to process large amounts of data consistently within the specified scope of task performance will learn the seed data, analyse new incoming threats, extract metadata, and automatically label them to build a large AI dataset.

Classification	Human	AI
Pros	<ul style="list-style-type: none"> • High accuracy • Flexibility • Ability to process unique cases • Decision-making 	<ul style="list-style-type: none"> • Automation and efficiency • Consistency and reliability (no intervention of subjectivity) • Continuous improvement and learning • Ability to process large amounts of data
Cons	<ul style="list-style-type: none"> • Highly time-consuming and costly • Difficulties in securing consistency and reliability • Limitation in manpower input • Dependence on analysts 	<ul style="list-style-type: none"> • High initial investment and development cost • Difficulties in fulfilling causal relationships • Limitations of a specific domain • Reliability and responsibility

Table 2: Comparison of AI dataset construction between human and AI.

Since AI can be gradually improved through repeated learning and feedback, it is possible to build a high-quality AI dataset by reflecting even the features of the latest threat data to reduce costs and increase efficiency. The quality (accuracy, consistency, conformity, etc.) of a dataset constructed in this way can be verified through periodic review by experts from industry, academia, and research institutes and through a third-party accredited testing agency. In addition, the effectiveness of our dataset was proved through pilot application to actual security infrastructure in the field that operates IT services.

Next, we will explain the construction process of our dataset. Broadly speaking, our project is carried out in two stages: dataset construction and verification.

Dataset construction process

i. Data collection stage

First, we collect data. The first thing to do to build a dataset based on threat information related to various real-life problems is to find the latest threat cases and collect available reference data regarding related IoCs. Therefore, we refer to about 174 issue keywords, collect KrCERT/CC reports, domestic and international threat information reports, SNS posts, etc., and inspect the quality and reliability of the contents of each reference. Then, we construct a reference dictionary by extracting issue keywords related to the selected reference contents, and secure a seed IoC list by parsing IoCs (strings; hashes, domains, IPs, and URLs) used directly or indirectly in the attack. Among the IoCs secured here, we must obtain the original file, not the string value, to create metadata through dynamic analysis. We collect raw data through KrCERT/CC Storage and MWS (malwares.com) API queries. The information on the initial seed IoCs (January 2018 – December 2022) that we have secured is as follows:

Classification	APT group	Botnet family	Ransom family	Social issue
Number of keywords	75	60	10	29
Number of seed IoCs	33,018	22,286	5,981	16,272

Table 3: Status of threat profiling dataset construction.

- APT group (75): FIN8, Lazarus, APT30, TA505, Turla, Kimsuky, Gamaredon, DustStorm
- Botnet family (60): Mirai, XORDDos, Emotet, TrickBot, QakBot, Dridex, Zloader, Kaiji
- Ransom family (10): Clop, Conti, LockBit, Maze, Revil, Hive, Blackcat, Mabniber
- Social issues (29): Covid-19, Russia-Ukraine War, Zoom, NorthKorea, WorldCup

Table 4: Examples of labels in threat profiling dataset.

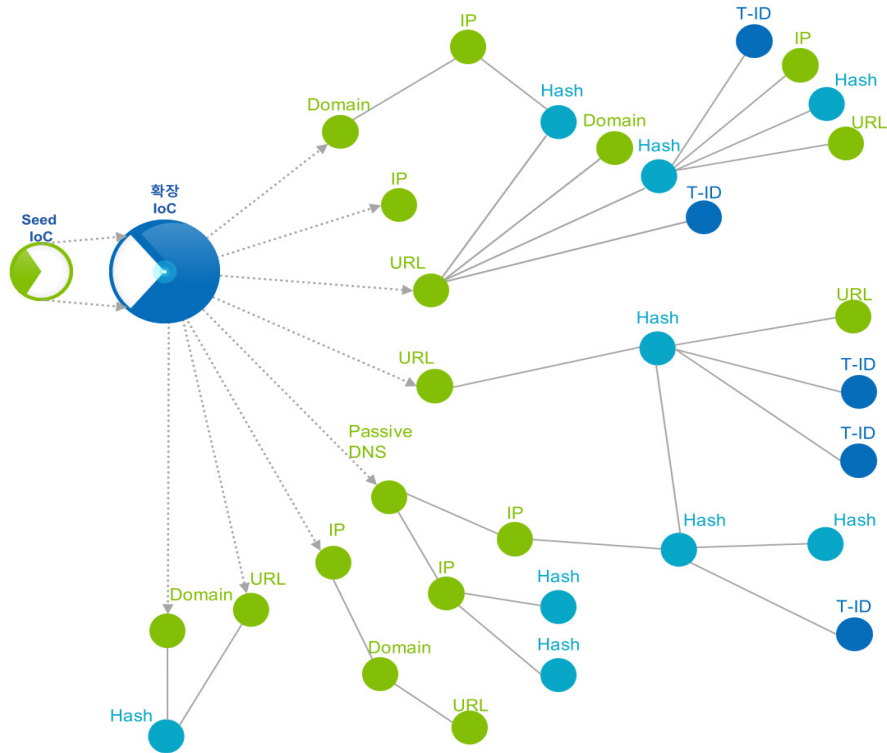


Figure 3: Concept of IoC depth tracking for enriching threat intelligence.

We secure a large meaningful dataset based on the IoCs used in the actual infringement incident by tracking the depth of related IoCs derived from the seed IoCs, as shown in Figure 3. Usually, we construct a dataset by tracing up to three depths based on the seed because, as the dimension of depth increases, the correlation with the seed IoCs becomes weaker. We convert the IoCs secured in this way into structured metadata consisting of various kinds of feature information that AI can learn through dynamic/static analysis and processing.

ii. Data analysis and processing stage; metadata generation

The purpose of this step is to analyse the collected threat data (malicious files) to generate metadata as well as to extract features included in the metadata. By processing the extracted feature information (advanced analysis), the AI model creates feature values that can identify the label and annotation (identifying threat type, TTP, family, and attack group information) of the input data. At this time, the metadata (JSON) of the malware is composed of about 130 pieces of feature information (fields) through static and dynamic analysis. Some data is converted to the STIX 2.1 and image (.bmp) format for scalability of dataset utilization, and is provided as well. Table 5 summarizes the top components of the metadata extracted for each file type.

Network-related IoCs (IPs, URLs and domains) have limitations in terms of property information that can be used in a dataset for AI learning due to the characteristics of data. However, as it is the threat issue keywords related to the IoCs (see Table 5) and information that makes it possible to infer the activity history, attributes of IP/domain metadata as shown in Table 6 are provided as well as the data on the SITX2.0 standard [6], which can be used for intelligence analysis of threat situations.

• DNS	• C&C server information
• Passive DNS	• Country information
• Whois	• Source of attack
• Hostname-based history	• Activity history
• Issue keywords: APT group, botnet family, social issue (or campaign)	

Table 6: Attributes of IP/domain metadata.

iii. Data annotation and labelling stage

The strength and weakness of AI is its dependence on learning data. In particular, in the case of AI modelling based on supervised learning, since the characteristics of data are learned depending on identified label information, labelling to

Type	Property	Metadata count	
Common properties (20)	• HASH (MD5, SHA1, SHA256 and SHA512)	4	
	• File information (name, type, MME type and size)	4	
	• CVE	1	
	• MITRE ATT&CK TTP	1	
	• Strings	1	
	• Vaccine diagnosis result	4	
	• Binary features (N-gram, etc.)	5	
Exec type structure properties (197)	PE (133)	• PE header	60
		• Count by API category	17
		• Number of resources	21
		• Cert info (name, thumbprint, serial number, etc.)	5
		• Entropy (rdata, reloc, text, rsrc, data area, etc.)	6
		• API call sequence	1
		• OP code block (OP code + operand)	1
		• Dynamic analysis behaviour information	22
	ELF (30)	• ELF header	7
		• Section info	6
		• Entropy (init, bss, text, data area, etc.)	5
		• API call sequence	1
		• OP code block (OP code + operand)	1
		• Dynamic analysis behaviour information	10
	APK (23)	• Package info (permission list, main activity, permission, etc.)	8
		• Cert info (name, thumbprint, serial number, etc.)	4
		• Dynamic analysis behaviour information	11
	IOS (11)	• Package info (name, app ID, uuid, etc.)	7
		• Cert info (name, thumbprint, serial number and date)	4
Document properties (28)	• Document information	4	
	• Inserted script information	1	
	• Macro information	1	
	• Dynamic analysis behaviour information	22	
Script (1)	• Script parsing	1	

Table 5: Attributes of malware metadata.

maintain versatility and consistency during the labelling of learning data is a very important task. Accordingly, we are using the AI-based auto labelling technique that enables labelling so that irregularities or noise are not reflected according to the analyst's ability or subjectivity. In addition, since the labelling method differs depending on the types of labels (threat type, threat actor, T-ID, etc.), we developed and applied a model suitable for each characteristic.

First, we label the threat type. As versatility is important at this time, we analysed about 50 anti-virus diagnostic name morphemes, selected threat types specified by a number of AV engines, and performed labelling. In order to secure a list of limited types of threat type labels, we selected and referred to 20 threat types specified in STIX 2 .1, and made a list.

Second, we identify and label (annotate) the characteristics of threat actors and TTPs. In order to identify an attacker or attack technique related to a malicious file, it is important to extract the attacker's characteristics (coding technique) from the file. Accordingly, we performed labelling by applying the depth tracking method mentioned in the data collection stage (see Figure 3) and the DBP (Deep Binary Profiler) technology [7], which was certified as a new technology using this dataset in 2021. Ordinary malware implements the attack technique and the goal that the threat actor wants to achieve through coding. Various attack techniques implemented through coding are compiled in the form of an exec file and executed in the target system to achieve the goal, and this is a reverse technique. In other words, based on the assembly

language (OP-code + operand) extracted through binary reverse engineering, the characteristics of each threat actor or attack technique are extracted and learned by the AI model, and even when new data is entered, it is automatically identified and labelled.



Classification	Threat type	Threat actor and TTPs
Naming standard	STIXv2.1	MITRE ATT&CK Matrix v12 [8]
Concept of labelling	<p>Through morphological analysis (frequency, uniqueness, etc.) of detection names extracted from more than 30 anti-virus engines, threat types and family names can be identified.</p> <p>During morphological analysis, detection names originating from the same engine were excluded and not included in the processing.</p>	<p>The characteristics extracted from malicious samples previously used by attack groups (such as coding habits, code structure, etc.) and the seed TTPs manually identified through human analysis are incorporated into a database and utilized to train an AI model called DBP.</p> <p>This engine is employed to identify attack groups and TTP (tactic, technique and procedure) information.</p>
Labelling process		

Table 7: Comparison of dataset labelling process.

iv. Dataset quality and effectiveness verification stage

Since the cybersecurity AI dataset is directly related to the safety of digital infrastructure that is closely related to our daily lives when applied in the field, we establish and verify a thorough system. In this project, the quality of the constructed dataset is verified through the following three procedures:

- **Self-verification:** With the constructed AI dataset, the AI model is implemented, and performance, e.g. detection rate and accuracy of classification, exceeding a certain level, through dataset learning is reviewed.
- **Third-party quality verification:** Quality objectivity is secured by asking a third-party professional organization to review whether the dataset is properly constructed according to the designed methodology, the effectiveness of the file, the completeness of the format, labelling consistency, etc.
- **Effectiveness verification:** Whether it is effective in responding to real threats is verified through pilot application to IT service institutions and companies, e.g. games, transportation and telecommunication companies.

In particular, in the case of the effectiveness verification procedure, we prove the effectiveness by operating the AI model trained on this dataset for a certain period of time and applying it to various industrial sites that operate IT services or infrastructure. At this time, as the AI requirements are different depending on the characteristics of the operational infrastructure environment for each company, it is important to clearly identify the characteristics of the environment through interviews with the people in charge of infrastructure security as well as through a simple technical approach, and select and apply an appropriate dataset. For example, games companies requested an AI dataset for detecting and automating classification of hacktools that degrade game quality, and email service providers requested a dataset for enhancing the performance of an engine that detects new and variant malware due to the rapid increase in email exploit attacks that have increased dramatically due to Covid-19.

Therefore, the process of verifying the dataset involves not simply verifying the quality of the data to improve the performance of the AI model, but also reviewing the bias, stability, and ethical considerations of the data. We are making efforts to improve the completeness of the dataset by repeatedly performing the three procedures mentioned above and gradually confirming the empirical aspects.

4. STRENGTHS OF OUR DATASET

Together with the active investment (cost, budget, manpower, technology, etc.) of the Korean government, we are building a complete dataset by mobilizing systematic planning capabilities and technology through cooperation (tightly coupled) with security companies with long experience in responding to infringement. Our project team has been carrying out this project by carefully analysing the strengths and weaknesses of the previously studied cases, accepting the strengths and improving on the limitations, and as a result, our dataset has the following strengths:

First, our dataset contains more than 24 file types, such as APK, iOS, ELF, PDF and DOC, as well as EXE and DLL. In the real world, not only executable file types, but also various types of malicious files such as HTML and documents are distributed in new and variant forms. In particular, with the recent increase in remote working due to the Covid-19 pandemic, the distribution of document-type malware is rapidly increasing. We have tried to expand the scope of AI model utilization by securing malware including various file types to reflect the real-world situation.

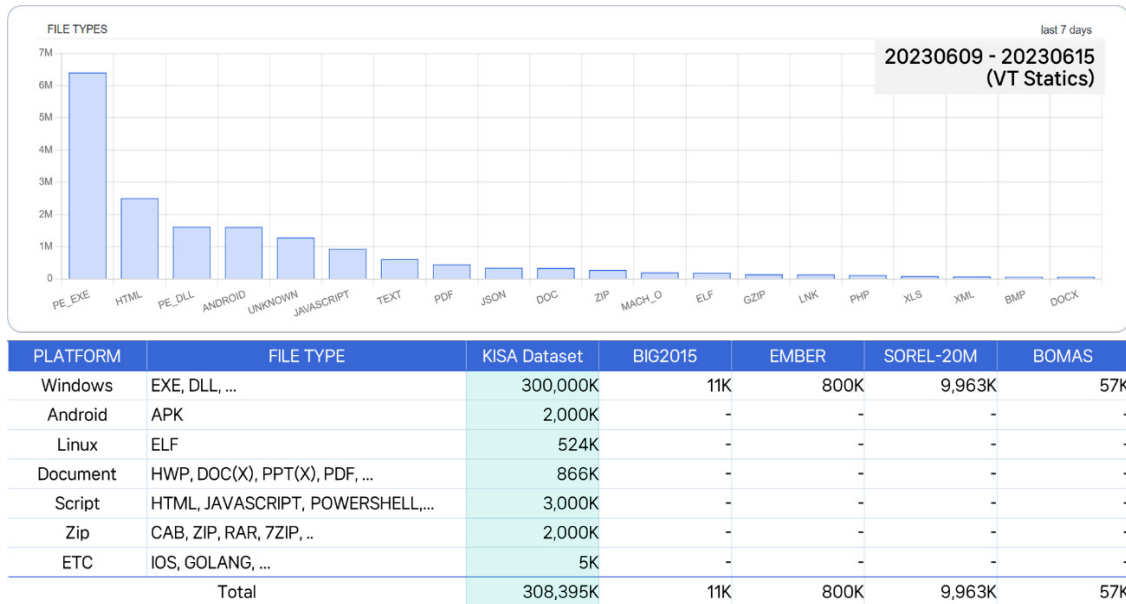


Table 8: The strength of our dataset #1 – huge size and various file types.

Second, we have been striving to solve these problems by using AI labels to build datasets, and constantly thinking about the following questions:

- Are the AI label keywords diverse enough to address realistic security threat issues due to various factors?
- Are the grounds and logic reasonable enough to secure the reliability and versatility of AI labelling?

Accordingly, by appropriately distributing and utilizing analysis manpower and AI technology, we have built an automation system that can improve the accuracy of AI labelling, and label large amounts of data within a given time. In addition, as the detection criteria and detection targets may be different for each company, we have configured the labels in a form that enables AI learning by classifying the file type, threat type, and family of the malware. At this time, the labels for the threat types are based on the types presented in STIX 2.1, and for families, we used the text mining technique to create labels by extracting keywords related to the names of the families from more than 50 anti-virus diagnosis names.

Third, by extracting various types of features, we have constructed useful metadata. By ensuring that the features, provided by previous projects, include not only bytes and entropy histograms, but also dynamic analysis feature information that takes a long time to extract, we have constructed metadata so that dataset users can try AI modelling from various perspectives.

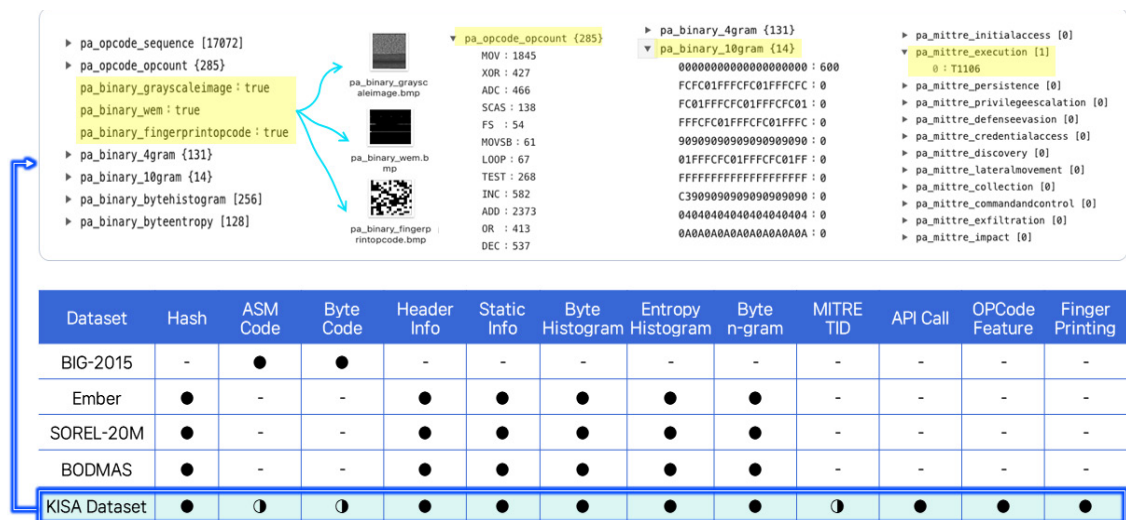


Table 9: The strength of our dataset #3 – a variety of feature vectors.

Table 10 summarizes the result of comparing our dataset with existing datasets. As the data about previous projects has already been disclosed at an excellent level, we were able to design our project by benchmarking them. However, in order to proactively respond to threats in the complex real world, we determined that an extended version of the existing datasets would be necessary, and we made improvements through this project.

Dataset	Host	Country	Malware Time	Malware Binaries	Feature Vectors	TTP	Threat Types	Families	# Threat Types	# Families	# Samples	# Benign	# Malware
BIG-2015	Kaggle (Microsoft)	US	Before 2015	0	-	-	-	●	-	9	11K	-	11K
EMBER	Elastic (endgame)	US	01/2017-12/2018	-	●	-	-	-	-	-	2,050K	750K	800K
UCSB-Packed	University of California	US	01/2017-03/2018	●	-	-	●	-	10	-	341K	109K	232K
SOREL-20M	Reversing Labs (Sophos)	UK	01/2017-04/2019	●	●	-	●	-	11	-	1,972K	9,762K	9,963K
BODMAS	University of Illinois	US	08/2019-09-2020	●	●	-	-	●	-	581	134K	77K	57K
KISA Dataset	KISA (SANDS Lab)	KR	08/2018-10/2021	●	●	●	●	●	17	6,828	300,000K	60,000K	240,000K

Table 10: Comparison of malware dataset.

5. EXAMPLE OF DATASET UTILIZATION AND EXPECTED BENEFITS

Since the uses of AI are not fixed, it is possible to perform AI modelling in various ways depending on the purpose, even with the same dataset. At this time, the technical approach is important for AI modelling, but it is also important to discover ideas about how to apply AI in our lives to increase efficiency. Therefore, from the start of this project in 2021 until last December, we discovered ideas for using the dataset and selected best practices through pilot application of the dataset to about 51 companies/institutions. In addition, we have been conducting various activities to intellectualize the infringement response system through AI dataset and technology circulation in the field of cybersecurity, while continuing to spread best practices, e.g. by holding meet-up days to share them and distributing casebooks. Now, we will introduce a few examples.

Understanding trends in threats based on social issues

The biggest insight we gained from this project was confirmation of the possibility of responding to current and future threats through learning from past data. For example, the same attacker has similar coding habits, techniques, access

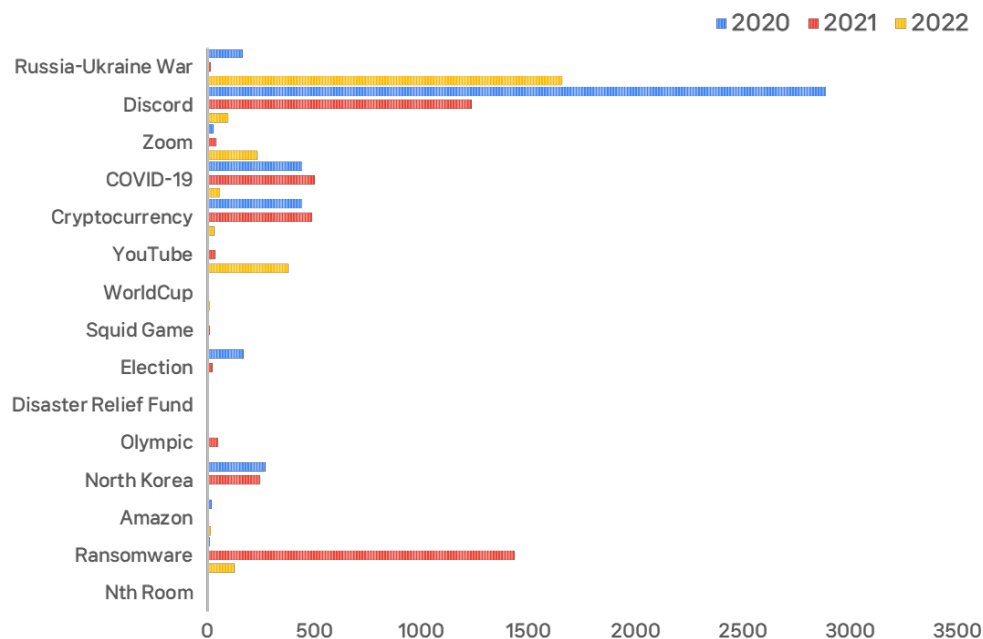


Figure 4: Statistics on attacks exploiting social issue keywords by year (2020-2022).

strategies, frequently used attack tools and subjects regarding attack targets, similar to the way in which newspaper articles written by the same reporter have similar topics of interest and narrative techniques. In addition, modern attackers do not simply show off their hacking skills or steal money, but often plan their attacks with political and social purposes. So the social issues we are experiencing in reality are also attractive subjects (traps) from the attacker’s point of view. This can be seen from articles or reports in which attacks related to social issue keywords appear every year. Figure 4 shows the statistical information from the threat reports that we have accumulated since 2020, and it can be seen that cyber threats increase in a similar trend according to the relevant issue of each year. We expect that it will be possible to respond proactively to future threats if we extract the characteristics of those that have occurred in the past, map labels for related issue keywords, and the AI model uses the dataset built through this to sufficiently learn even the trends in social issues.

Example of AI-based identification of attack groups

In the past, it was common to use IoCs to respond to cyber threats. However, as modern cyber threats are becoming more diverse and complex, attackers are using a variety of techniques and methods to evade IoCs and circumvent detection. Accordingly, many companies have recently moved away from detecting/blocking incoming IoC information by simply identifying whether it is malicious or normal, and try to identify the intention of the attack, attack groups and technique by understanding the contextual situation of the threat. This is because they want to prevent further damage by identifying even the inside story of an attack that threatens the company, and to establish a policy or make a decision to proactively respond to it.

Three companies with these requirements participated in a trial and demonstrated an AI-based attack group profiling model built with this threat profiling dataset. The demonstration structure is shown in Figure 5, and Table 11 shows the information on the industrial sector mainly serviced by each demonstration site and the status of the samples requested for demonstration. In this structure, the demonstration model is requested to analyse the unknown malicious files collected from the equipment (EDR, UTM and anti-spam) of each demonstration site through APIs, and the attack group and attack technique information similar to the requested sample is returned. The experimental results are shown in Figures 6 and 7 (as this paper deals with the value and meaning of the dataset, a detailed description of the AI model is omitted).

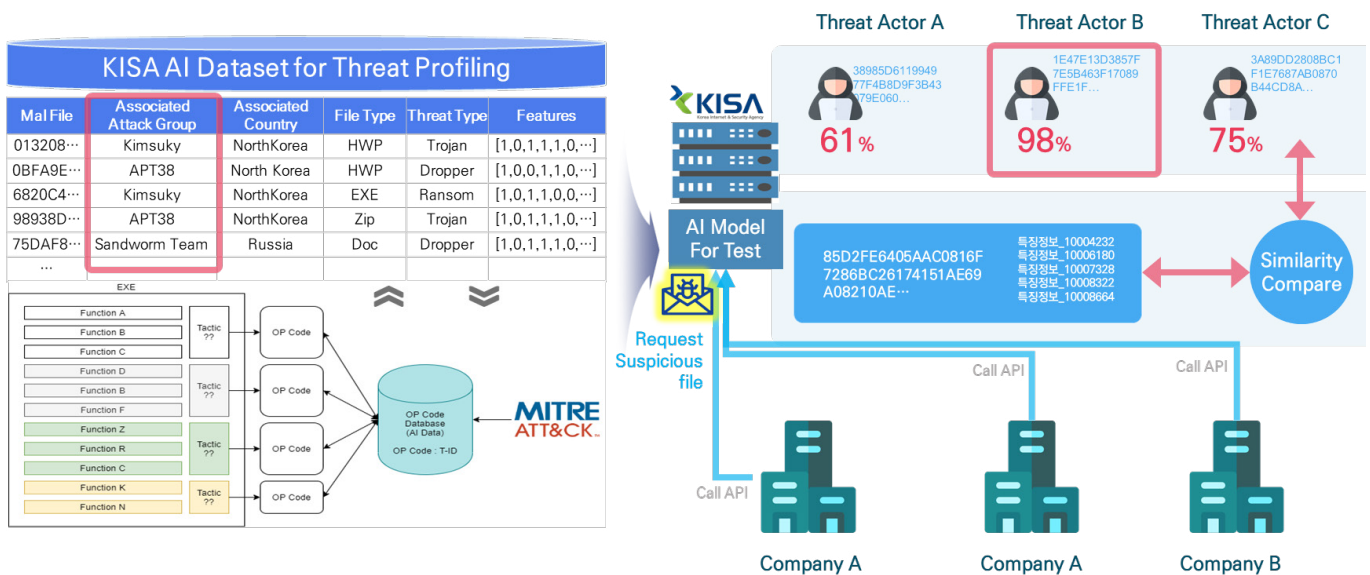


Figure 5: Demonstrative structure of our dataset applied.

Category	Company A	Company B	Company C
Samples	107,148	149,323	1,205
Target system	EDR	UTM	Anti-spam solution
Target sector	Healthcare, financial, education & research	Government, financial, defence, education, telecommunications services	Government, defence, non-profit

Table 11: Information on companies participating in dataset validation.

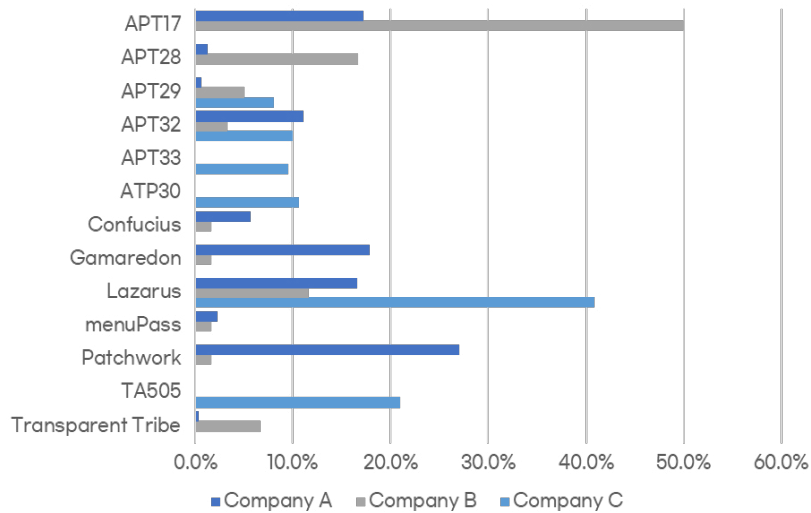


Figure 6: Results of attack group identification using the validated AI model.

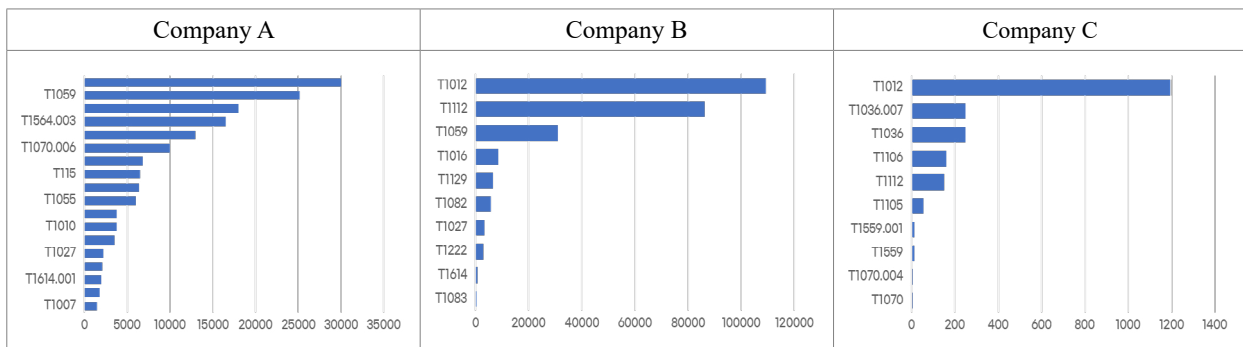


Figure 7: Results of TTPs identification using the validated AI model.

Looking at the above results, it can be seen that a large number of attack groups targeting the industrial sectors served by each participating company are detected. For example, in the case of Company C, the most frequently identified groups are Lazarus (40.8%) and TA505 (21%). These groups mainly carry out initial attacks by using sophisticated attack techniques to attach document-type malware disguised as recruitment requests and job descriptions to emails. Compared to Companies A and B, the fact that many samples, collected from email security services, were identified in association with each other can be interpreted as a meaningful result. In addition, the top identified attack technique T1012 (Query Registry) is also a technique found in elaborately crafted document-type malware that bypasses detection by hiding its name or file path. This shows that the AI model can identify the characteristics or techniques of attack groups to a certain extent without the intervention of people.

As this AI model is built for demonstration, there may be a margin of error. But we believe that if it is more elaborately tuned, it can contribute to resolving the difficulties encountered at infringement response sites. The participants in the demonstration said that this would contribute to efficient use, and time and manpower management in the initial analysis stage.

6. CONCLUSION

An IoC is fact-based data, and when accumulated, it can become a valuable resource as it generates information and knowledge about a threat. We have confirmed these values and possibilities through demonstration, and we are expanding and advancing this project with a vision of realizing cyber resilience by creating an environment for the use of datasets in the field of cybersecurity and intellectualizing all stages of responding to infringement. In other words, we will reduce human workload by improving the efficiency of security work through AI, and intensively invest human resources in tasks that are difficult for AI to do, such as security policy setting, decision making, and advanced analysis.

In other words, at a certain point in time, threat information must be interpreted, and humans must make important decisions based on the interpreted information. Provided with accurate and sufficient data, humans can make correct decisions with ease. In the past, we shared only IoC information for blocking threats, but fragmented, incomplete and insufficient data makes it difficult to make correct decisions. Automated technologies like AI with sufficient datasets provide information with a sufficient entropy level in the existing response mechanism of providing fragmentary information, thereby enabling a mid-to long-term response policy.

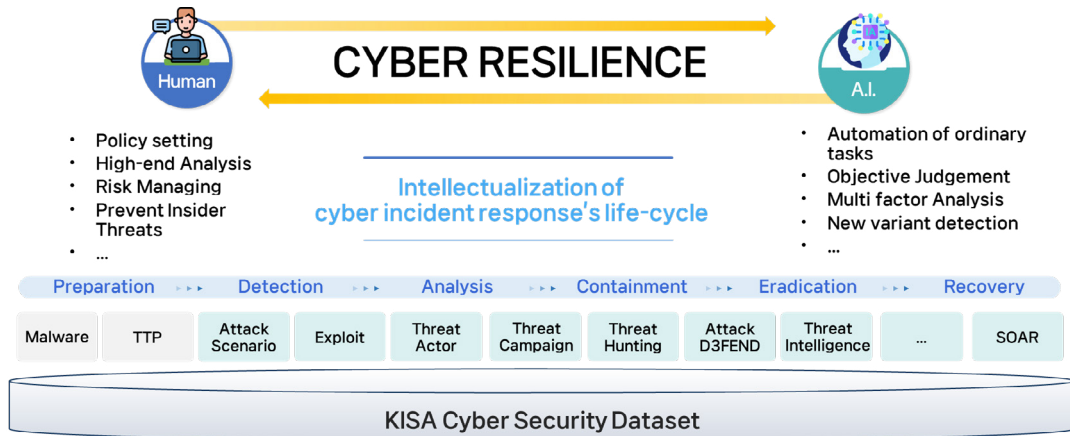


Figure 9: Our dataset project vision and future directions.

We hope that our dataset will be fully utilized here. We will continue to carry out activities to discover the value and meaning of the dataset we have built, and we will widely publicize and develop the dataset we have built through active cooperation with domestic and foreign companies/institutions. We hope that by allowing many people to use it, they will discover new ideas about using AI in the field of collective intelligence-based security, and AI will become gradually more intelligent and advanced, e.g. utilizing the dysfunction of AI, and automatically judging the situation of complex cyber threats, and contributing to the realization of a safe digital world through prediction of future threats and preemptive responses.

REFERENCES

- [1] McDonough (1963). Information Economics and Management Systems.
- [2] Verizon. 2023 Data Breach Investigations Report. <https://www.verizon.com/business/resources/reports/dbir/>.
- [3] Ronen, R., et al. Microsoft malware classification challenge. arXiv preprint arXiv:1802.10135, 2018.
- [4] Anderson, H.S.; Roth, P. Ember: an open dataset for training static pe malware machine learning models. arXiv preprint arXiv:1804.04637, 2018. <https://arxiv.org/abs/1804.04637>.
- [5] Harang, R.; Rudd, E. M. SOREL-20M: A large scale benchmark dataset for malicious PE detection. arXiv preprint arXiv:2012.07634, 2020.
- [6] STIXv2.1. <https://oasis-open.github.io/cti-documentation/resources#stix-21-specification>.
- [7] SANDS Lab. Multidimensional Metadata Extraction Analysis based Non-executable Malware Profiling and Detection Technology. https://www.netmark.or.kr/sub4/pop_tech_detail.asp?recv_seq=20144468.
- [8] MITRE. ATT&CK Matrix. <https://attack.mitre.org/matrices/enterprise/>.