Int. J. Nonlinear Anal. Appl. 13 (2022) 1, 4037-4052
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.6230

Concept and difficulties of advanced persistent threats (APT): Survey

Eman J. Khaleefa[a,*], Dhahair A. Abdulah[a]

[a] *Diyala University, Diyala, Iraq*

(Communicated by Madjid Eshaghi Gordji)

Abstract

Previously confined to nation-states and associated institutions, dangers have increasingly penetrated the private and business sectors. Advanced Persistent Threats (APTs) are the type of threats that every government and established organization worries and seeks to counter. While state-sponsored APT assaults will always be more sophisticated, the increasing prevalence of APT strikes in the corporate sector complicates matters for corporations. Existing security solutions are becoming ineffective as attack tools and techniques evolve at a breakneck pace. While defenders attempt to safeguard every endpoint and connection in their networks, attackers come up with new ways to breach their targets' systems. In this scientific study, we will discuss the issue (APT) and what it includes in terms of obstacles or difficulties, as well as the current state of progress in this field. Additionally, we will present an overview of the most commonly used dataset support (APT) for algorithm assessment and highlight the approaches and strategies used.

Keywords: Advanced Persistent Threats, Attacks, Businesses, Security, Challenges, Dataset, World Wide Web.
2010 MSC: 68M25

1. Introduction

The Internet has developed from a system for serving an interconnected set of static pages to a powerful, diverse, and vast platform for application delivery and information dissemination. Businesses have increasingly put essential resources and sensitive data online [38]. Unfortunately, as the web's power and popularity have grown, so has the number and influence of cybercriminals [39].

[*] Corresponding author

Email addresses: scicompms2106@uodiyala.edu.iq (Eman J. Khaleefa), dhahair@sciences.uodiyala.edu.iq (Dhahair A. Abdulah)

Received: November 2021 *Accepted:* January 2022

This results While people benefit from the incredible convenience of the Internet, they are always at risk of cyber-attacks as the Internet and artificial intelligence (AI) grow. There are no exceptions when it comes to smart grid and industrial Internet technology. Internet Emergency Center data shows that, as a result of (CNCERT) [34]. APT (Advanced Persistent Threats) is a severe Internet menace, and the software allows attackers to take control of infected PCs and steal sensitive data from afar [47]. As a generalization, APT assaults are getting more prevalent on the internet these days. Regrettably, detection is challenging [62]. It is a form of continuous hacking and a set of covert approaches targeted against a particular institution with high-value data, such as the government, military, or financial industry.

APT attacks are carried out with the intent of stealing data rather than causing damage to the company or network. After successfully breaching the network, an attacker installs APT malware on the target workstation [68]. That means The intention of an APT attack is to steal data rather than cause damage to the network or organization [45]. It is a kind of cybercrime directed at companies and governments. APTs must maintain a high degree of stealth for a lengthy period of time in order to be effective. Even after critical systems have been penetrated and initial goals have been accomplished, the attack objectives often extend beyond immediate financial gain, and compromised systems continue to function [64]. According to the United Kingdom's National Cyber Security Centre, it is a "targeted cyber-attack in which a hacker accesses a system and stays undiscovered for an extended period of time" [8]. From this definition alone, the high likelihood of malevolent actors adopting a range of infiltration strategies to compromise targets can be concluded [16]. There have been cyber attacks since the beginning of the Internet, and they've evolved dramatically over time, from worms and viruses to malware and botnets. "Advanced Persistent Threat (APT)" has emerged in recent years as a new sort of threat. The term "APT" was first used to describe cyber incursions against military organizations, but it has since evolved and is no longer restricted to the military sphere [58].

2. Characteristics APT

In recent years, the word has come to denote complex attacks involving far-flung resources aimed at certain organizations in order to achieve a specified goal. Each word in APT, in particular, has its own unique meaning [66]. Offers the following definitions for each word in table 1:

Table 1: Definitions Of Word APT [52].

NO.	Name	Description
1	Advanced	The attacker boasts considerable technical skills and uses a number of threat vectors, including malware, vulnerability scanning, spear phishing, and social engineering, to exploit weaknesses inside the target.
2	Persistent	APTs frequently occur in stages over an extended period of time, including identifying the organization's vulnerabilities, exploiting those vulnerabilities via multiple threat vectors, expanding control once access is gained, and continuing the attack, which might also result in long-term planning to avoid detection.
3	Threat	The attacker is motivated and capable of successfully committing an attack. The objective is to amass significant assets such as money and information from designated organizations such as government agencies, critical infrastructure systems, and financial institutions

In order to obtain financial advantage, or merely to demonstrate their mettle by wrecking havoc on a company's reputation, an attacker or a gang of attackers may have attempted to bring down the business they targeted. In any of these assaults, the perpetrators made no attempt to conceal their identities or their behavior [69]. These types of attacks still exist, but there is a new style of attack that has grown in popularity in recent decades. This sort of attack is described by a gang of attackers operating slowly and discreetly in order to accomplish their objective, which is often data theft without being noticed. An attack (APT) is a type of attack (APT) that may use well-known techniques to acquire access to a target institution's networks, but its tools are new to the attacker [56]. In order for an attacker to remain active in the network for an extended period of time, they need sophisticated tools, as the term suggests. Traveling between systems within the organization's network, they acquire critical information and carefully export it to their command-and-control center in a low-profile manner [46]. It is common for APTs to be conducted by well-funded attackers that have the resources they need for a pre-determined amount of time. When the financing organization has all the essential information, the assault is over [60]. Regardless of the scenario, the company that was the victim of an APT attack would have suffered significant damage, possibly even permanent loss, as is often the case when the attack goes unnoticed until all of the organization's data has been hacked. APT victims are routinely questioned about their inability to detect the attack even after installing powerful intrusion detection and prevention systems [9]. APT can be illustrated by the following figure 1, which represents its life cycle:

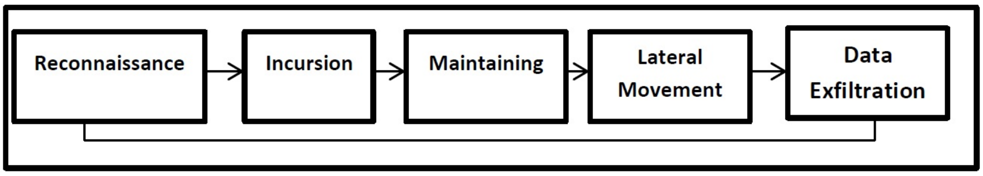

Figure 1: APT life cycle [46].

The stages of the life cycle can be summarized in the table 2:

Table 2: Definitions Of Word APT [52].

RF	Name of Stage	Description Of Stage
[13]	Reconnaissance	The reconnaissance phase, during which cyber enemies acquire intelligence on a target organization, is the initial step taken by cyber adversaries. To obtain as much knowledge about the target organization's network as possible, this entails both human and technological aspects. This is used to find weak spots in the organization's network and infiltrate them. Intelligence reconnaissance in this phase entails obtaining information from both technology and weak human links.
[30]	Incursion	Throughout this phase of intrusion, the APA makes a number of efforts to gain access to the organization's network. Spear phishing and other social engineering techniques are used to deliver targeted malware using SQL injection in this method. A zero-day flaw in the software system is also used to detect network holes.
[7]	Maintaining	Remote administration tools allow cyber attackers to keep access to a payload once they have gained access to an organization's network (RAT). There is a command and control server outside of the organization that links to the RAT. Encryption of an HTTP connection between the host and an external command and control server is common. Disguising itself and going undetected, malware is able to get beyond network firewalls and other security measures.
[37]	Lateral Movement	The APT virus spreads across the network to additional uninfected hosts. The other host often has more privileged access, which raises the likelihood of sensitive content and data exfiltration being included.
[48]	Data Exfiltration	Data exfiltration is the final step in the APT cycle. The host delivers the data it has gathered to an external source or cloud during this phase. One burst of this technique may be used, or it may be applied gradually without the user's knowledge.

APT assaults target education, banking, technology, space exploration and aviation, power supply, chemistry, telecommunications, medicine, and consultancy [17]. APT attacks take advantage of security flaws in a wide range of devices, including Internet of Things (IoT) devices [6]. Actors may capitalize on current events that stimulate the public's interest, resulting in a new scope. COVID-19's present dilemma has provided an excellent environment for actors to launch their assaults. The bait in this instance was advice about the situation of healthcare in numerous nations, and as a result, spear-phishing, remote access tool assaults, and ransomware were used [41]. Finally, The distinguishing characteristics of APTs are:

1. precise aims and unambiguous objectives.
2. attackers who are extremely organised and well-resourced
3. a long-term effort characterized by repeated attempts.
4. Techniques of assault that are subtle and elusive.

And the following table 3 elaborates on each of these characteristics:

Table 3: APT Characteristics [25, 37].

RF	Name	Description
[16]	precise aims and unambiguous objectives	APT attacks are very focused and are always directed at a particular objective. Of the top ten target industries, the bulk of them are governments or large enterprises with significant intellectual property value in fields including education, finance, high technology, government, consulting, and energy, chemicals, telecommunications, and aeronautics. Traditional attacks spread far and wide in order to maximize their chances of success and harvest, but an APT attack focuses solely on pre-defined targets, restricting the attack range to a smaller area. Digital assets that provide a competitive advantage or strategic benefit, such as national security data, intellectual property, and trade secrets are typically targeted by APTs while personal information, such as credit card data, or generically valuable information that facilitates financial gain, is typically targeted by traditional threats.
[53]	Attackers Who Are Extremely Organised And Well-Resourced	APTs are often comprised of a group of skilled hackers that collaborate. Governments and commercial enterprises may hire them as cyber mercenaries or assign them to a government or military cyber unit. They are well-equipped on both the financial and technological fronts. This enables them to work for extended periods of time and has provided them with access to zero-day vulnerabilities and attack tools (through development or procurement). When they are state-sponsored, they may even have military or state intelligence on their side.
[12]	A long-term effort characterized by repeated attempts.	APT assaults can go undetected on a target's network for months or even years before they are uncovered by the victim. While prior attempts to complete the mission have failed, APT actors continue to attack their targets and change their efforts to achieve the goal. This is an unique threat from conventional attacks, which typically target a wide variety of victims and move on to a less secure target if they are unable to infiltrate the original target.
[50]	Techniques of assault that are subtle and elusive	APT assaults are stealthy, having the ability to stay undetected by blending into business network traffic and engaging only when necessary to accomplish the attacker's objectives. To evade signature-based detection, APT attackers may disguise network traffic using zero-day vulnerabilities and encryption. This is in contrast to conventional assaults, in which attackers often use ”smash and grab” techniques to gain the attention of the defenders.

3. APT Responsible

Governments and nation-states are increasingly using cyberspace to carry out attacks. Because of these players' considerable cybernetic skills, concerns about election tampering or disruption of

Table 4: Famous Responsible In APT.

RF	Responsible Name	Description
[13, 27]	China	Chinese cyber-attacks have been identified as targeting industrial espionage and intellectual property theft. This actor's most consistent cyber threat has been APT1.
[22]	United States	Individuals like this one have the ability to conduct the most advanced cyberattacks. Since the attacks were successful in causing harm, it can be assumed that a large amount of money and effort was put into developing and perfecting this sort assault. Geopolitical objectives have always been at the heart of the APT's mission. Stuxnet, the world-famous cyberattack, was one such operation. Its goal was to severely disrupt Iran's nuclear program's SCADA (Supervisory Control and Data Acquisition) systems.
[15]	Russia	This actor is a major player in state-sponsored APT activities. These organizations have been the subject of in-depth investigations as a result of their involvement in high-profile invasions. APT28 targeted German government employees with spear-phishing assaults, according to Microsoft. Employee credentials and malware-infected websites have been targeted by this team.
[36]	Iran	With several cases carried out by different groups in the Middle East, this actor owns the country's biggest assault potential. Due to the new enhancements to APT33's infrastructure, analysts have been following its activities. Aviation and energy companies with connections to petrochemical manufacturing have been the group's primary targets. Recent malware operations have concentrated their efforts on organizations in the United States, the Middle East, and Asia.
[60]	North Korea	This actor's cyber organizations have carried out a variety of operations, including traditional espionage, financial intrusions, and damaging strikes. The malware WannaCry is one example used by this perpetrator.

electrical supplies in other countries are causing widespread public worry. Table 4 summarizes the well-known doers:

APT attacks have become more careful and damaging since the discovery of (Stuxnet) [21], high-profile systems can be easily hacked while avoiding many of the more complex procedures employed to defend the computer environment. A lot of these threats are currently overlooked. Once they've been discovered, many of these dangers return with only minor alterations. An attack on the financial, private, and intellectual property of individuals has occurred [62]. The class APT attack type and to how belong in table 5 can be divided accordingly.

Table 5: Classic APT Attack Cases [62].

Year	Target zone	Organization
2009	China	APT-C-39
2012	South-East Asian countries, China	Ocean Lotus
2012	Middle East and Europe	APT33 (doubtful)
2014	Global, SWIFT	APT38
2016	Ukraine	APT28(doubtful)
2017	Korea	Hades
2018	North America, Europe	APT28
2018	China	Blue Mushroom
2018	China, Pakistan	BITTER
2018	China	Dark Hotel

There are two distinct sorts of attacks: conventional and unconventional. Figure summarizes the distinctions between conventional threats and APTs in terms of numerous attack features 2:

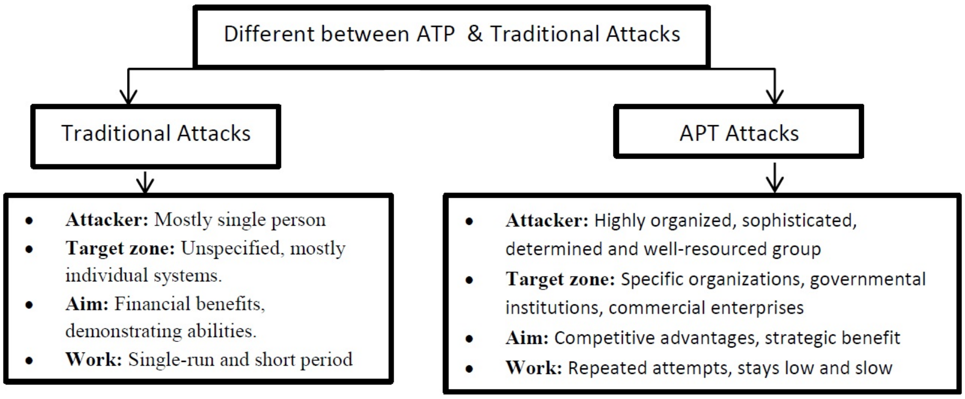
Different between ATP & Traditional Attacks
Traditional Attacks
APT Attacks
Attacker: Mostly single person
Target zone: Unspecified, mostly individual systems.
Aim: Financial benefits, demonstrating abilities.
Work: Single-run and short period
Attacker: Highly organized, sophisticated, determined and well-resourced group
Target zone: Specific organizations, governmental institutions, commercial enterprises
Aim: Competitive advantages, strategic benefit
Work: Repeated attempts, stays low and slow

Figure 2: Different between ATP & Traditional Attacks [7, 34]

Relying on the above diagram, there can be criteria for assessing whether the attacks were intentional APT or not [7, 30, 45]:

- **This attack may have been avoided in a variety of ways:** Assaults that are unexpected and extremely likely (in relation to the target surroundings) require the bare minimum of countermeasures and security procedures to prevent them from occurring.
- **This assault required little change on the part of attackers:** There may be a problem with the target environment's defensive system if attackers' efforts to achieve their goal do not need considerable adaptation or extensive evasive movements when confronted with defensive measures.
- **This assault lacked uniqueness in its variants:** The effectiveness of an APT attack is usually attributable to the attack methodologies or approaches being novel. If the attack techniques or approaches are not novel, they will be detected via the use of established tools and processes.

While the activities used in APT assaults are consistent across all of these attack categories, they are either too broad or overly specific. When it comes to dealing with APT attacks, we can break them down in to five distinct phases, each of which can represent any APT attack regardless of intent, as shown in table 6:

4. Famous Methods of APT

APTs employ a wide range of strategies and tactics. Using the target's profile, social engineering, and spear-phishing, attackers can persuade victims to download a malicious file via email or spear-phishing [2]. The attacker then compromises the computer and exploits the network to get access to the organization's other systems. The most ”advanced” APT organizations differentiate themselves by the exploitation of zero-day vulnerabilities and previously discovered infection vectors [36]. This strategy enlists the help of multiple government agencies from several nations in order to steal secret information for an extended period of time without being discovered [54]. Depending on the target, the strategies frequently employed to carry out an APT assault are altered or mixed [23]. The following table 11 are some famous instances of these techniques [46]:

Table 6: Stages Of APT Attacker [20, 28, 38].

Stage Number	Stage Name	Description
One	Reconnoitering	The target is understood, and the more you explore, the more efficient the attack will be (first important stage).
Tow	Establish Foothold	This stage denotes penetration and entrance into the objective, as well as stability. In order to achieve their goal, they need to get access to the target's network (second import stage)
Three	Stay Undetected	The attackers will have to traverse the target's network laterally if they want to compromise critical components or steal sensitive data.
Four	Exfiltration/Impediment	This stage involves operations such as data retrieval and delivery to the attackers' command and control center when their goal is to obtain corporate data. If the attackers' goal is to weaken or destroy essential components of the target organization, this is where they will conduct their attacks.
Five	Post-Exfiltration/Post-Impediment	This step comprises duties such as completing the exfiltration process, deactivating further critical components, and destroying evidence in order to guarantee a clean withdrawal from the organization's network. That is, either aim for stability and stillness, as well as information withdrawal, or leave.

Table 7: Methods Of APT [7, 18, 32, 62].

RF	Techniques Name	Description
[35]	Social engineering	Rather than launching an ad hoc attack on a system, this strategy aims to persuade a user to risk information systems by enticing them to hand up personal information in order to carry out a hostile attack.
[4]	Spear-phishing	This is a method used to get the login passwords, financial information, or other sensitive data of a specific company.
[55]	Watering hole	This is a method of attempting to access a specific organization's user credentials, financial information, or other secret information.
[49]	Drive-by-download	When a rogue web page is browsed, this approach unintentionally downloads and executes malicious software. "Invisible" malware is downloaded without the user's awareness by exploiting security weaknesses, browser exploits, or embedded plugins such as "ActiveX, Java/JavaScript, or Adobe Flash Player".
[63]	Anomalous Flow-based Detection	Unexpected behavior may be discovered by utilizing network traffic analytic models, which is the underlying principle underpinning anomaly detection in network traffic.
[26]	Machine Learning-based Detection	Detection of APT attacks has always been difficult due to the sheer volume of data involved, and machine learning now plays a vital part in this process. For the machine learning-based detection technique, the primary objective is to identify potential threats by utilizing the processing power of machine learning to identify and record behavioral patterns. Human assistance is not required in the execution of this procedure.

5. Famous dataset in APT

The data set is the cornerstone upon which an efficient model is built [65]. The following is a table 8 showing what is the most important data set available in a topic:

Table 8: Famous Dataset in APT

RF	Data set name	Year	Description	Free
[40]	DARPA98	1998 AND Update In 2015	Darpa is a collection of communications between source and destination IP addresses. This dataset comprises a variety of attacks from various IP addresses. Results of the DARPA 1998 Offline Intrusion Detection Evaluation, by Richard Lippmann et al., updated in 2015.	√

[70]	kDD,99	1999	Annotated results of KDD-99, the Fifth International Conference on Knowledge Discovery and Data Mining's Third International Knowledge Discovery and Data Mining Tool Competition. It was the goal of the competition to design a network intrusion detector capable of distinguishing between "bad" connections, which are sometimes referred to as intrusions or attacks, and "normal" connections. This database contains a standardized collection of auditable data, including a variety of simulated intrusions in a military network environment.	√
[71]	UNIBS	2011	The edge router of the University of Brescia campus network was traced throughout the course of three days of work (September 30th, October 1st, and October 2nd). Twenty computers running the GT client daemon are responsible for their creation. Tcpdump was used to collect the data on the faculty router, which is a Linux machine with a dedicated 100Mb/s uplink that links our network to the Internet. The router's internals are connected to a specific ATA controller, which stores the trace data on a dedicated hard disk. In the end, I had over 27GB of data, with almost all of it being TCP and UDP traffic, with about 79000 individual trips. P2P programs like BitTorrent and Edonkey create traffic that is included in the service (FTP, SSH, and MSN). additional information is available here. The anonymized and payload-stripped traces take up around 2.7GB of storage space.	☒
[72]	NSL-KDD	2015	The NSL-KDD is a data set that has been suggested to remedy some of the issues raised in the research about the KDD'99 data set. Despite the fact that this updated version of the KDD data set retains some of the issues discussed by McHugh and may not be an exact representation of existing real-world networks, it can still be used as an effective benchmark data set to aid researchers in comparing different intrusion detection methods due to the scarcity of publicly available data sets for network-based IDSs. Additionally, the NSL-KDD train and test sets have a sufficient number of records. This benefit enables studies to be conducted on the full set without having to randomly choose a small sample. As a consequence, numerous research initiatives' results are analyzed. It enhances the KDD '99 dataset.	√
[44]	UNSW-NB15	2015	The UNSW-NB 15 dataset's raw network packets were generated using the IXIA PerfectStorm program at UNSW Canberra's Cyber Range Lab to produce a combination of genuine current regular activities and synthetic contemporary attack behaviors. 100 GB of raw traffic was captured using the tcpdump program (e.g., Pcap files). Ffuzzers, analyses, backdoors, denial-of-service attacks, exploits, generics, reconnaissance, shellcode, and worms are all included in this collection. To construct a total of 49 attributes connected with the class label, Argus and Bro-IDS tools are combined with twelve algorithms. In the UNSW-NB15 features.csv file, you'll find information about these attributes. Backdoors, DDoS assaults, exploits, fuzzers, analysis, general, reconnaissance, shellcode, and worms are all included in the dataset. Extra attributes are included in the data set, which may be accessed in both packet-based and flow-based forms.	☒
[73]	NGIDS-DS	2016	In a 2016 network environment, seven variables were obtained from raw network data and nine from log files. It maintains both packet-based network communications and host log files. The collection includes attack families such as backdoors, denial-of-service, exploits, generics, reconnaissance, shellcode, and worms.	☒
[74]	TRAbID	2017	It contains sixteen distinct IDS assessment scenarios, each of which was collected in a simulated environment with one honeypot server and one hundred clients. Each scenario lasted 30 minutes, for a total of 8 hours. Port scanning and denial-of-service (DoS) assaults are two types of malicious network activity.	☒
[75]	CIC-IDS2017	2017 and last update in 2021	Examples of common threats, both benign and current, may be found in the CICIDS2017 dataset, which is an excellent representation of actual data (PCAPs). A study of network traffic using CICFlowMeter is also included, which contains labeled flows based on the time stamp, IP addresses and port numbers of the sources and destinations, protocols and attack vectors that were found (CSV files). Additionally, a definition of the extracted properties is supplied. We have generated realistic innocuous background traffic by profiling the abstract behavior of human interactions using our recommended B-Profile technology. For this dataset, we modeled the abstract behavior of 25 users utilizing the HTTP, HTTPS, FTP, SSH, and email protocols.	√

[76]	CIC-IDS2018	2018	Network and system logs are included. Network traffic and log data were taken over an eight-day period in a large simulated network environment for CSE-CIC-2018, whereas CICIDS2017 used traffic collected over a five-day period in a small-scale simulated network (the victim organization has five departments and includes 420 machines and 30 servers; the attacking infrastructure includes 50 machines). In addition to packet and flow-based entries, both datasets include 80 attributes culled from recorded traffic. Brute force, Heartbleed, botnets, DDoS, online attacks and network penetration from the inside are all depicted in the labeled network data set.	√
[57]	Unified Host and Network Data Set	2018	Data from the Los Alamos National Laboratory's corporate network is aggregated into the Unified Host and Network Dataset, which is a subset of these events. As a precautionary measure, LANL's working IT environment has been secured by anonymizing the data values (anonymized). There are unique identities assigned to both a host and a network, making it possible to examine and investigate both at the same time. Values like well-known network ports, system-level user names (not linked to specific individuals), and system-level hosts could not be deidentified. Replicated services like as Active Directory servers, email servers, and automated vulnerability detection systems were included in some circumstances.	☒

6. Detection And Prevention From APT

Cyber security researchers have proposed a few important research ideas in the battle against APT that they say are state-of-the-art and complete tactics [33]. However, a closer investigation shows that the proposed architecture is unintelligible and incapable of adapting to the complex and ever-changing cyber threat scenario [23]. Table 9 shows a complete examination of various approaches to APT detection and prevention strategies proposed by renowned researchers. To present a thorough view of the APT threat landscape, we summarized the strengths and weaknesses of each technique.

Table 9: Famous Techniques Prevention APT.

RF	Techniques Name	Working Way	Limited
[10]	Honey-pot systems	They are dummy systems or servers that are deployed alongside the production systems on your network. When utilized as enticing targets for attackers, honeypots may provide extra security monitoring opportunities for blue teams while diverting the adversary's attention away from their intended target.	● Methodology for detecting post-infiltration. ● just typical cyber-attacks are logged. ● There are no real-time detection and prevention systems.
[67]	Intrusion Kill Chains for APT Detection (IKC)	Examining system event logs and comparing them to IKS	● The passive approach is primarily concerned with post-infiltration detection techniques. ● No preventative action in real time ● Detection effort is time demanding ● Probability of a false-positive result is rather high.

[14]	Big data analytics	A good method for detecting APTs is to use big data analysis. The vast volume of data to sift through in search of anomalies is a hurdle in detecting APTs. That is, analysis of the network's flow topology for the purpose of pattern matching.	● Methodology for detecting infiltration after it has occurred. ● Human analysts collaborate to conduct analysis of critical threats, which may be ineffective and result in a greater rate of false positives. ● There is no protection in real time.
[43]	Mechanisms of collaboration for security	Implementing an Open Source SIEM (OSSIM) system to monitor malware activity	● Methodology for detecting infiltration after it has occurred. ● Keep an eye out for any unusual access techniques to system software DLLs that might cause the deployed application to do additional work and hence be less efficient. ● There is no way to avert an attack in real time.
[24]	Context-based framework	To improve the user experience of an existing medical information system, a context-based information system is devised and a mobile application is developed. It is not possible to gain access to the inner workings of a real-world medical information system, hence it is not used. Additionally, the signature and policy are matched to different attack occurrences.	● After identification of infiltration, a passive technique is used. ● APT assaults are not detected and prevented in real time.
[62]	APT attack detection using attack intelligence	Pattern matching between behavior and event, deep packet inspection	● Inability to update the database due to a lack of APT datasets ● Methodology for detecting invasion after it has occurred

7. Learning Techniques in APT

Labeled data is obtained when the correct answer to a data-related question is known. Data that isn't labeled is generated when the right answer isn't known with certainty [29]. The ability of machine learning algorithms to learn from available data is what gives them their power. There are two types of machine learning models [48]: Supervised learning and unsupervised or semi-supervised learning. In figure 3, showing the type:

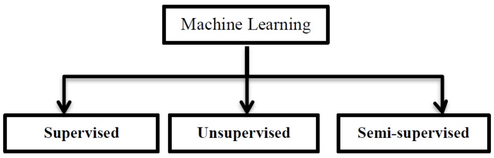

Figure 3: Type Of Machine Learning [29].

These types can be summarized from the following table 10, which shows the difference between the two most important types:

Table 10: supervised vs. Unsupervised [11, 19, 31].

Supervised	**Unsupervised**
Definition: Data sets that have been tagged are employed in this form of learning. Using these datasets, algorithms may be trained or ”supervised” such that they are able to correctly identify data or predict future outcomes. The model's correctness can be checked, and it may be improved over time, using clearly defined inputs and outputs.	**Definition**: Machine learning techniques are used to examine and cluster unlabeled data sets. Without the help of humans, these algorithms discover hidden patterns in data (thus the term ”unsupervised”).
Types of problems: ● Classification: To accurately assign test data into certain groups, such as distinguishing apples from oranges, problems utilize an algorithm. A supervised learning algorithm may also be used in the real world to classify spam and transfer it to a different folder from your inbox. Classification methods include, for example, support vector machines, decision trees, and random forests. ● Regression: In this case, an algorithm is employed to determine the link between the dependent and independent variables. A regression model is a useful tool for predicting numerical values, such as a company's sales revenue estimates, from a variety of data sources. Linear regression, logistic regression, and polynomial regression are all popular methods for doing regression analyses.	**Types of problems**: ● Clustering: is a data mining approach that uses similarities and differences to classify unlabeled data When using K-means clustering, for example, the number K represents the size and granularity of the groups formed by related data points. This method is useful for a wide range of tasks, including market segmentation, image compression, and many more. ● Association: Many criteria are used to find correlations between variables in the dataset in an unsupervised learning strategy. These approaches are often used in market basket analysis and recommendation engines, such as recommendations based on ”Customers Who Bought This Item Also Bought.” ● Dimensionality reduction: This learning approach is used when the number of features (or dimensions) in a dataset is too large. The data is protected while the quantity of inputs is kept to a bearable level. Autoencoders, for example, use this technology to remove noise from visual input in order to improve picture quality.
Goals: The purpose of supervised learning is to anticipate outcomes given new data. You are already aware of the kind of outcomes you can expect.	**Goals**: Unsupervised learning techniques are used to extract valuable information from large amounts of new data. Computers use machine learning to identify what is unusual or intriguing in data sets.
Applications: Many different applications for supervised learning models may be found, but a few examples include spam detection and sentiment analysis.	**Applications**: Recommendation engines and customer personas are great examples of uses for unsupervised learning.
Complexity: Supervised learning is a straightforward machine learning technique that is often computed using programming languages such as R or Python.	**Complexity**: To deal with huge amounts of unclassified data, you'll require advanced unsupervised learning methods. Because they require a large training set to get desired results, unsupervised learning models are computationally intensive.
Drawbacks: It takes time to train supervised learning models, and the labels for input and output variables require knowledge.	**Drawbacks**: Without human interaction to check the output variables, unsupervised learning algorithms can produce radically erroneous findings.

Labeled data: Predictions on the data are generated frequently and the algorithm ”learns” from the training dataset by adjusting for the proper answer. Despite the fact that supervised learning models are more accurate than unsupervised learning models, they still require human input to accurately identify the data. So, based on the time of day and the weather, a supervised learning model can estimate how long your journey will take. There are a few steps before this can be accomplished, though.	**Labeled data**: On the other hand, unsupervised learning models operate independently to reveal the structure of unlabeled data. Notably, verifying output variables still requires human interaction. Online buyers frequently buy many goods at once when using an unsupervised learning model. A data analyst, on the other hand, would have to verify that a recommendation engine can use items like baby outfits, diapers, applesauce, and sippy cups.

Semi-supervised learning, on the other hand, is a machine learning technique that combines the two previous approaches. Despite the fact that data scientists may give an algorithm highly labeled training data, the model is free to explore the data and develop an understanding of the set [3].

Cluster analysis is the most common non-supervised learning method, and it is used to evaluate exploratory data in order to identify hidden patterns or categorize them into data. The clusters are built using a similarity measure expressed in terms of Euclidean or probabilistic metrics [48].

Machine learning and deep learning [1] are currently being used to identify pictures, audio processing, APR and other information processing applications [7], Additionally The most famous learning algorithms can be summarized in following figure 4:

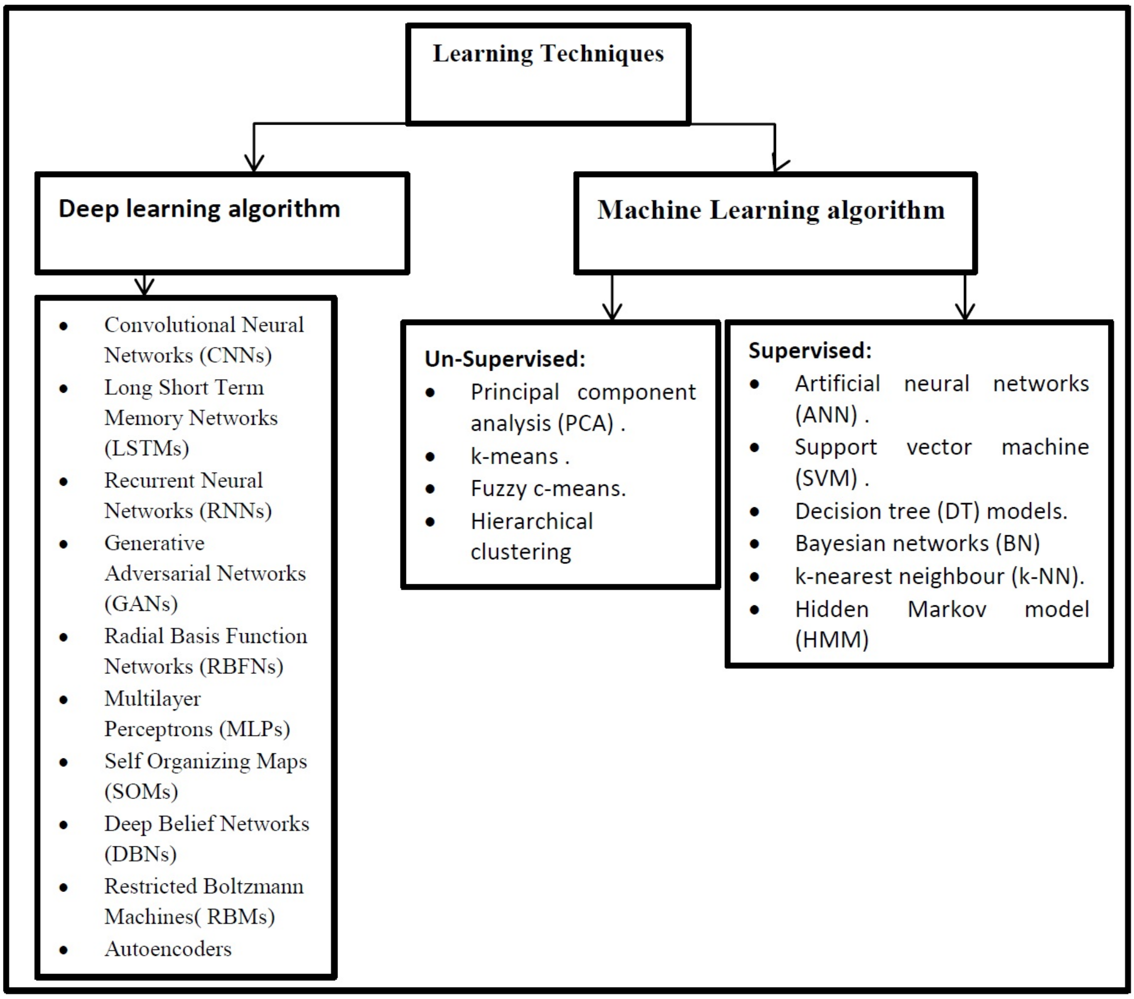
Learning Techniques
Deep learning algorithm
Machine Learning algorithm
Convolutional Neural Networks (CNNs)
Long Short Term Memory Networks (LSTMs)
Recurrent Neural Networks (RNNs)
Generative Adversarial Networks (GANs)
Radial Basis Function Networks (RBFNs)
Multilayer Perceptrons (MLPs)
Self Organizing Maps (SOMs)
Deep Belief Networks (DBNs)
Restricted Boltzmann Machines(RBMs)
Autoencoders
Un-Supervised:
Principal component analysis (PCA) .
k-means .
Fuzzy c-means.
Hierarchical clustering
Supervised:
Artificial neural networks (ANN) .
Support vector machine (SVM) .
Decision tree (DT) models.
Bayesian networks (BN)
k-nearest neighbour (k-NN).
Hidden Markov model (HMM)

Figure 4: Classification Of Machine Learning And Deep Algorithms [29, 46, 59]

8. Comparing Previous Studies

In this section, we will review the most important previous works of this type (survey OR review) about the position of APT in a brief and concise form about what they achieved and what new ideas they presented in table (11):

Table 11: Comparing previous studies.

RF	Year	Present dataset	Present application	Present attack	Present algorithm	Contribution
[61]	2015	☒	☒	√	☒	They established a taxonomy of both APT mechanisms and cyber espionage consequences. The taxonomy clarifies the various components of a developing problem and demonstrates that cyber espionage knows no bounds.
[51]	2016	☒	√	√	☒	The purpose of this article is to illustrate the basic attack patterns of an APT as well as the defense tactics used to counter APT attacks. As a result, readers will see that current protection methods are not completely ready to deal with such well-organized attacks.
[42]	2017	☒	☒	√	√	This article provides an overview of the life cycle of advanced persistent threats (APT) assaults as defined by security experts, which means a developed life cycle model guided by attackers' objectives rather than their activities.
[2]	2018	☒	☒	√	☒	The purpose of this article is to show twenty-five researchers highlight 12 mitigating approaches for defending information systems against APT.
[5]	2019	☒	☒	√	√	
[52]	2020	√	☒	√	√	The generation of APT datasets for automated detection systems is the topic of this paper. Based on the underlying architecture, it covers two use cases.
	Proposed survey	√	√	√	√	Presentation of a scientific paper that contains all the main details of the subject (APT), which is the concept of its life cycle, components, applications, algorithms, types, official sponsorship, datasets, attacks, and the details it contains in a brief and summary form for referential purposes.

9. Conclusion

Resolute, persistent, and well-financed attackers with the purpose of obtaining essential data or disrupting critical components of their target company are known as Advanced Persistent Threats. These attacks, unlike targeted attacks, use sophisticated tools and/or strategies. We have provided the reader with a comprehensive overview of APT and information on how APTs work in classifying APT's defensive techniques, which have included monitoring, detection, and mitigation. We also provide technical background on current APT detection and mitigation procedures, evaluation procedures to measure the efficiency of APT's defensive strategies, classification of the most important dataset used, as well as APT's official sponsor and country from which it came. He investigated it and made a comparison between these scientific papers, noting that this field is very broad and includes a variety of topics, the most recent of which was the Corona pandemic and its epidemic. Impact on patient information theft. It is an ideal environment for researchers to delve deeper into the topic and that APT is constantly evolving and that there are many types of us that categorize malicious attacks on the basis of or among them on the basis of surveillance.

References

[1] D.A. Abdullah, *Objective flow-shop scheduling using PSO algorithm*, Diyala J. Pure Sci. 1 (2013) 140-–153.

[2] O. Adelaiye, A. Ajibola and S. Faki, *Evaluating advanced persistent threats mitigation effects: A review*, Int. J. Inf. Secur. Sci. 7(4) (2018) 159—171.

[3] D. Ahfock and G.J. McLachlan, *Semi-supervised learning of cassifiers from a statistical perspective: A brief review*, arXiv, (2021) 1-–25.

[4] A. Aleroud and L. Zhou, *Phishing environments, techniques, and countermeasures: A survey*, Comput. Secur. 68 (2017) 160—196.

[5] A.K. Al Hwaitat, S. Manaseer and R.M.H. Al-Sayyed, *A survey of digital forensic methods under advanced persistent threat in fog computing environment*, J. Theor. Appl. Inf. Technol. 97(18) (2019) 4934—4954.

[6] S. Al Salami, J. Baek, K. Salah and E. Damiani, *Lightweight encryption for smart home*, Proc. - 2016 11th Int. Conf. Availability, Reliab. Secur. ARES (2016) 382-–388.

[7] A. Alshamrani, S. Myneni, A. Chowdhary and D. Huang, *A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities*, IEEE Commun. Surv. Tutorials 21(2) (2019) 1851—1877.

[8] G. Austin, *Grading National Cybersecurity*, Springer Handbooks, 2018.

[9] M. Auty, *Anatomy of an advanced persistent threat*, Netw. Secur. 2015(4) (2015) 13-–16.

[10] H. Bari, *Protecting an Enterprise Network through the Deployment of Honeypot*, Bangladesh University, Post Graduate Thesis, 2021.

[11] M.J. Baxter, *A review of supervised and unsupervised pattern recogniton in archaeometry*, Archaeometry 48(4) (2006) 671-–694.

[12] A. Berady, V.V.T. Tong, G. Guette, C. Bidan and G. Carat, *Modeling the operational phases of APT campaigns*, Int. Conf. Comput. Sci. Comput. Intell. 2019, pp. 96–101.

[13] G. Brogi and V.V.T. Tong, *TerminAPTor: Highlighting advanced persistent threats through information flow tracking*, 8th IFIP Int. Conf. New Technol. Mobil. Secur. NTMS, 2016, pp. 1-–6.

[14] A.A. Cardenas, P.K. Manadhata and S.P. Rajan, *Big data analytics for security*, IEEE Secur. Privacy 11 (2013) 74–76.

[15] J. Chen, C. Su, K.H. Yeh and M. Yung, *Special issue on advanced persistent threat*, Futur. Gener. Comput. Syst. 79 (2018) 243-–246.

[16] C. Çınar, M. Alkan, M. Dörterler, I.A. Doğru, *A study on advanced persistent threat*, 3rd Int. Conf. Comput. Sci. Eng. (UBMK), 2018, pp. 116—121.

[17] N. De, *Advanced Persistent Threats*, 2015.

[18] B. Dimitrios, *APT Methods for Passive and Active Portfolio Management*, Msc Thesis in Banking and Financial Management, University of Piraeus, 2002.

[19] O. El Aissaoui, Y.E.A. El Madani, L. Oughdir and Y. El Allioui, *Combining supervised and unsupervised machine learning algorithms to predict the learners' learning styles*, Procedia Comput. Sci. 148 (2019) 87-–96.

[20] E. Etuh, F.S. Bakpo and E.A. H, *Social media network attacks and their preventive mechanisms: A review*, CoRR (2021) 59—74.

[21] N. Falliere, L.O. Murchu and E. Chien, *W32. stuxnet dossier: Symantec security response*, Symantec Secur. Response, Version 1.4, (2011) 1—69.

[22] H. Geng, G. Geng, X. Gao and J. Ma, *Dynamic defense strategy against advanced persistent threat with insiders*, Trans. Nonferrous Met. Soc. China 5(3) (2015) 113-–118.

[23] I. Ghafir, M. Hammoudeh, V. Prenosil, L. Han, R. Hegarty, K. Rabie and F.J. Aparicio-Navarro, *Detection of advanced persistent threat using machine-learning correlation analysis*, Future Gen. Comput. Syst. 89 (2018) 349–359.

[24] P. Giura and W. Wang, *A context-based detection framework for advanced persistent threats*, Proc. 2012 ASE Int. Conf. Cyber Secur. CyberSecurity, 2012, pp. 69—74.

[25] W. Han, J. Xue, Y. Wang, F. Zhang and X. Gao, *APTMalInsight: Identify and cognize APT malware based on system call information and ontology knowledge framework*, Inf. Sci. 546 (2021) 633—664.

[26] M.M.H. Henchiri and S. Wani, *Innovative architectural framework design for an effective machine learning based APT detection*, Int. J. Digital Inf. Wireless Commun. 11(1) (2021) 12—22.

[27] M. Hund, *ASEAN plus three: Towards a new age of pan-East Asian regionalism? A skeptic's appraisal*, Pacific Rev. 3(16) (2013) 383-–417.

[28] S. Hussain, M. Bin Ahmad and S.S.U. Ghouri, *Advance persistent threat–A systematic review of literature and meta-analysis of threat vectors*, Adv. Intell. Syst. Comput. 1158 (2021) 161-–178.

[29] C. Janiesch, P. Zschech and K. Heinrich, *Machine learning and deep learning*, Electron. Mark. 31(3) (2021) 685—695.

[30] I. Jeun, Y. Lee and D. Won, *A practical study on advanced persistent threats*, Commun. Comput. Inf. Sci. 339 (2012) 144-–152.

[31] W. Jiang, J. Chen, X. Ding, J. Wu, J. He and G. Wang, *Review summary generation in online systems: frameworks for supervised and unsupervised scenarios*, ACM Trans. Web 15(3) (2021) 1-–33.

[32] J.H. Joloudari, M. Haderbadi, A. Mashmool, M. Ghasemigol, S.S. Band and A. Mosavi, *Early detection of the advanced persistent threat attack using performance analysis of deep learning*, IEEE Access 8 (2020) 186125-–186137.

[33] A. Khalid, A. Zainal, M.A. Maarof and F.A. Ghaleb, *Advanced persistent threat detection: A survey*, 3rd Int. Cyber Resilience Conf. (CRC), IEEE, 2021, pp. 1–6.

[34] M.B. Khan, *Advanced persistent threat: Detection and defense*, arXiv, (2020).

[35] K. Krombholz, H. Hobel, M. Huber and E. Weippl, *Advanced social engineering attacks*, J. Inf. Secur. Appl. 22 (2015) 113-–122.

[36] A. Lemay, J. Calvet, F. Menet and J.M. Fernandez, *Survey of publicly available reports on advanced persistent threat actors*, Comput. Secur. 72 (2018) 26—59.

[37] M. Li, W. Huang, Y. Wang, W. Fan and J. Li, *The study of APT attack stage model*, IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. ICIS 2016, Proc. 2016, pp. 1—5.

[38] P. Li, X. Yang, Q. Xiong, J. Wen and Y.Y. Tang, *Defending against the advanced persistent threat: An optimal control approach*, Secur. Commun. Networks, 2018 (2018).

[39] S. Li, Q. Zhang, X. Wu, W. Han and Z. Tian, *Attribution classification method of APT malware in IoT using machine learning techniques*, Secur. Commun. Networks, 2021 (2021).

[40] R.P. Lippmann, R.K. Cunningham, D.J. Fried, I. Graf, K.R. Kendall, S.E. Webster and M.A. Zissman, *Results of the DARPA 1998 offline intrusion detection evaluation*, MIT Lincoln Laboratory, (1999).

[41] P. Mahadevan, *Cybercrime threats during the COVID-19 pandemic*, The Global Initiative Against Transnational Organized Crime, (2020).

[42] B.I.D. Messaoud, K. Guennoun, M. Wahbi and M. Sadik, *Advanced persistent threat: New analysis driven by life cycle phases and their challenges*, 2016 Int. Conf. Adv. Commun. Syst. Inf. Secur. ACOSIS 2016 - Proc. 2017, pp. 1–6.

[43] N.A.S. Mirza, H. Abbas, F.A. Khan and J. Al Muhtadi, *Anticipating advanced persistent threat (APT) countermeasures using collaborative security mechanisms*, Proc. - 2014 Int. Symp. Biometrics Secur. Technol. ISBAST 2014, (2015) 129—132.

[44] M. Nour, *The UNSW-NB15 Dataset*, UNSW Canberra, 2021.

[45] V. Prenosil and I. Ghafir, *Advanced persistent threat attack detection: An overview*, Int. J. Adv. Comput. Netwo UBMK 2018 - 3rd Int. Conf. Comput. Sci. Eng.rks Its Secur. 4(4) (2014).

[46] S. Quintero-Bonilla and A.M. del Rey, *A new proposal on the advanced persistent threat: A survey*, Appl. Sci. 10(11) (2020).

[47] M. Rakhi and R. Patel, *A review on detecting APT malware infections based on traffic analysis and DNS*, Int. J. Trend Res. Dev. 2(5) (2015) 149—153.

[48] B. Sabir, F. Ullah, M.A. Babar and R. Gaire, *Machine learning for detecting data exfiltration*, ACM Comput. Surv. 54(3) (2021) 1-–32.

[49] S.C. Satapathy, K.S. Raju, J.K. Mandal and V. Bhateja, *Proceedings of the Second International Conference on Computer and Communication Technologies: IC3T 2015*, Springer Link, 2016.

[50] S. Sibi Chakkaravarthy, D. Sangeetha and V. Vaidehi, *A survey on malware analysis and mitigation techniques*, Comput. Sci. Rev. 32 (2019) 1—23.

[51] M.A. Siddiqi, A. Mugheri and K. Oad, *Advance persistent threat defense techniques: A review*, pjcis J. 1(2) (2016) 53-–65.

[52] B. Stojanović, K. Hofer-Schmitz and U. Kleb, *APT datasets and attack modeling for automated detection methods: A review*, Comput. Secur. 92 (2020) 101734.

[53] T.N. Sun, C. Teodorov and L. Le Roux, *Operational design for advanced persistent threats*, Proc. - 23rd ACM/IEEE Int. Conf. Model Driven Eng. Lang. Syst. Model. 2020 - Companion Proc., (2020) 362—371.

[54] P.S. Suryateja, *Threats and vulnerabilities of cloud computing: A review*, Int. J. Comput. Sci. Eng. 6(3) (2018) 297-–302.

[55] Y. Tanaka, M. Akiyama and A. Goto, *Analysis of malware download sites by focusing on time series variation of malware*, J. Comput. Sci. 22 (2017) 301—313.

[56] C. Tankard, *Advanced persistent threats and how to monitor and deter them*, Netw. Secur. 2011(8) (2011) 16-–19.

[57] M.J. Turcotte, A.D. Kent and C. Hash, *Unified host and network data set*, In Data Science for Cyber-Security, (2019) 1–22.

[58] M. Ussath, D. Jaeger, F. Cheng and C. Meinel, *Advanced persistent threats: Behind the scenes*, In 2016 Ann. Conf. Info. Sci. Syst. (CISS), IEEE, (2016) 181–186.

[59] R. Wagner, M. Fredrikson and D. Garlan, *An advanced persistent threat exemplar*, CARNEGIE-MELLON UNIV. PITTSBURGH PA PITTSBURGH United States, (2017).

[60] X. Wang, K. Zheng, X. Niu, B. Wu and C. Wu, *Detection of command and control in advanced persistent threat based on independent access*, 2016 IEEE Int. Conf. Commun. ICC 2016, (2016).

[61] G. Wangen, *The role of malware in reported cyber espionage: A review of the impact and mechanism*, Info. 6(2) (2015) 183-–211.

[62] K. Xing, A. Li, R. Jiang and Y. Jia, *A review of APT attack detection methods and defense strategies*, Proc. -2020 IEEE 5th Int. Conf. Data Sci. Cyberspace, DSC 2020, (2020) 67—70.

[63] C.D. Xuan, M.H. Dao and H.D. Nguyen, *APT attack detection based on flow network analysis techniques using deep learning*, J. Intell. Fuzzy Syst. 39(3) (2020) 4785-–4801.

[64] L.X. Yang, K. Huang, X. Yang, Y. Zhang, Y. Xiang and Y.Y. Tang, *Defense against advanced persistent threat through data backup and recovery*, IEEE Trans. Netw. Sci. Eng. 8(3) (2021) 2001—2013.

[65] Z.S.B. Zainudin, *A Case Study Of Advanced Persistent Threats on Financial Institutions in Malaysia*, Msc thesis, International Islamic University Malaysia, 2017.

[66] Z.S. Zainudin and N.N.A. Molok, *Advanced persistent threats awareness and readiness: A case study in Malaysian financial institutions*, Proc. 2018 Cyber Resil. Conf. CRC 2018, (2018) 1—3.

[67] R. Zhang, Y. Huo, J. Liu and F. Weng, *Constructing APT attack scenarios based on intrusion kill chain and fuzzy clustering*, Secur. Commun. Networks, 2017 (2017).

[68] G. Zhao, K. Xu, L. Xu and B. Wu, *Detecting APT malware infections based on malicious DNS and traffic analysis*, IEEE Access 3 (2015) 1132-–1142.

[69] Z. Zulkefli, M. Mahinderjit-Singh and N. Malim, *Advanced persistent threat mitigation using multi level security access control framework*, Lecture Notes in Computer Science, 2015.

[70] Kdd99, “kdd99,” kdd, 1998. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[71] UNIBS, “UNIBS,” 2011. http://netweb.ing.unibs.it/ ntw/tools/traces/.

[72] NSL-KDD, “NSL-KDD,” NSL-KDD, 2015. https://www.unb.ca/cic/datasets/nsl.html.

[73] NGIDS-DS, “No Title,” NGIDS-DS, 2017. https://research.unsw.edu.au/people/professor-jiankun-hu.

[74] TRAbID, “TRAbID,” 2017. https://secplab.ppgia.pucpr.br/trabid.

[75] “CIC-IDS2017,” CIC-IDS2017. https://www.unb.ca/cic/datasets/ids-2017.html.

[76] CIC-IDS2018, “CIC-IDS2018,” 2018. https://www.unb.ca/cic/datasets/ids-2018.html.