



APT-Dt-KC: advanced persistent threat detection based on kill-chain model

Maryam Panahnejad¹ · Meghdad Mirabi¹

Accepted: 8 November 2021 / Published online: 12 January 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Advanced persistent threat attacks are considered as a serious risk to almost any infrastructure since attackers are constantly changing and evolving their advanced techniques and methods. It is difficult to use traditional defense for detecting the advanced persistent threat attacks and protect network information. The detection of advanced persistent threat attack is usually mixed with many other attacks. Therefore, it is necessary to have a solution that is safe from error and failure in detecting them. In this paper, an intelligent approach is proposed called “APT-Dt-KC” to analyze, identify, and prevent cyber-attacks using the cyber-kill chain model and matching its fuzzy characteristics with the advanced persistent threat attack. In APT-Dt-KC, Pearson correlation test is used to reduce the amount of processing data, and then, a hybrid intrusion detection method is proposed using Bayesian classification algorithm and fuzzy analytical hierarchy process. The experimental results show that APT-Dt-KC has a false positive rate and false negative rate 1.9% and 3.6% less than the existing approach, respectively. The accuracy and detection rate of APT-Dt-KC has reached 98% with an average improvement of 5% over the existing approach.

Keywords Advanced persistent threat · Cyber-kill chain · Bayesian classification · Fuzzy clustering · Support vector machine

✉ Meghdad Mirabi
meghdad.mirabi@gmail.com

Maryam Panahnejad
mahdaneh.panahnejad@gmail.com

¹ Department of Computer Engineering, Faculty of Engineering, Islamic Azad University, South Tehran Branch, Tehran, Iran

1 Introduction

With the growth and development of cyber threats and attacks in computer networks, the need for modern methods to detect intrusion in these networks and defense against cyber threats and attacks has become a significant challenge in computer systems [1, 2].

In general, intrusion detection systems are responsible for identifying and detecting any unauthorized use of resources in the system, as well as abuse or harm caused by both internal and external users. Detection and prevention of intrusion is considered as one of the main mechanisms in providing the security of various computer systems and it is generally used along with firewalls as a complement [3, 4].

Generally, an intrusion detection system can be classified based on intrusion detection method, architecture, and type of intrusion response in the network [4–6]. Basically, intrusion detection systems include three general functions: monitoring and evaluating, detecting, and responding to the security threats. Moreover, different types of intrusion detection methods include abnormal behavior detection and abuse detection (i.e., signature-based detection). There are also several types of architectures for intrusion detection systems that can be classified into three classes: host-based (HIDS), network-based (NIDS), and distributed (DIDS) intrusion detection systems. While a firewall allows traffic to pass through its destination, intrusion detection technologies perform complex analyses on network threats and vulnerabilities, and they can detect and remove attacks that are within the network's legal traffic passing the firewall [7]. Given that there are intrusion prevention approaches, these methods may not always be able to resist against all available cyber-attacks [8]. The task of intrusion detection systems in computer systems is to detect threats and anomalies and notify such anomalies to system administrators to take the appropriate actions to eliminate and prevent them. The simplest form of intrusion detection systems monitors a variety of heterogeneous resources in the system and collects data to examine them to detect anomalies and threats, including data collection from the system and monitoring different types of heterogeneous resources [9, 10].

Among the existing cyber security attacks, advanced persistent threat (APT) attack is one of the newest and modern cyber security attacks that has sacrificed many individuals and organizations. This attack includes a set of complex and long-term actions taken against specific individuals, organizations, or companies [11]. In a typical cyber security attack, an attacker tries to quickly enter the network, steal information, and exit the network so that the intrusion detection systems are less likely to detect such an attack. However, in an APT attack, the goal is to achieve the continuous access to the system and data. To intrude into computer systems using unknown access, an attacker must constantly rewrite the codes and apply sophisticated escape methods. Some complex APT attacks require full-time monitoring and management. Therefore, this usually attacks defense and financial organizations that have confidential and sensitive data and information. APT attacks usually infiltrate into the network through a spear phishing attack which is a social engineering style attack. The next step in this attack is to find a valid username and password to log in to software systems [12–14].

In general, an APT attack involves complex efforts carried out by a group of hackers, focusing on a single goal. The aim of these efforts is to penetrate undetectable software systems for a long time with the lowest level of tracking [11, 14, 15]. It must be known that a single technology or process cannot stop APT attacks, and traditional security methods are not able to defend against them. Therefore, different layers of defense, information about the threat, and advanced skills are needed to counter APT attacks.

The cyber-kill chain is a continuous model and process that demonstrates how activities related to malicious targets are carried out. This chain penetrates computer networks as a specific sequence, and if any of the steps in this chain is blocked, the attack fails [16]. In this paper, an APT detection approach based on Kill Chain model is proposed called “APT-Dt-KC” for detecting APT attacks by adapting the cyber-kill chain model with the fuzzy features of APT attacks. In fact, this paper proposes a more effective, accurate, and as early as possible approach to detect an APT attack by classifying security alerts based on the cyber-kill chain model. In this regard, Pearson correlation algorithm is used to correlate and pre-process a wide range of data. Then, Bayesian algorithm is applied to train and evaluate the threshold values, and finally, the fuzzy prioritization is exploited as a fuzzy analytic hierarchy process to detect and classify all types of attacks. Therefore, the main contributions of this paper are summarized as follows:

- We apply Pearson correlation test in APT-Dt-KC to correlate and preprocess data to reduce the amount of processing data in the process of detecting an APT attack. To the best of our knowledge, this method has not yet been used to pre-process a wide range of data in the detection of APT attacks based on the cyber-kill chain model.
- We exploit a new hybrid approach based on the cyber-kill chain model to detect APT attacks. In APT-Dt-KC, Bayesian classification algorithm is used to train and evaluate threshold values in the classification structure. The fuzzy analytic hierarchy process is also employed to fuzzy prioritization and attacks classification.
- We evaluate the performance of APT-Dt-KC in terms of intrusion detection rate, training time, accuracy, and computational efficiency with DT-EnSVM approach in [17] by performing several experiments using KDD Cup 99 dataset. DT-EnSVM is a support vector-based approach which combines both the ensemble learning and data transformation techniques in the process of APT attack detection.

The rest of paper is organized as follows: In Sect. 2, the concepts and definitions of APT attacks, cyber-kill chain model as well as applied methods in this paper are presented. In Sect. 3, the existing approaches for network intrusion detection, especially techniques for detecting the APT attacks are explained. In Sect. 4, the proposed approach (i.e., APT-Dt-KC) for detecting the APT attacks using the cyber-kill chain model is explained in more details. In Sect. 5, the proposed approach is simulated, and the experimental results are compared with DT-EnSVM approach in

[17]. Finally, in Sect. 6, after a general conclusion, the suggestions are presented as future works.

2 Concepts and definitions

This section describes the concepts and definitions related to the APT attack, the cyber-kill chain model, and the applied techniques in this paper.

2.1 Advanced persistent threat (APT) attack

The APT attack uses a variety of techniques to infiltrate the system and collect sensitive information. In this type of attacks, when an attacker realizes that he has been detected, he changes the method of attack and uses other methods to infiltrate the system [15]. Another feature of this attack is its complexity, which abuses the zero-day vulnerability. This attack utilizes various social engineering techniques to infiltrate the system [18]. A zero-day attack is a computer attack that exploits an unknown vulnerability in the software. This means that the exploiting a zero-day attack (i.e., software that uses a security hole to launch an attack) is used or shared by attackers before the developer of the target software becomes aware of the vulnerability. The zero-day attacks occur during the vulnerability window period. The vulnerability window includes the following steps [15, 18]:

- Developers produce software that has unknown vulnerability.
- Attackers discover vulnerability points before developers take some actions for solving them.
- An attacker writes exploits that are either unknown or known to developers, but these vulnerability points have not yet been completely closed.
- Developers become aware of the vulnerability and programmers try to eliminate the vulnerability by developing some codes.
- Development is free to improve the vulnerability. It means that using plug-in elements and customization of the detection process is applicable to reduce the vulnerability for the service provider and there are no restrictions in this respect.

Conceptually, when an event occurs by a zero-day attack, users who make custom improvements to cover the vulnerability points and effectively cover the damage window are different from those users who readily cover the vulnerability points by the ready-made software. Meanwhile, some users may not be aware of vulnerability points so that these points remain unprotected. Therefore, the length of vulnerability windows depends only on the size of vulnerability points obtained by the user. It is difficult to measure the vulnerability window length because attackers do not report a vulnerability point when they discover it. Developers do not want this information to be released for commercial or security reasons. They may not be aware that they will be attacked on zero-day while repairing the vulnerability. In general, the specific characteristics of an APT attack include achieving targets, using complex techniques, having a good chance

of exploiting zero-day vulnerabilities, intruding the target continuously, and being established for as long as possible [15, 18, 19].

2.2 Cyber kill-chain model

The cyber-kill chain model is developed to define the various stages of a cyber-attack. This model can be used to analyze, detect, and prevent different types of cyber security attacks [20, 21]. In general, there are seven phases of cyber-kill chain model: (1) Reconnaissance, (2) Weaponization, (3) Deliver, (4) Exploitation, (5) Installation, (6) Command and Control, and (7) Actions on objectives [20–25]. As shown in Fig. 1, these seven consecutive steps in the cyber-kill chain provide information on the adversary's tactics, techniques and procedures (TTPs).

2.2.1 Reconnaissance

Reconnaissance is the first phase of the cyber-kill chain model during which an attacker collects the network or endpoint information about a target. The endpoint can be an individual, an organization, or part of a target network's hardware/software. In this phase, the attacker performs hidden investigations about the existence of the target and identifies the potential methods and ways of network failure. Investigations in this phase also provide information on what types of malicious objects can be deployed on the target network, without being detected by cyber-security defense. Furthermore, the backdoor is found in the target network. In addition, attackers determine an appropriate set of intrusion targets located in the target network. If the purpose is to steal personal information, the attacker must identify a way to investigate to establish a bilateral link, so that he can first enter the network, find the information of interest, and then steal it from the outside of network. In the case that the purpose is to destroy the network, he can do it in another way. Regardless of this, the attacker seeks to find the network or users' vulnerabilities for illegal use and access. An example of a unilateral link is spam phishing emails, which, in addition to their traditional purpose, have been stealing credentials to provide malware and sending malicious files as attachments to a specific user. This example shows how an attacker exploits an email service to attack a network. An example of a bilateral link is to find open ports using a port scanner, which after finding an open port, allows authorized bilateral communication through Telnet connections.

2.2.2 Weaponization

The second phase of the kill chain model is weaponization, during which the attacker creates a deliverable malware payload. The attacker uses the information collected



Fig. 1 Lockheed martin cyber kill chain model [22]

during the reconnaissance phase to plan what vulnerabilities should be delivered for exploitation and how it should be delivered for use with the discovered backdoor. There are two types of malware payloads in this phase [23]:

- Malware payloads that do not require to communicate with the attacker, such as viruses and worms.
- Malware payloads that need to communicate with the attacker to receive command and control signals or send the stolen information to the attacker.

These are known as Remote Access Trojans (RAT). The RAT requires both the client and the command-and-control servers. The RAT client is the destination that receives the actual malware payload and is configured to communicate with the command-and-control server. This server is located on the Internet and controlled by the attacker. For example, in the reconnaissance phase, the attacker notices that the email system does not allow sending and receiving *.exe files but allows the.pdf * file in emails of a particular university. On the other hand, the attacker notices that the professors regularly receive and open PDF files by emails from their students. Therefore, the attacker creates a RAT file with the ability to communicate with the command-and-control server and embeds this RAT file in a PDF file, called myCV.pdf, which is sent as an attachment via a phishing email.

2.2.3 Delivery

After the malware payload has been developed and the backdoor for payload reception has been identified, the delivery phase will be performed. Malware can be delivered by tricking or forcing the user to interact with the malware, or it can be delivered automatically by exploiting the weak points of protocols or software packages. For example, an email can have an attachment file to deliver the malware payload. Delivery is an important part of ensuring a successful attack without being detected by the existing security mechanisms. Therefore, enemies design their attacks in such a way that their attacks are not tracked, and they must hide the source of the attack from the security and criminology experts. In addition, enemies use a variety of delivery methods to increase their success rate. It is very difficult to find an exploitative malware that does not require user interaction because they use an inherent defect in the protocol, program, or software to deliver the malware payload. This inherent defect is called software vulnerability and requires software patch to reduce vulnerability [22].

2.2.4 Exploitation

After the successful delivery of the malware payload to the target computer, the exploitation phase begins by installing the malware inside the target computer. The following conditions must be met to begin the malware installation [24]:

- Malware must have the necessary permissions to be installed on the target computer.

- The target computer's operating system or software must be able to install the malware without additional requirements. For example, a malware built for the Linux operating system cannot be installed on a Windows operating system.
- The anti-malware defense of the target computer should not be able to detect this malware; otherwise, the attack will fail due to broken cyber chains.

The exploitation phase does not actually perform the installation, but rather prepares the environment for the installation phase of the cyber-kill chain model. However, this phase is not far from the installation phase, as all installation phases must be performed by the exploitation phase. In order to deliver the malware payload for installation, there must be a form of software or hardware error that the malware payload can be exploited for installation or execution. Such errors are called Common Vulnerabilities and Exposure (CVE).

2.2.5 Installation

Computer infection starts during the installation phase. If the malware is in the form of an executable file or malicious activity based on a code injection or an internal threat, there is no need for installation phase. However, if the malware needs to be installed on the target computer, then the delivery phase should place the dropper or downloader on the destination computer and complete the exploitation phase by disabling security services and finding traps in the operating system to begin malware installation. At this phase, the malware is installed and the installed files support libraries and operating system files or download those files from the downloader and dropper packages. In addition, the malware installation updates the operating system files using the authorized permissions. Malware changes the appearance of their files by changing the file format or hiding the files from user access. Advanced malware can also modify their footprint memory to prevent detection by sandbox algorithms or behavior-based anti-malware systems. The installation phase is not only performed on the backdoor of the target victim but also ensures that the attackers are able to communicate permanently with the victim computer. It should be noted that this phase does not begin to communicate with the command-and-control mechanism for malware activity. Moreover, this phase is different from the exploitation phase, it means that the exploitation phase specifies that the malicious packages are ready to be installed and all the needs of the exploitation phase are met, while in the installation phase, the real malware payload starts to establish a base inside the victim computer in a local manner.

2.2.6 Command and control

The command-and-control phase of the cyber-kill chain model is necessary because of the following reasons [15]:

- To steal information (i.e. passwords, financial data, intellectual property, etc.) from the target computer using some tools such as Key Loggers, Zeus, and Trojan.Coinbitclip.

- To send instructions of the malware to the target computer to connect the malware to other parts of the target computer, execute malware, or enable encryption for ransomware activities.

It should be noted that the endpoint defense mechanisms and network monitoring services play an important role in detecting illegal connection to the network and this phase is the last step to prevent the malicious activities. There are two major types of communication-based command and control servers:

- Servers that contain meta-information about compromise nodes via proactive messages.
- Servers that actively communicate with target nodes by issuing commands to victim node and perform more malicious activities.

In addition, command-and-control servers can be classified into direct and indirect communication classes [15]:

- In direct communication, the malware in the victim node contains a list of command and controls of the IP servers, so that if a particular IP is blocked, the malware communicates with another IP. This feature is called durability.
- In indirect communications, attackers use legitimate intermediary nodes to communicate. A group of nodes is compromised to establish a communication link, while the source of the link is hidden from the victim's view. Thus, a botnet is created to re-establish the communication path from the victim to the source.

Attackers may have different ways to connect to the destination node. They may use email protocols for malware payload delivery. They may utilize one or more HTTP connections to establish an output link. In addition, they can use different compression mechanisms to data exfiltration. The attacker constantly changes his tactics and techniques to escape detection. However, their common feature is network traffic, and if the endpoint security mechanism fails to detect the presence of a communication malware, the network defense can do this [24].

2.2.7 Actions on objectives

This is the last phase of the cyber-kill chain model, and it is responsible for carrying out the attack against targets. In this phase, if the malware is on the target computer, then the execution of the program function starts either through the instructions of the command-and-control server or independently. This phase is also known as the explosive phase that has successfully completed the cyber-kill chain model. The following are the main classes of actions that can be taken in this phase:

- Data Exfiltration: Stealing the confidential information from the network.
- Ransomware: Attackers get hostage the victim information by encryption, as well as block the network resources and demand compensation by changing credits or using encryption.

- **Cyber Terrorism:** Attackers do damage the system by deleting data or destroying files completely.

In general, the cyber-kill chain model is a continuous model that demonstrates how the attacker reaches the target node. This model is based on this hypothesis that attackers attempt to infiltrate computer networks in a sequential, incremental, and advanced manner. In this structured model, if each phase of the cyber-kill chain model is blocked, the attack taken by the attacker will not be successful. In fact, cyber-security experts are looking for early detection of cyber threats with a cyber-kill chain model. Seven consecutive phases in the cyber-kill chain model provide information on enemy's tactics, techniques, and procedures (TTPs) [25]

2.3 Bayesian inference and its relationship with the cyber-kill chain model

Bayesian classification uses Bayes theorem to predict the occurrence of any event. Bayesian classifiers are the statistical classifiers with the Bayesian probability understandings. The accuracy of Bayesian classifier is high, and its value can be significantly increased by using special functions. The learning method in this approach is supervised learning, so that this approach considers the prediction variables independently, therefore, it is called simple Bayes or Naïve Bayes [26].

For simple Bayesian classification, feature attributes separation, including feature identification and their classifications are necessary. After classification, the data forms a set of feature attributes, at the same time learning samples are generated simultaneously. The classification results obtained from the classifiers are mainly determined by the feature attributes and the quality of learning samples.

The behaviors of cyber-attacks can be considered by exploiting Bayesian classification approach. In this structure, the malicious behavior is used to indicate the events of a cyber-attack [11]. When n types of attack behaviors (behavior 1, behavior 2, ..., behavior n) are captured, the probability of an APT attack is greater than or equal to Qn .

Generally, Bayesian inference system architecture consists of the following modules [11]:

- **Alarm Processing Module:** The whole system, which is constantly running, is driven by alerts. Alarm processing module includes alarm receiver.
- **Information Integration and Analysis Module:** This module performs fuzzy search in parallel according to the received alert. Search results are stored in the internal cache. The module performs the corresponding functions, alarm classification, and priority allocation.
- **Bayesian Learning Module:** Learning includes the statistics of the sample and Bayesian engine tuning. The statistics of the sample include previous probabilities called APT and NOAPT, so conditional probabilities of APT attack with single or multiple behaviors are received by the defender. Bayesian engine tuning involves the behavior weight, and the participation of a behavior in the evaluation of APT attack. The learning module not only performs the task of learning,

but also tests the system. The configuration information is given to the Bayesian engine through the continuous test and storage. Information is stored for in-depth analysis in case of discrepancies with the result of detected and predicted system.

- **Bayesian Engine:** After receiving the results from the Bayesian learning module, the Bayesian engine is ready to receive a variety of event alerts, such as the alerts for APT attacks. Many real-time alerts are first collected in the information integration and analysis module. Then, along with the baseline information, alerts are issued by the Bayesian evaluation program. This process divides the events in the attack into three cases: APT cases, threat-less cases, and gray threat cases. Gray event indicates that it cannot be detected by current information and all events are stored in default state. Gray results are predicted according to the thresholds. The threshold value is adjusted after several simulation tests.

2.4 Analytic hierarchy process (AHP)

The analytic hierarchy process (AHP) is a powerful and flexible method among the multi-criteria decision-making methods by which complex problems can be solved at different levels. This is called a hierarchical model since it is a tree model in a hierarchy manner. The AHP method combines both objective and subjective evaluations into an integrated structure based on scales with pair comparisons and helps analysts to organize the essential aspects of a problem into a hierarchical framework [27, 28].

The advantages of this method are as follows: evaluating the consistency of decision makers' judgments, creating pairwise comparisons in choosing the optimal solution, the ability to consider the criteria and sub-criteria in evaluating options, and the ability to achieve the best solution through pairwise comparisons.

Generally, AHP is a way to help the decision-making process and focuses on the importance of a decision maker's intuitive judgments as well as the stability of comparing alternative options in the decision-making process. One of the advantages of this method is that it regularly organizes tangible and intangible factors and offers a structural but relatively simple solution to decision problems [27]. In this method, the decision-making problem is divided into different levels of objectives, criteria, sub-criteria and options. Therefore, various options are involved in decision making and it is possible to analyze the sensitivity of criteria and sub-criteria. Sensitivity analysis means what changes occur in the ranking of options as the weight of the criteria changes. Another advantage of this method is to determine the amount of compatibility and incompatibility of the decision. AHP method simplifies the complex issues by analyzing them [28]. The selection of criteria is the first part of AHP method. Candidates are evaluated based on the identified criteria. The word "alternatives" and "candidates" are used interchangeably. The reason for calling this method hierarchy is that we must first start from the goals and strategies of organization at the top of the pyramid and by expanding them, identify the criteria to reach the bottom of the pyramid. This method is one of the most widely used methods for ranking and determining the importance of factors, which is used to prioritize each of the criteria using pairwise comparisons. It is difficult to form a matrix of pairwise

comparisons if there are many alternatives. The aim of AHP method is to select the best option based on different criteria through pairwise comparison. This method is also used to weigh criteria. Since increasing the number of elements in each cluster makes paired comparisons difficult, decision criteria are usually subdivided into sub-criteria [29].

- Criterion: There is a parameter that is selected as a quality component.
- Option: An item that is selected from the available items.

The following models are used as the most widely used models in the AHP method:

- Goal-Criterion
- Goal-Criterion-Sub-Criterion
- Goal-Criterion-Option
- Goal-Criterion- Sub-Criterion-Option

In the AHP method, you may want to weigh only the criteria. There may be sub-criteria and the goal is to determine the weight of the sub-criteria. In this process, the evaluation of the relative importance of decision criteria and comparison of decision options according to each criterion is performed by pairwise comparisons, which includes the following three steps [29]:

- Create a comparison matrix at each level of the hierarchy, starting from the second level.
- Calculate the relative weights for each element of the hierarchy.
- Estimate the adjustment rate to check the compliance of the arbitration.

2.5 Pearson correlation method

Pearson correlation is a method based on parametric statistics that shows the intensity and direction of the relationship between two variables. This method, like other correlation methods, considers the relationships of variables in pairs. That is, if you measure the relationship between two variables A and B with or without the presence of a variable such as C, the value of this relationship is still the same. In examining the correlation of two variables, if both variables have relative scale, the Pearson moment correlation coefficient is used. If the correlation coefficient of the population ρ and the correlation coefficient of the sample n are the volume n of the population r , r may be obtained randomly. For this purpose, the correlation coefficient significance test is used. This test examines whether two variables are random and independent. In other words, whether the correlation coefficient of the society is zero or not. This coefficient calculates the value of correlation between two distance or relative variables which is between $+1$ and -1 . If the obtained value is positive, it means that the changes in two variables occur in the same direction, namely as each variable increases, so does the other variable; conversely, if the value of r becomes

negative, it means that the two variables also act in the opposite direction; that is, by increasing the value of one variable, the values of the other variable decrease, and vice versa. If the obtained value is zero, it shows that there is no relationship between two variables. In the case that it is $+1$, the correlation is positive and complete, while if it is -1 , the correlation is negative and complete [30].

If the data distribution of the variables is normal, the Pearson correlation coefficient is used to measure the correlation. The Pearson correlation coefficient, which is also known as the Pearson moment correlation coefficient, correlation coefficient and bilateral correlation coefficient, is used to calculate the value and amount of linear relationship between two variables. The range of correlation coefficients varies from $+1$ to -1 . The closer this value is to $+1$, the stronger and positive the relationship between the two variables. In other words, as each of variable increases, the other variables increase, and vice versa. In the case that the closer the value of this coefficient is to -1 , the stronger and negative the relationship between the two variables. In other words, with the increase in each variable, the other variables decrease, and with the decrease in each variable, the other variables increase [30]. Therefore, there are:

- Null hypothesis: The correlation coefficient between two variables is zero.
- Alternative hypothesis: The correlation coefficient between two variables is not zero.

Then, in the proposed method, this coefficient is used to create correlations between the components and while improving the correlation in the data, the level of redundancy is also reduced.

3 Related works

Many intelligence algorithms have been applied to improve the detection capability of an intrusion detection system. Among these methods, ensemble learning has received an increasing interest and shown to obtain better performance than single learning methods. Besides, the intrusion detection performance is also highly dependent on the quality of training data. In [17], an effective intrusion detection framework based on SVM ensemble with feature augmentation is proposed. Specifically, the quality-improved technique is used to provide concise, high-quality training data and SVM ensemble is applied to build intrusion detection model.

Singh et al. [31] have proposed the application of Security Onion (SecOn) to develop the network security monitoring (NSM) and intrusion detection system (IDS) in the context of SCADA cyber physical security. They have applied a cyber kill-chain model to demonstrate the different stages of attacks and associated mechanisms.

In the current development of online social networks and information technologies, the capture of group privacy may lead to individual privacy violations. In this regard, Kim et al. [32] have studied the privacy kill chain, which uses group privacy as a tool to capture the individual privacy. They have shown how the kill chain

makes the need to protect group privacy possible from a social, legal, ethical, commercial, and technical perspectives.

Shameli-Sendi et al. [33] have proposed a new approach for automated response systems to assess the value of the loss that could be suffered by a compromised resource. This approach uses a feedback-based mechanism that can measure the quality of response and provides a great assist to represent the risk level of different applications. Note that the proposed approach uses a new online mechanism to activate and deactivate the responses based on the online risk effect. Furthermore, it can determine the main factors in risk assessment and efficiency calculation with very high complexity.

Duncan et al. [34] have proposed a hybrid method of attack tree and kill chain to identify multiple indirect detection in cloud computing. An attack tree, as defined by Schneier [35], represents attacks as a tree structure in which the root node is the attack target and the leaf nodes are actions or events. In particular, the use of attack trees makes it possible to identify all possibilities of detecting self-attacks in cloud environment. Basically, the attack tree coverage at the top of the kill chain, increases the chances of indirect detection from the tree itself, as well as allowing the cloud provider to determine how much an attack has progressed after a suspicious activity.

Hoffman [36] has proposed to start using Markov processes to model some cyber-attacks. In the proposed method, two example theory models of cyber-attack recursion cycles are presented, which is called the cyber-kill chain model with iterations. The proposed models are based on homogeneous continuous time Markov chains. There are no special steps such as start and end in all available solutions to describe the cyber-attack kill chains. Therefore, a generalized cyber-attack life cycle has recently been proposed including two additional phases [37]. The first step is to identify the attacker's requirements. The last step in the cyber-attack process is to finish the attack along with removing the traces of the attackers' activities. The aim is to provide analytical stochastic models and theory of cyber kill chains with iterations. The steps of modeled cyber-kill chains are understood according to steps given in [38] and [39]. It is also assumed that some phases of cyber-attack processes may be skipped by the attackers; cyber-attacks may be assigned by attackers, or they may be stopped by cyber defense systems at any time, and then the new cyber-attacks may begin. The proposed models are based on homogeneous Markov chains.

In [40], the problem of customizing a dynamic quarantine and recovery (QAR) scheme for an organization is addressed in such a way that the APT impact can be minimized. In this approach, the expected effect of APT under a QAR scheme is estimated based on an epidemic model.

In [41], the decision tree learning methods (i.e., Bayesian network and deep learning) are used to detect and classify different types of APT attacks. The proposed method can improve the detection accuracy by analyzing the data through a deep learning model. Moreover, it can improve deep learning training by testing the existing data through the Maxout method and cross-validation to avoid over-fitting and increase generalizability.

A theoretical framework is built in [42] for describing an APT information-based attack over the internal network. In particular, the mathematical framework of this approach includes an initial input model for selecting entry points

and a targeted attack model for information collection, strategy decision-making, weaponization, and lateral movement.

A new APT defense mechanism known as the DBAR-based APT defense mechanism is proposed in [43], which can overcome the main drawbacks of the DAR-based APT defense mechanism. It is expected that it is efficiently applicable on the Software-Defined Network (SDN) patterns. In [43], the problem is reduced to a differential game problem based on a new dynamic model that describes the expected security situation of the organization network. Therefore, it finds a cost-effective DBAR strategy in terms of Nash equilibrium.

In [44], a multiple machine learning classifier is used to identify APT attacks by matching the alerts at different stages of the cyber-kill chain model. The proposed approach uses feature selection to reduce the impact of this on the overall prediction accuracy to detect APT attacks.

A detection method is proposed in [45] to detect APT attacks based on abnormal behaviors of network traffic using machine learning. Abnormal behavior of APT attacks in network traffic, which includes domain and IP, is evaluated, and classified based on a random forest classification algorithm to draw conclusions about the behavior of APT attacks. This method combines the behavior and characteristics of IP and domain.

Table 1 provides a general comparison of the existing approaches, along with performance measurement criteria, scalability level, detection accuracy, and additional useful features.

It can be expressed that we increase the accuracy, reduce the positive and negative false rates and finally improve the detection rate in our proposed approach (i.e., APT-Dt-KC). Since APT-Dt-KC uses an intrusion detection method based on fuzzy hierarchical structure, it is expected that the detection of attacks will be performed with better accuracy, false positive and false negative rates.

4 The proposed approach

In this section, we describe APT-Dt-KC approach for detecting the APT attacks based on the cyber-kill chain model. The flowchart of our proposed approach is shown in Fig. 2. In general, it is necessary to create a database of real-time alerts. This database is used in the process of detecting the APT attacks in APT-Dt-KC. As shown in Fig. 2, the data preprocessing phase is performed using Pearson correlation test. This method is used to optimize the input information for training by Bayesian algorithm. Next, Bayesian algorithm classifies data based on three components of detection threshold, prediction threshold and gray results. Then, the rest of the preprocessed data is used as experimental components and their outputs are employed as input parameters in the analytical hierarchy process to prioritize attacks. The outputs of this phase include a classification with a known priority for attacks. In the following, we describe these phases in more detail.

Table 1 Comparison of the existing approaches

Ref	Method	Performance evaluation criteria	Scalability	Detection accuracy
Proposed approach	Using Pearson correlation test, Bayesian classification algorithm and fuzzy hierarchical analysis approach for timely identification and classification of APT attacks	Accuracy Detection Rate False Alarm Rate False Positive Rate False Negative Rate Training Time	High	Ideal
[17]	Intrusion detection framework based on SVM ensemble with feature augmentation	Accuracy Detection Rate False Alarm Rate False Positive Rate False Negative Rate	High	High
[31]	Application of SecOn security system for intrusion detection system and network security monitoring along with SCADA cyber physical security	Precision Accuracy	High	Medium
[32]	Investigating the privacy kill chain: protecting online individual and group privacy	Accuracy	Medium	High
[33]	Using a new online mechanism to activate and deactivate responses based on the impact of online risk	Monitoring Overhead Detection Rate	Medium	Medium
[34]	Presenting a hybrid method of attack tree and kill chain to identify multiple indirect detection measures	Speed Rate Detection Rate	High	Medium
[36]	Application of Markov models of cyber killing chains with iterations	Efficiency Detection Rate	Medium	High
[40]	Using QAR scheme for an organization to reduce the impact of APT	Precision Detection Rate	Medium	High
[41]	Applying decision tree learning methods, Bayesian network and deep learning to timely detect and classify APT attacks	Precision Monitoring Efficiency Accuracy False Negative Rate F-Measure	Low	Medium
[42]	Establishing a theory framework to describe an APT information-based attack on the internal network	Precision Accuracy	High	Medium

Table 1 (continued)

Ref	Method	Performance evaluation criteria		
		Detection accuracy	Scalability	Detection accuracy
[43]	Presenting a new APT defense mechanism called DBAR-based APT defense mechanism	Detection Overhead	Medium	Medium
[44]	Using multiple machine learning classifiers such as Naïve Bayes, Bayes Net KNN, Random Forest and Support Vector Machine (SVM)	Latency		
		Precision		
[45]	A method of detecting APT attacks based on abnormal behaviors of Network traffic using machine learning	Accuracy	Low	Medium
		Detection Rate		
		F-measure		
		False Alarm Rate		
		Accuracy	High	Medium
		Precision		
		Recall		
		F-measure		
		False Positive Rate		
		True Negative Rate		
		False Negative Rate		

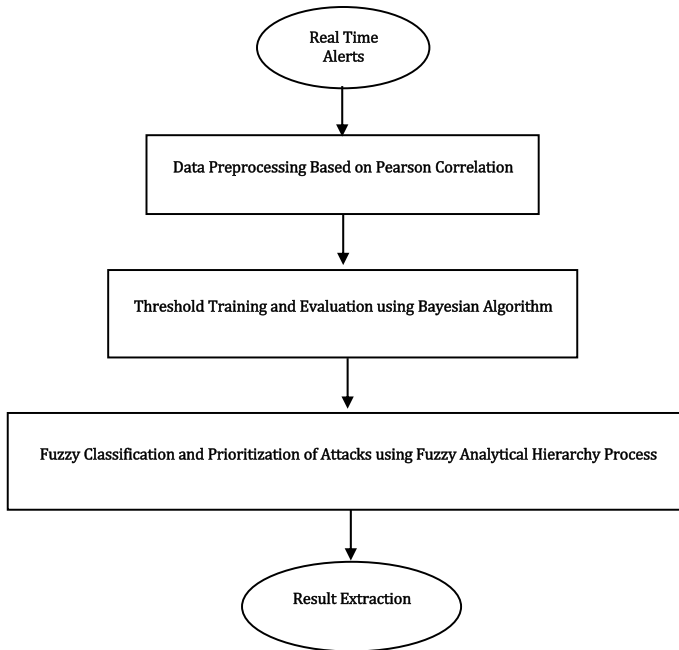


Fig. 2 Flowchart of APT-Dt-KC approach

4.1 Pre-processing and correlation of parameters

In general, different attacks have different evaluation parameters. Therefore, several parameters must be evaluated to detect the APT attacks. The existing studies have shown that various parameters have been introduced for the evaluation and detection of cyber-attacks, which depending on the type and severity of the attacks. Some of these parameters are used for the evaluation of cyber-attacks but the most parameters are left unused. It should be noted that examining the volume of many parameters may increase the overhead and ultimately reduce the efficiency of the proposed solution for intrusion detection [46, 47].

In this paper, the correlation coefficient method is used in the preprocessing and correlation phase to reduce the number of parameters used to detect the APT attacks. The correlation coefficient is a statistical tool to determine the type and degree of relationship between two quantitative variables. Also, it is one of the factors used to determine the correlation of two variables. It indicates the intensity of relationship as well as the type of relationship. This coefficient is between -1 to 1 . In the case that there is no relationship between two variables, it is equal to zero [46].

The correlation between two variables can be measured using a variety of different computational methods. Pearson correlation coefficient, Spearman correlation coefficient, and Tau Kendall correlation coefficient are the most common methods for the calculation of correlation between variables. In general, there are three cases as follows: (1) If both variables are with rank scale, Tau-Kendall correlation coefficient is used. (2)

If both variables are proportional and contiguous scale, Pearson correlation coefficient is used. (3) If both variables are proportional and discrete scale, Spearman correlation coefficient is used. Therefore, due to the correlation between the parameters, the best method is Pearson correlation coefficient. Perhaps the most widely used application of bivariate correlation statistical index is Pearson moment correlation coefficient, commonly called the Pearson correlation and denoted by r [46, 47]. Pearson coefficient shows how much linear relationship exists among quantitative variables. The main application of Pearson correlation coefficient is when the variables are parametric; that is, they have a normal distribution, and they are at a distance/relative level. Meaning of distances and levels are the Euclidean distance between the parameters and the available rates for these parameters. If the ratings for two parameters are on the same level, the distance can be traversed and measured. In general, Pearson correlation coefficient between two random variables equal to their covariance divided by the product of their standard deviations. For a statistical population, the correlation coefficient can be defined as Eq. (1) [48]:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

where cov is covariance, σ_X is the standard deviation of the variable X, σ_Y is the standard deviation of Y, μ_X is the mean of X, μ_Y is the mean of the Y, and finally E represents the mathematical expectation.

Usually, Pearson correlation coefficient for a statistical sample with n data pairs is defined as (X_i, Y_j) using Eq. (2) [7]:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \tag{2}$$

Equation (2) can be summarized as Eq. (3).

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right) \tag{3}$$

where parameters \bar{X} , \bar{Y} , s_X and s_Y represent the mean of X, mean of Y, the standard deviation of X, and the standard deviation of Y, respectively, defined by Eqs. (4), (5), (6) and (7):

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \tag{4}$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \tag{5}$$

$$s_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (6)$$

$$s_Y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (7)$$

where X_i and Y_i are equal to the existing parameters.

After determining the level of correlation and direction, it is necessary to evaluate the intensity of relationship. To interpret this intensity, various classifications are proposed depending on the given applications. These classifications are used to correlate data and remove useless data from the mass of data. By correlating data, a very large portion of useless evaluations for unnecessary data can be removed. Therefore, those parameters used to detect attacks are correlated using the Pearson correlation method, and the unused parameters are discarded. The interpretation of relationship intensity in Pearson correlation is shown in Table 2.

4.2 Threshold training and evaluation using Bayesian algorithm

Generally, Bayesian decision theory is a probable method for reasoning. It is assumed that the given variables follow a certain probabilistic distribution. These probabilities and observed data can be used to make decisions.

The simple Bayesian classification model (NBC) is based on Bayesian decision theory, which is a simple Bayesian probability model. This classification is simple in execution, fast in classification, and high in accuracy and it is one of the most widely used classification models in machine learning [49].

The advantages of this algorithm are (1) Classifying experimental data using this algorithm is easy and fast. (2) When the condition of independence is met; a naïve Bayesian classifier performs better than other models such as logistic regression and requires a low amount of training data.

Given that the data set has K features, it is assumed that values of K features are discrete. The aim of classification is to predict the type of each item within the test set that is part of the data set. Other part is the learning set, whose aim is to make simple Bayesian learning. For a particular sample whose features are in

Table 2 Interpretation of relationship intensity in pearson correlation

Intensity of relationship	Interpretation
0.8–1	Very Strong Relationship
0.6–0.8	Strong Relationship
0.4–0.6	Medium Relationship
0.2–0.4	Low (or weak) Relationship
0–0.2	Lack of Relationship or Insignificant Relationship

the range of a_1 to a_k , the calculated probability for the given sample of class c_i is, $P(C = c_i | A_1 = a_1, \dots, A_k = a_k)$. It is obvious that according to Bayesian decision theory, we have Eq. (8) [49]:

$$P(C = c_i | A_1 = a_1, \dots, A_k = a_k) = \frac{P(A_1 = a_1, \dots, A_k = a_k | C = c_i)P(C = c_i)}{P(A_1 = a_1, \dots, A_k = a_k)} \tag{8}$$

where $P(C = c_i)$ is the prior probability and can easily be considered as a part of learning set. In the given dataset, $P(A_1 = a_1, \dots, A_k = a_k)$ is the same for each class of c_i , and we assume that the values of the features are independent. Therefore, Eqs. (9) and (10) are as follows:

$$P(A_1 = a_1, \dots, A_k = a_k) = 1 \tag{9}$$

$$P(A_1 = a_1, \dots, A_k = a_k | C = c_i) = P(A_1 = a_1 | C = c_i) \dots P(A_k = a_k | C = c_i) \tag{10}$$

By replacing Eqs. (9) and (10) in Eq. (8), the method used in simple Bayesian classification is applied in this paper. In other words, we have Eq. (11):

$$V_{NBC}(x) = \arg \max P(C = c_i) \prod P(A_j = a_j | C = c_i) \tag{11}$$

where $V_{NBC}(x)$ represents the output of the simple Bayesian classification and $\arg \max P(C = c_i)$ is the maximum of prior probability. Also, $P(A_j = a_j | C = c_i)$ is equal to the calculated probability for the given sample of class c_i .

Theoretically, simple Bayesian classification has the lowest incorrect classification rate compared to other classification algorithms [26, 49]. However, it is difficult to assume that the actual network behaviors are independent. In general, each computer network has its own unique features, which can directly affect the results of intrusion detection methods. Therefore, a weighted feature is assigned to the simple Bayesian classification to give different weights to each attribute that affect these relationships in the simple Bayesian classification. Different weights have different results in the simple Bayesian classification, and these weights have many effects on intrusion detection methods. The main point of simple Bayesian classification in an intrusion detection system is the way of determining the weights of different features. The probability calculated in the above topic is used to determine the threshold value. Therefore, the results are expressed in three modes (i.e., gray threshold, prediction threshold, and detection threshold) by evaluating the threshold values. The prediction mode represents the process by which current intrusion information is detected. The detection mode determines the process by which intrusion information is fully identified and detected. The gray mode indicates that intrusion is not detectable with current information.

4.3 Fuzzy analytical hierarchy process

APT-Dt-KC uses a fuzzy AHP [7] to classify different alerts and determine the correlation among parameters to detect an APT attack. AHP uses multiple attribute decision making (MADM) method for decision making and select an option solution among several options. This process reflects human's natural behavior and thought. This method analyzes complex problems, modifies them simply, and then begins to solve them [50].

In this method, input data must be fuzzified at the first. Various membership functions are used for the numerical values fuzzification. One of the common methods is to determine membership functions using the mathematical relations, because in this case neutrality is maintained. Membership functions are defined in different forms, the most common of which are triangular membership function and bell membership function [27]. The advantage of the triangular membership function is that if it is used, more theoretical arguments can be used to prove theories while the most important feature of the bell function is that it is closer to the human thinking. Membership degree $A(x)$ indicates the membership degree of the element x in the fuzzy set A . If the membership degree of an element is zero, that member is completely out of the set while if the membership degree is equal to one, that member is completely in the set. In the case that the membership degree is between zero and one, this number indicates a gradual membership degree. In this paper, the triangular membership function is used for the fuzzification of values. The triangular membership function can be computed using Eq. (12) [28]:

$$\mu A(x) = \begin{cases} 0 & x < a_1 \\ \frac{x-a_1}{a_2-a_1} & a_1 \leq x \leq a_2 \\ \frac{a_3-x}{a_3-a_2} & a_2 \leq x \leq a_3 \\ 0 & x > a_3 \end{cases} \quad (12)$$

where a_1 , a_2 , and a_3 represent the x coordinates in the fuzzy set $A(x)$. The value 0 indicates the lowest membership degree, while the value 1 indicates the highest membership degree for a fuzzy set. It should be noted that the values between a_1 and a_2 show an increase in the membership degree, while the values between a_2 and a_3 represent a decrease in membership degree in a fuzzy set [27].

In general, the hierarchical analysis process includes the following steps:

- AHP Modeling Process: In this step, all the decision-making goals are considered as a hierarchy of decision elements that are related to each other. Decision elements are decision-making criteria and options. Figure 3 shows the structure of AHP in which the desired goals can be achieved using various available options and criteria.
- Creating a pairwise comparison decision matrix: According to each criterion (parameter), the pair comparison matrix of different options is built. Moreover, the pair comparison matrix of criteria is created to obtain the weight of decision-making criteria. In this method, the elements at each level are compared with

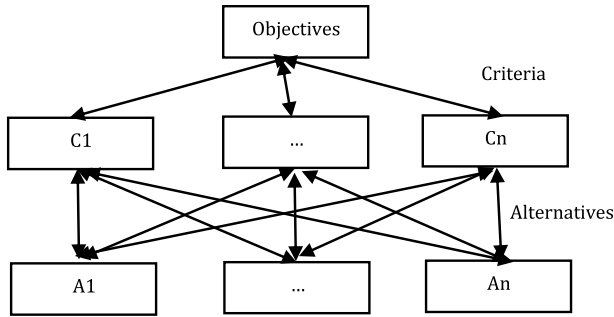


Fig. 3 The structure of AHP

the corresponding elements in the higher level and their weights are calculated. These weights are called the mean weights and they are combined to achieve the final weight of each option. Now, consider a matrix $n \times n$ where n represents the number of criteria. The decision matrix A and the relationship between its elements can be defined using Eqs. (13) and (14), respectively.

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \tag{13}$$

$$a_{ij} = \frac{1}{a_{ji}} \text{ and } a_{ii} = 1 \tag{14}$$

The element of row i and column j in matrix A which is a_{ij} indicates the importance of objective i relative to objective j . This importance is measured with an integer in the range of 1 to 9. Since the importance of each factor relative to itself is one, then the diagonal elements of the matrix of pairwise comparisons are equal to one, and the other elements of the matrix are different and inverse of each other. Therefore, this matrix is a square and invertible matrix in which the importance of the factors relative to each other is determined according to Table 3. The standard preference table can have different modes. It should be noted that increasing the number of modes in this table leads to an increase in computational overhead. Therefore, the selection of these values must be done accurately based on the balance between accuracy and efficiency.

- Calculate the criteria weight and score of each option: Now, a pairwise comparison matrix of criteria is generated (i.e., Matrix A). There are different methods for obtaining the weight of the criteria in this matrix. One of the methods to approximate the weight of the criteria in this matrix is normalization that takes place during the following steps:
 - Step 1: For each column of matrix A , the following operations are performed: The sum of the elements of each column is calculated. Then, each element

Table 3 Standard preference table

Value	Importance
1	Same
3	Slightly Better
5	Better
7	Much Better
9	Completely Better
2, 4, 6, 8	Intermediate values, for example 8 means that the importance of the criterion is between the two states of much better and completely better

of column i is divided by the sum of the elements in the column. Finally, the new matrix is the result of a matrix whose the total elements in each column are equal to one.

- o Step 2: To approximate the weight of each criterion, the average of each row is calculated and the vector $n \times 1$ is obtained. This vector determines the weight of each criteria, denoted by w that indicates the weight of each parameter for testing and comparison. It is a value between 0 and 1, and the sum of that weight must be 1. Therefore, the importance of each option for each criterion is obtained. For example, if a problem contains n criteria and m options, the matrix $n \times m$ is obtained by performing this calculation which indicates the importance of all options for the given criteria.
- o Step 3: This step includes the compatibility test; for this, two parameters of consistency index (CI) and consistency ratio (CR) are involved. These two parameters are calculated using Eqs. (15) and (16), respectively.

$$CI = \frac{X - n}{n - 1} \quad X = \frac{1}{n} \sum_{i=1}^n \frac{(A \times W)_i}{w_i} \tag{15}$$

$$CR = \frac{CI}{RI} \tag{16}$$

where CI , RI , and CR represent the consistency index, random index, and consistency ratio, respectively, while X is the largest eigenvalue of the n -order matrix and W is the weight of each parameter. It has a value between 0 and 1, and the sum of that weight must be 1. The random indexes of different number of criteria have different values as shown in Table 4. If the consistency ratio is less than 0.1, the result is acceptable and matrix A is fully compatible; otherwise, it returns to step 1 and the performed operations are repeated.

Table 4 Relationship between random index values and number of criteria

n	2	3	4	5	6	7	8	9	10
RI	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.51

5 Performance evaluation

In this section, we first explain our simulation environment and its related settings and then we show the experimental results along with their analysis.

5.1 Experimental settings

Our proposed approach (i.e., APT-Dt-KC) was simulated along with the proposed approach in [17] (i.e., DT-EnSVM) using MATLAB 2015. All the experiments were performed on a computer with an Intel Core i7-6700 @ 3.40 GHz processor with 16 GB RAM and Windows 7.

In all the experiments, KDD Cup 99 dataset [17] generated by MIT laboratory has been used to evaluate our proposed approach. This dataset, which is the most common and standard dataset for evaluating the intrusion detection systems since 1999, is based on data recorded from DARPA 98 project. This dataset contains approximately 4 GB of raw binary data collected by TCPDump software from the traffic of network. It also contains approximately 5 million records with connection vectors, each with a size of 100 bytes and 41 attributes, and a tag that includes normal or attack modes. Due to the large volume of this dataset, most studies have used 10% reference subset provided in the standard dataset. Therefore, we also used 10% of dataset in our experiments. This dataset includes 494,021 records with 23 attacks and 2 normal classes. In this dataset, data records are classified into four main classes of attacks, including the Denial of Service (DoS) attacks, obtaining initial access (R2L), improving the level of access and performance (U2R), and probing. In our experiments, these attacks were used as an example for the APT attacks according to the fuzzy features of cyber-kill chain and the adaptation of these kinds of attacks with some phases of the cyber kill chain model. In this classification, we have considered the attacks of DoS, R2L, U2R, and probing for information collection (“Reconnaissance”), intrusion (“weaponization” and “Delivery”), deployment (“Exploitation” and “Installation”) and information stealing (“Command and Control” and “Action on Objectives”) phases. To distinguish normal communication from attacks, each network communication data component in KDD99 was represented by 41 attributes. In addition, each data component in KDD99 was marked as the attack alert or the normalized alert. It should be noted that there were 23 types of cyber-attacks that could be used in an APT attack, each of which belongs to one of the four given attack classes.

The general structure of APT-Dt-KC approach consists of five functional components and three memory components. The functional components in this structure include data processing, classification, detection, training, and decision making based on the fuzzy based classification structure. Since the data forms used in the process of detecting the APT attacks have different modes and the training data set changes dynamically in both approaches (i.e., Bayesian classification method and fuzzy AHP method), three types of memory were considered in APT-Dt-KC approach. A memory was used to sort all the candidate data for training; a cache

memory was employed to link the training data of each procedure in Bayesian algorithm; finally, a cache memory was applied to link the training data during the process of detecting the APT attacks. The types of sorted data in these three memories were displayed as small point, large point, and object, respectively. Other structures used in the simulation were as follows:

- **Row Processing Structure:** In APT-Dt-KC approach, row data were converted to point-type data after processing and stored on the normal memory. Then, they were used as the input training data for Bayesian algorithm.
- **Structure of Training Algorithm:** In this structure, data were converted to large point data and stored on the cache of device memory. This structure implements Bayesian algorithm training process until the selected data points form the final training structure.
- **Hierarchical Classification:** Data in this structure were converted to data types that could be evaluated by the fuzzy structure and stored on the memory of decision-making and classification model. This structure aligns the decision-making process with the marked objects.
- **Creation of a Detection-Based Classification Structure:** It implemented APT-Dt-KC approach based on classifier using the Bayesian and AHP method.
- **Testing the Detection-Based Classification Structure:** This structure tested the classifier stored on the memory using a test dataset and the statistical results obtained from the analysis of this data were represented in the form of graphs.

It is obvious that Bayesian algorithm achieves the best detection rate when the amount of training data from two classes is balanced. In our experiments, the amount of normal data and different ratios of intrusion types are set to deal with the problem of imbalance. Let D represents a training dataset. The distribution for each class in D is shown in Table 5.

To compare the performance of APT-Dt-KC approach, the classifiers generated by Bayesian algorithm must be tested separately. Since general comparisons and their results are not related to the amount of training data and the distribution of each data class, only 10% of the training data was used. Assume that $T1$ represents the amount of experimental data used to compare the available approaches. The distribution of $T1$ -related classes is also given in Table 6. To

Table 5 Class distribution in network communication records

Class	10% dataset of KDD99	D training dataset	T1 test set	T2 test set
Normal	97,277	200	1000	10,000
DoS	391,458	60	500	40,000
U2R	52	30	52	52
R2L	1126	60	1000	100
Probe	4107	40	500	400
Total	494,021	390	3052	50,552

Table 6 Confusion matrix for *T1* dataset

Classes	Normal	DoS	U2R	R2L	Probe
<i>a. Support vector machine classifier with feature improvement in DT-EnSVM approach</i>					
Normal	345	17	3	48	15
DoS	0	485	0	0	7
U2R	7	0	38	15	0
R2L	39	289	25	653	5
Probe	5	256	26	5	84
<i>b. A classifier based on Bayesian algorithm in APT-Dt-KC approach</i>					
Normal	320	11	1	31	11
DoS	0	162	0	0	4
U2R	6	0	36	14	0
R2L	26	156	22	254	3
Probe	4	250	12	4	71

consider this trend, another dataset, called *T2*, was considered. The distribution of all classes in *T2* dataset was almost similar to the distribution of the entire 10% KDD99 dataset. To eliminate the differences among the features and prevent the overcoming of features with a large value compared to the features that have a lower numerical range, when data were normalized, the preprocessing was performed on data using Pearson correlation. In our experiments, 90% of the given data was selected to test the intrusion detection model and another 10% was used to train the model. Therefore, each subset had an equal chance in the training and testing by running the model 10 times. In the phase of evaluation, the performance of APT-Dt-KC approach and DT-EnSVM approach was evaluated to enhance the features using KDD99 as the dataset and *T1* as the experimental data. In this phase, the feature improvement of training scheme could be used directly for multi-fold problems to evaluate DT-EnSVM approach. In the training process, the intrusion detection system trained a single classifier that provides a model for each class of attacks. A suitable value was required for the similarity coefficient β , which could be estimated between 0.2 and 0.4. In our experiments, the value of β was set to 0.25. The coefficient β indicates the similarity of congestion in the dataset. By exploiting this structured based classifier, all data in *T1* could be classified.

Finally, the performance of APT-Dt-KC approach was evaluated and compared with DT-EnSVM approach in terms of training time, detection rate, precision rate, accuracy, false positive rate, false negative rate, false alarm rate (i.e., FAR) and detection rate (i.e., DR). A false positive is a warning event that sounds when an attack has not occurred, while a false negative indicates that we have encountered a problem in detection of the real attack. All the experiments were performed 5 times, and finally, the experimental average values were used as the result of experiments.

5.2 Experimental results

In DT-EnSVM approach, the support vector-based classification process was applied to the training scheme and 4 classifiers were trained to differentiate each class from the existing training attacks. Data records in *TI* datasets have been tested using a support vector machine classifier. It was formed based on a voting strategy in such a way that each decision on each binary classification was considered as a vote. Finally, the final decision was generated for the classes using the maximum number of votes. In the case that the maximum votes did not form a single class, the data points were labeled as unknown. In contrast, in APT-Dt-KC approach, before performing the process of training and testing, the data were pre-processed and optimized for future processing by Pearson correlation test. Then, the data were included in the training cycle as a set of processing features. Since input data included both normalized and optimal data, redundancy in the preprocessing process was significantly reduced. In our experiments, 10% of data was trained and the remaining 90% was evaluated. Based on the evaluation results, the confusion matrix could be extracted for APT-Dt-KC approach and DT-EnSVM approach by improving the features. Generally, a confusion matrix is a matrix in which the performance of relevant algorithms is represented. Such a representation is commonly used for supervised learning algorithms, although it also applies to unsupervised learning. This matrix is used to determine the value of evaluation indicators such as precision and accuracy, which will be discussed later. The confusion matrix extracted from the support vector machine classifier test with the feature improvement and our proposed approach for *TI* dataset are shown in Tables 6a and b, respectively. Based on the confusion matrices, APT-Dt-KC approach is more efficient compared to DT-EnSVM approach.

As explained in Sect. 4, a combination of Bayesian algorithm and fuzzy AHP is applied in APT-Dt-KC approach. In the training strategy of APT-Dt-KC approach, the parameter *M* is responsible for controlling the convergence of these two algorithms. The detection rate and training time for *TI* dataset are shown in Fig. 4 when this parameter was set in ranges of 1 to 6 with intervals of 1. Considering the efficiency and accuracy of this parameter, the value of *M* was set to 4.

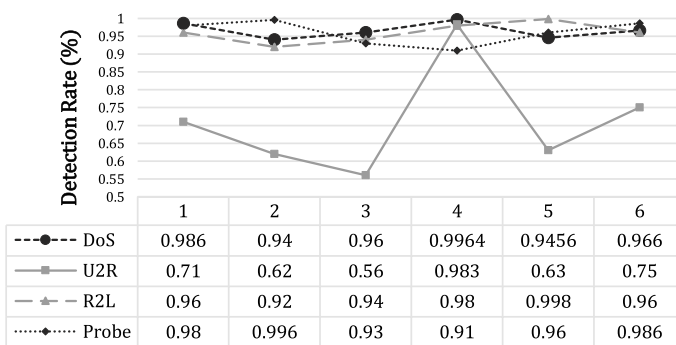


Fig. 4 Detection rate and training time with different values of *M*

The value of M represents the data points that will be used for training in the next step of Bayesian algorithm in APT-Dt-KC approach and DT-EnSVM approach.

The confusion matrix extracted from DT-EnSVM approach and APT-Dt-KC approach for $T2$ dataset are shown in Tables 7a and b, respectively. The purpose of evaluating this dataset is to analyze the scalability of APT-Dt-KC approach in a larger amount of data. Based on Table 7, the confusion matrix for APT-Dt-KC approach has a better efficiency in the classification phase than DT-EnSVM approach.

The comparative results for both datasets $T1$ and $T2$ are shown in Tables 8 and 9, respectively. Based on the results in Tables 8 and 9, APT-Dt-KC approach has higher efficiency than DT-EnSVM approach in terms of detection rate, false positive rate, and false negative rate. Also, the training time in APT-Dt-KC approach

Table 7 Confusion Matrix for $T2$ Dataset

Classes	Normal	DoS	U2R	R2L	Probe
<i>a. Support vector machine classifier with feature improvement in DT-EnSVM approach</i>					
Normal	7854	74	10	32	41
DoS	115	3545	55	88	541
U2R	5	2	14	152	7
R2L	75	19	25	963	85
Probe	51	18	44	266	362
<i>b. A classifier based on Bayesian algorithm in APT-Dt-KC approach</i>					
Normal	6985	74	10	32	30
DoS	3	241	52	1	62
U2R	0	2	21	85	0
R2L	4	1	4	84	1
Probe	0	12	6	121	158

Table 8 Comparison of evaluation results in $T1$ dataset

Criteria/Approach	DT-EnSVM	APT-Dt-KC
Training time (s)	3.338	2.982
Detection rate (%)	78.180	89.52
False positive rate(%)	2.776	1.985
False negative rate (%)	0.322	0.296

Table 9 Comparison of Evaluation Results in $T2$ Dataset

Criteria/Approach	DT-EnSVM	APT-Dt-KC
Training time (s)	5.452	4.511
Detection rate (%)	94.86	98.65
False positive rate (%)	6.01	4.652
False negative rate (%)	1.23	0.896

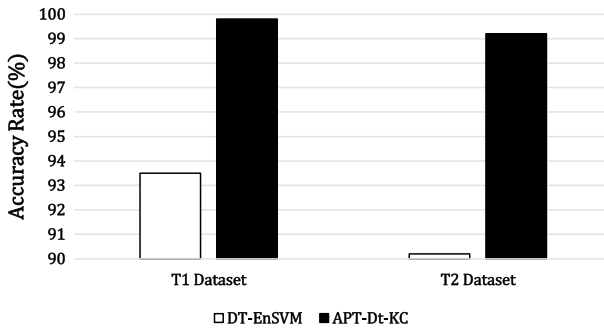


Fig. 5 Accuracy rate of attack detection in T1 and T2 datasets

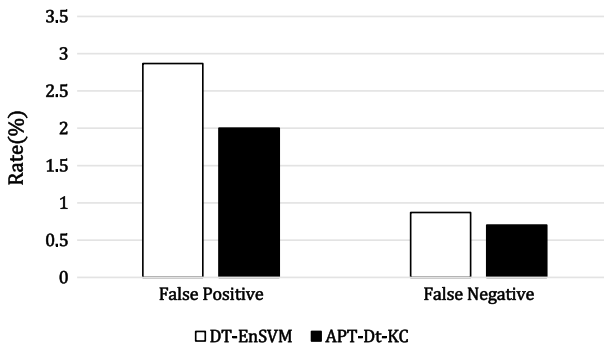


Fig. 6 Rates of false positive and false negative in the phase of attack detection in T1 dataset

is better than DT-EnSVM approach due to the elimination of redundancy of the evaluated alerts.

The rates of detection accuracy for APT-Dt-KC approach and DT-EnSVM approach in *T1* and *T2* datasets are shown in Fig. 5. Based on the experimental results, the accuracy rate of APT-Dt-KC approach is higher than DT-EnSVM approach in both *T1* and *T2* datasets.

After the attack classification process, the level and priority of attacks should be identified. Since that APT-Dt-KC approach applies AHP algorithm to consider the level of attacks, it can rank different attacks. Therefore, all the attacks can be classified, prioritized, and evaluated as most likely to occur. The experimental results in the phase of attack detection in *T1* dataset are shown in Fig. 6. As shown in Fig. 6, the rate of false positive and the rate of false negative for APT-Dt-KC approach are less than DT-EnSVM approach.

The values of mean detection rate for APT-Dt-KC approach and DT-EnSVM approach in *T1* dataset are shown in Fig. 7. The experimental results show that the final detection rate for APT-Dt-KC approach is better than DT-EnSVM approach. Considering that the detection rate in the classification process also had a better

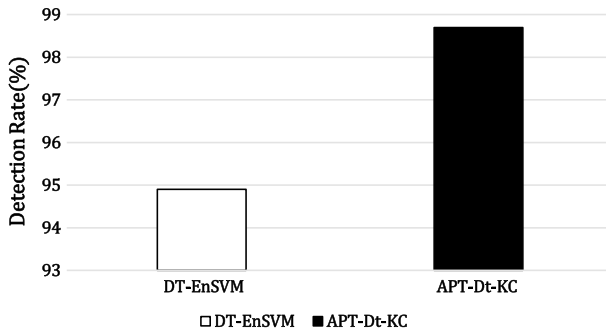


Fig. 7 Rate of mean detection in T1 Dataset

level for APT-Dt-KC approach, it can be concluded that APT-Dt-KC approach provides better detection in whole process.

The use of data preprocessing process and Bayesian algorithm, which has a better convergence rate and accuracy than the support vector machine, increases the efficiency of APT-Dt-KC approach and improves the detection process over DT-EnSVM approach.

To evaluate the scalability of APT-Dt-KC approach, T2 dataset was used in our experiments. The values of false positive rate and false negative rate in T2 dataset for APT-Dt-KC approach and DT-EnSVM approach are shown in Fig. 8. As it can be seen from Fig. 8, these values for APT-Dt-KC approach are smaller than DT-EnSVM approach. It means that APT-Dt-KC approach has higher accuracy than DT-EnSVM approach in large datasets.

The values of mean detection rate in APT-Dt-KC approach and DT-EnSVM approach in T2 dataset are shown in Fig. 9. As shown in Fig. 9, the experimental results show the efficiency of APT-Dt-KC approach compared to DT-EnSVM approach on the existing scale.

Table 10 shows the values of accuracy, false alarm rate (FAR) and detection rate (DR) in APT-Dt-KC approach and DT-EnSVM approach. Based on the

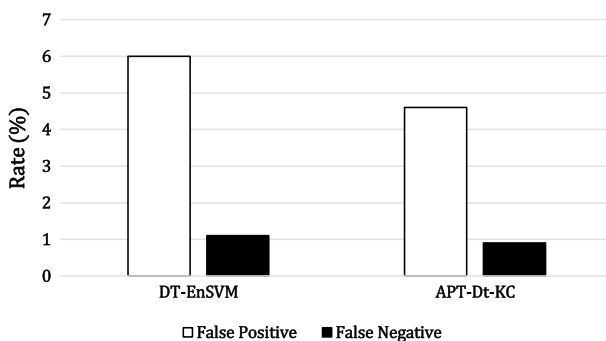


Fig. 8 Rates of false positive and false negative in T2 dataset

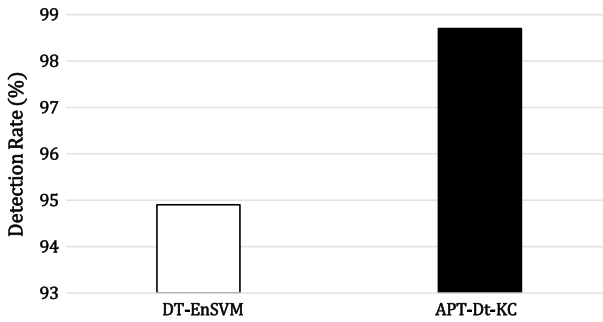


Fig. 9 Mean detection rate in T2 dataset

Table 10 Comparison of Experimental Results for Other Metrics

Criteria/Approach	DTEnSVM	APT-Dt-KC
Accuracy	98.3	99.1
False alarm rate (FAR)	99.09	99.65
Detection rate (DR)	0.31	0.28

experimental results, it is determined that APT-Dt-KC approach has better precision rate, FAR and DR compared to DT-EnSVM approach. The reason of these improvements is due to improvement of false positive rate and false negative rate in APT-Dt-KC approach.

6 Conclusion and future works

In this paper, a new approach called APT-Dt-KC was proposed for detecting the APT attacks. APT-Dt-KC approach used the prioritization and selection based on Bayesian algorithm classification to select training data and significantly increase the efficiency in terms of runtime. It also used an AHP to solve multi-criteria optimization problem using a prioritization-based classification method. Therefore, it could improve the mean detection rate for all data classes. APT-Dt-KC approach had lower training overhead compared to DT-EnSVM approach since it only used data points around the boundaries of two classes to construct the classifier. The training process for intrusion detection using DT-EnSVM approach was suitable for multi-criteria cases with various problems, but it had more difficulty in balancing the training and test data. However, APT-Dt-KC approach overcomes the existing problems. Finally, in APT-Dt-KC approach, the training and testing processes could be done in parallel since the training and testing phases are designed as separate modules.

Given that our proposed approach uses AHP technique to solve multi-criteria optimization problem, it has high computational overhead. Thus, it can be claimed that the problem with our approach on large datasets is computational overhead. In the future, this problem can be considered by using other optimization techniques. Moreover, the evolutionary algorithms such as bee colony can be combined by

APT-Dt-KC approach to improve the efficiency and flexibility of an intrusion detection system. In addition, other classifiers such as neural network-based methods can be applied in APT-Dt-KC approach as a new path to detect the APT attacks.

References

1. Alazzam H, Sharieh A, Sabri KE (2020) A feature selection algorithm for intrusion detection system based on Pigeon inspired optimizer. *Expert Syst Appl*. <https://doi.org/10.1016/j.eswa.2020.113>
2. Quincozes SE, Albuquerque C, Passos D, Mossé D (2021) A survey on intrusion detection and prevention systems in digital substations. *Comput Netw*. <https://doi.org/10.1016/j.comnet.2020.107679>
3. Bostani H, Sheikhan M (2017) Modification of supervised OPF-based intrusion detection systems using unsupervised learning and social network concept. *Pattern Recogn* 62:56–72
4. Hassan MM, Gumaei A, Alsanad A, Alrubaian M, Fortino G (2020) A hybrid deep learning model for efficient intrusion detection in big data environment. *Inf Sci* 513:386–396
5. Condomines JP, Zhang R, Larriue N (2019) Network intrusion detection system for UAV ad-hoc communication: From methodology design to real test validation. *Ad Hoc Netw*. <https://doi.org/10.1016/j.adhoc.2018.09.004>
6. Martinez CV, Vogel-Heuser B (2021) A host intrusion detection system architecture for embedded industrial devices. *J Franklin Inst* 358:210–236
7. Setiawan B, Djanali S, Ahmad T, Aziz MN (2019) Assessing centroid-based classification models for intrusion detection system using composite indicators. *Procedia Comput Sci* 161:665–676
8. Rahouma K, Ali A (2019) Applying intrusion detection and response systems for securing the client data signals in the Egyptian optical network. *Procedia Comput Sci* 163:538–549
9. Dong Y, Wang R, He J (2019) "Real-Time Network Intrusion Detection System Based on Deep Learning," in: 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, pp 1–4
10. Zhou Y, Mazzuchi TA, Sarkani S (2020) M-AdaBoost-A based ensemble system for network intrusion detection. *Expert Syst Appl* 162:2020. <https://doi.org/10.1016/j.eswa.2020.113864>
11. Zimba A, Chen H, Wang Z (2019) Bayesian network based weighted APT attack paths modeling in cloud computing. *Futur Gener Comput Syst* 96:525–537
12. Lee M, Choi J, Choi C, Kim P (2017) APT attack behavior pattern mining using the FP-growth algorithm," in: 2017 IEEE Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, USA, pp 1–4
13. Hasan K, Shetty S, Ullah S (2019) Artificial Intelligence Empowered Cyber Threat Detection and Protection for Power Utilities, in: 2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC), Los Angeles, USA, pp 354–359
14. Wang Q, Cai X, Tang Y, Ni M (2021) Methods of cyber-attack identification for power systems based on bilateral cyber-physical information. *Int J Elect Power Energy Syst*. <https://doi.org/10.1016/j.ijepes.2020.106515>
15. Bhatnagar D, Som S, Khatri SK (2019) Advance Persistent Threat and Cyber Spying - The Big Picture, Its Tools, Attack Vectors and Countermeasures, in: 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, pp 828–839
16. Eggers S (2021) A novel approach for analyzing the nuclear supply chain cyber-attack surface. *Nucl Eng Technol* 53:879–887
17. Gu J, Wang L, Wang H, Wang S (2019) A novel approach to intrusion detection using SVM ensemble with feature augmentation. *Comput Secur* 86:53–62
18. Zulkefli Z, Singh MM, Shariff ARM, Samsudin A (2017) Typosquat cyber crime attack detection via smartphone. *Procedia Comput Sci* 124:664–671
19. Cho DX, Nam HH (2019) IA method of monitoring and detecting APT attacks based on unknown domains. *Procedia Comput Sci* 150:316–323
20. Dargahi T, Dehghantaha A, Nikkha Bahrami P, Conti M, Bianchi G, Benedetto L (2019) A Cyber-Kill-Chain based taxonomy of crypto-ransomware features. *J Comput Virol Hack Tech* 15:277–305
21. Tankard C (2011) Advanced Persistent threats and how to monitor and deter them. *Netw Secur* 2011:16–19

22. Khan MS, Siddiqui S, Ferens K (2018) "A Cognitive and Concurrent Cyber Kill Chain Model", 2018, in: *Computer and Network Security Essentials*, Springer, pp 585-602
23. Bryant B, Saiedian H (2017) A novel kill-chain framework for remote security log analysis with SIEM software. *Comput Secur* 67:198–210
24. Yadav T, Rao AM (2015) "Technical Aspects of Cyber Kill Chain", In: *Security in Computing and Communications*. (SSCC 2015), Communications in Computer and Information, Vol 536. Springer, https://doi.org/10.1007/978-3-319-22915-7_40
25. Mohsin M, Anwar Z (2016) "Where to Kill the Cyber Kill-Chain: An Ontology-Driven Framework for IoT Security Analytics," in: 2016 International Conference on Frontiers of Information Technology (FIT), Islamabad, 2016, Islamabad, Pakistan, pp 23-28
26. Kiwiaa D, Dehghantanhaa A, Choob K-KR, Slaughter J (2018) A cyber kill chain based taxonomy of banking Trojans for evolutionary computational intelligence. *J Comput Sci* 27:394–409
27. Verma R, Chandra S (2020) "A Fuzzy AHP Approach for Ranking Security Attributes in Fog-IoT Environment," in: *A Fuzzy AHP Approach for Ranking Security Attributes in Fog-IoT Environment (ICCCNT)*, Kharagpur, India, pp 1–5
28. Ogundoyin SO, Kamil IA (2020) A Fuzzy-AHP based prioritization of trust criteria in fog computing services. *Appl Soft Comput* 97:106789
29. Liu Y, Eckert CM, Earl C (2020) A review of fuzzy AHP methods for decision-making with subjective judgements. *Expert Syst Appl* 161:113738
30. Kalaiselvi B, Thangamani M (2020) An efficient Pearson correlation based improved random forest classification for protein structure prediction techniques. *Measurement* 162:107885
31. Singh VK, Callupe SP, Govindarasu M (2019) Testbed-based Evaluation of SIEM Tool for Cyber Kill Chain Model in Power Grid SCADA System", in: 2019 North American Power Symposium (NAPS), Wichita, KS, USA
32. Kim J, Baskerville RL, Ding Y (2020) Breaking the privacy kill chain: protecting individual and group privacy online. *Inf Syst Front* 22:171–185
33. Shamel-Sendi A, Dagenais M (2014) ARITO: cyber-attack response system using accurate risk impact tolerance. *Int J Inf Secur* 13:367–390
34. Duncan A, Creese S, Goldsmith M (2019) "A Combined Attack-Tree and Kill-Chain Approach to Designing Attack-Detection Strategies for Malicious Insiders in Cloud Computing", in: 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), Oxford, United Kingdom, United Kingdom, pp 1–9
35. Schneier B (2019) *Attack Trees*, Dr. Dobb's Journal, 24
36. Hoffmann R (2019) "Markov Models of Cyber Kill Chains with Iterations", in: *International Conference on Military Communications and Information Systems (ICMCIS)*, Budva, Montenegro
37. Hoffmann R (2018) The general cyber-attack life cycle and its continuous time Markov chain model. *Ekonomiczne Problemy Usług* 10:121–130
38. Hutchins EM, Cloppert MJ, Amin RM (2011) "Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains", *Leading Issues in Information Warfare and Security Research*, pp 78–104
39. Martin L (2015) "Seven Ways to Apply the Cyber Kill Chain with a Threat Intelligence Platform," [Online]. Available: <https://www.lockheedmartin.com/content/dam/>
40. Yang L, Li P, Yang X, Xiang Y, Jiang F, Zhou W (2019) "Effective Quarantine and Recovery Scheme Against Advanced Persistent Threat", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp1–5
41. Joloudari JH, Haderbadi M, Mashmool A, Ghasemigol M, Band SS, Mosavi A (2020) Early detection of the advanced persistent threat attack using performance analysis of deep learning. *IEEE Access* 8:186125–186137
42. Yan D, Liu F, Jia K (2019) "Modeling an Information-Based Advanced Persistent Threat Attack on the Internal Network," in: *IEEE International Conference on Communications (ICC)*, Shanghai, China, pp 1–7
43. Yang L-X, Huang K, Yang X, Zhang Y, Xiang Y, Tang YY (2020) "Defense against advanced persistent threat through data backup and recovery," *IEEE Transactions on Network Science and Engineering*, pp 1–1
44. Ahmed Y, Asyhari AT, Rahman MA (2021) A cyber kill chain approach for detecting advanced persistent threat. *Comput Mater Continua* 67:2497–2513
45. Xuan CD (2021) Detecting APT Attacks Based On Network Traffic Using Machine Learning. *Journal of Web Engineering*. <https://doi.org/10.13052/jwe1540-9589.2019>

46. Immink KAS, Weber JH (2014) Minimum Pearson distance detection for multilevel channels with gain and/or offset mismatch. *IEEE Trans Inf Theory* 60:5966–5974
47. Rummel RJ (1976) *Understanding Correlation, Life Time of Correlations and its Applications*. Wydawnictwo Niezależne, pp 5–21
48. Mu Y, Liu X, Wang L (2018) A Pearson's correlation coefficient based decision tree and its parallel implementation. *Inf Sci* 435:40–58
49. Kyrimi E, Neves MR, McLachlan S, Neil M, Marsh W, Fenton N (2020) Medical idioms for clinical Bayesian network development. *J Biomed Inform* 108:103495
50. Mohammadi H (2012) "Strategic Decision Making in Resource Selection", 2, 1-12

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.