Cybercrimes Investigation and Intrusion Detection in Internet of Things Based on Data Science Methods

Ezz El-Din Hemdan and D. H. Manjaiah

Abstract In recent times, Internet of Things (IoT) has paying attention from different organization ranging from academia to industry. The IoT is an internetworking of connecting and integrating several types of devices and technologies that comprising sensors, Radio Frequency Identification (RFID), cloud computing, the Internet, smart grids, and vehicle networks, and many other devices and new technologies. The IoT becomes a subject for illegal and criminals activities. Cyber criminals and terrorists are highly qualified persons in the computer, network, digital systems and new technologies. An enormous amount of data is gathered about criminals and their behavior from different data sources over the Internet can be processed using data science methods to monitor and trace them in real-time and online. The massive amount of data needs new fast and efficient processing tools and techniques for data extracting and analyzing in less period of time. Data science methods can be used for this purpose to investigate and detect a different type of severe attacks and intrusions. This chapter introduce principles of Digital Forensics, Intrusion Detection and Internet of Things as well as exploring data science concepts and methods that can help the digital investigators and security professionals to develop and propose new data science techniques and methods that can be adapted to the unique context of Internet of Things environment for performing intrusion detection and digital investigation process in forensically sound and timely fashion manner.

Keywords Cybercrimes • Digital forensics • Intrusion detection
Internet of things • Data science methods

E. E.-D. Hemdan (✉) · D. H. Manjaiah
Department of Computer Science, Mangalore University, Mangalore, India
e-mail: ezzvip@yahoo.com

D. H. Manjaiah
e-mail: manju@mangaloreuniversity.ac.in

A. K. Sangaiah et al. (eds.), *Cognitive Computing for Big Data Systems Over IoT*,
Lecture Notes on Data Engineering and Communications Technologies 14,
https://doi.org/10.1007/978-3-319-70688-7_2

1 Introduction

Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining and knowledge discovery, and predictive analytics. There are several advantages and benefits of applying data science methods in cybercrime investigation and intrusion detection over Internet of Things (IoT). This is to look for crucial information that can be used in the digital investigation and help refute or support a claim or put together a missing piece, this has seen a rapid increase in the field of digital investigation, and intrusion detection and prevention. Internet of Things has paying attention from different organization ranging from academia to industry. A Huge amount of data is congregated about criminals and their behavior from different data sources in the IoT environment through using data science methods to observing and tracing them. New fast and efficient processing tools and techniques are required for data extracting and analyzing in less period of time. Data science methods can be used for this purpose to investigate and detect a different type of severe attacks and intrusions. Digital Investigators, examiners and system administrators can use numerous innovative statistics, machine learning, data mining, and predictive analytics to recognize data patterns from the gigantic collected large data from IoT devices to discover any digital evidence about hackers or detect and trace them through their criminal activities by identifying suspicious behavior patterns to identify threats that are likely to happen.

Presently, existing anomaly detection is often associated with high false alarm with moderate accuracy and detection rates when it's unable to detect all types of attacks properly as well as digital investigation of cybercrimes is become two important area in information security. Thus, this chapter focus on applying both of them over the Internet of Things based on using data science methods and approaches that can help to improve detecting and investigating crimes in efficient and effective manner. This chapter will explore and identify challenges and opportunities of digital forensic and intrusion detection and how can apply data science approaches and cognitive methods to fight and investigate serve attacks and crimes over the Internet of Things environment in forensically sound and timely way.

This chapter introduce principles of Digital Forensics, Intrusion Detection and Internet of Things as well as exploring data science concepts and methods that can help the digital investigators and security professionals in developing and proposing new techniques that can be adapted to the unique context of Internet of Things environment which can help in performing intrusion detection and digital investigation process in forensically sound and timely fashion manner.

The remainder of this chapter is structured as follows: Sect. 2, presents an overview about digital forensics, intrusion detection, and the internet of things as well as exploring data science concepts and methods while cybercrimes investigation in the internet of things is presented in Sect. 3. Section 4 provides intrusion detection in the internet of things while applying data science methods for the cybercrimes investigation and intrusion detection in the internet of things is introduced in Sect. 5. Finally, the chapter conclusions and future directions in this innovative subject are presented in Sect. 6.

2 Background

This section provides basics of digital forensics, intrusion detection, and Internet of Things (IoT) along with identifying data science concepts and methods.

2.1 Digital Forensics

Digital forensics is a branch of forensics science that concern with finding and collecting digital evidence then analysis and examine them to find any traces related to crimes against digital systems. Digital forensics has many directions such as Computer Forensics, Mobile Forensics, Network Forensics and Cloud Forensics. This section discusses digital forensics definition as well as digital forensics investigation process which the digital investigators follow it during the investigation of crimes to reconstruct the crime events that occurred.

2.1.1 Digital Forensics Definition

The process of collecting, identifying, preserving and examining digital evidence is known as 'Digital Forensics'. One of the popular definition for the digital forensics is introduced by first Digital Forensic Research Workshop (DFRWS). The DFRWS defined the digital forensics as: "*The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations*" [1].

2.1.2 Digital Forensic Investigation Process

Criminals and attackers after committing their cybercrimes some trails that remain behind them. Collecting, extracting and preserving digital evidence from the crime scene need careful strategies to handle and manage them to become ready for presenting in the court of law. In digital forensics, there are four crucial steps for performing the digital forensic process as shown in Fig. 1 as follows [2]:

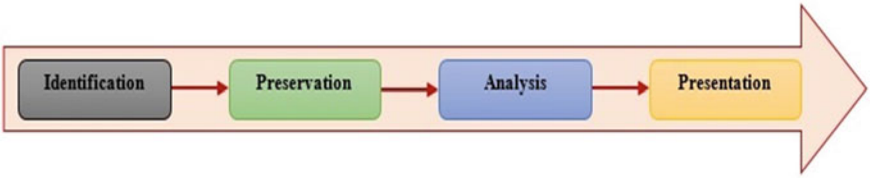

Fig. 1 Digital forensic investigation process

- *Identification*: It is the identification of sources of digital evidence, which will be required to prove the committed crime.
- *Preservation*: In the preservation process, the digital investigator preserves the collected digital evidence such as Hard Disks, Laptops, Tablets and Mobile Phones.
- *Analysis*: In the analysis process, the examiners and digital investigators interpret and correlate the evidential data to come to a summary and conclusion, which can prove or disprove civil, or criminal actions.
- *Presentation*: In this process, the digital investigators make a forensic report to summarise their findings from the analysis process. This report should be suitable to present to the court of law.

2.2 Intrusion Detection System

Intrusion Detection System (IDS) is can be a software or hardware system that observers the system or actions of the network for policy criminal and malicious activities and produces reports to the central administration system. The main focus of intrusion detection systems is to recognize the possible incidents, logging information about them and in report attempts. Furthermore, organizations use the intrusion detection systems for other objectives, such as detecting problems with security policies, deterring individuals and documenting present threats from infringing security policies. The intrusion detection systems have become an important addition to the security infrastructure of nearly each organization. Several procedures can be used to identify criminals and intrusions but each one is specific to a particular method. The main aim of intrusion detection system is to detect and identify the attacks professionally. In addition, it is equally imperative to detect attacks at an early stage in order to reduce their impacts.

2.2.1 Categorization of Intrusion Detection Systems

Intrusion Detection System is used to analyze packet traffic to match it with any anomalies found in comparison with the normal traffic. For anomalous traffic, the

IDS tries to identify the pattern of common threats and alerts the system administrator. Ever since an IoT-based platform is frequently a high-speed network, it essential be protected using an entirely automated intrusion detection system. The intrusion detection system briefly classified into two basic categories:

- *Network Intrusion Detection System*: This type tries to secure all machine systems in the network.
- *Host-based Intrusion Detection System*: This type tries to secure a single host. A highly scalable intrusion detection system is able to provide support for proficient utilization of recent high-performance architectures.

Detection models are divided into statistical or signature based models [3]. The statistical model maintains profiles regarding applications, hosts, users, and connections. It then matches current activity with the attributes of the profile for any anomalies. In the other side, the signature-based model compares the traffic against a collection of existing signatures. Also, there are three detection approaches that are used by host-based or network intrusion detection systems which are used to analyze events and discover attacks as follows [4, 5]:

- *Signature based System*: This model is also known as misuse detection system which still the utmost method and focuses on the identification of known bad patterns and searching for similar activities such as vulnerabilities or acknowledged intrusion signatures. As with each system that uses a blacklist approach, it is vulnerable to attacks for which the signature is unknown, such as zero-day exploits or use of encoding, obfuscation or packing methods.
- *Anomaly based System*: This system work on searching unusual behavior on network traffic as well as observing system behavior to fix whether an observed activity is anomalous or normal, according to a heuristic analysis, can be used to identify unknown attacks. Anomaly detection based IDS model has the ability to detect attack indications without specifying attack models, but these models are very sensitive to false alarms.
- *Specification based IDS*: This type of IDS is like anomaly detection system. In this system, the normal behavior of the network is defined by manually, so it gives less incorrect positives rate. This system attempts to excerpt best between signature-based and anomaly based detection methods by trying to clarify deviations from normal behavioral patterns that are produced neither by the training data nor by the machine learning techniques. The development of attack specification is done manually so it takes more time.

2.3 Internet of Things

Recently, Internet of Things (IoT) has an urgent economic and societal impact for the future construction of communication and network systems to exchange information between things and people. The novel planning of future will be eventually, “everything will be connected and intelligently controlled”. The idea of IoT is becoming more relevant to the real world due to the development of mobile devices, cloud computing, embedded and ubiquitous communication technologies, data science and data analytics. The IoT made up of devices connected to the Internet to collect information about the environment using sensors connected to devices (i.e. things). These devices communicate and interact together to acquire, process and storage information in smart and intelligent manner.

With the IoT, millions of devices are connected to each other which need to exchange information through the network (i.e. Internet) with the need to massive capabilities such as processing, storage, and high bandwidth. These capabilities can be delivered through using cloud computing technology. Researchers and scientist who are working in the IoT can use the cloud computing services to design and develop applications that can create of smart environments like Smart Cities. The devices that are used in the IoT system produces enormous data (i.e. big data) which often need to leverage the technology of cloud computing to scale cost effectively. Big data analytic is an important direction nowadays to help business to predict about future and so make correct decisions in business marketing.

2.3.1 Internet of Things Operations

In Internet of Things, there are various operation phases include collection phase, transmission phase, and processing, management and utilization phase [6] as follows:

- *Collection Phase*: The principal aim is to collect data about the physical environment. Sensing devices and technologies for short range communication are combined to reach this objective. Devices of the collection phase are usually small and resource-constrained. Communication technologies and protocols for this phase are designed to operate at limited data rates and short distances, with constrained low energy consumption and memory capacity. Due to these characteristics, collection phase networks often are referred to as Low power and Lossy Networks (LLN).
- *Transmission Phase*: The goal of this phase is to transmit the data collected during the collection phase to applications and, therefore, to users. Here, technologies such as Ethernet, WiFi, Hybrid Fiber Coaxial (HFC) and Digital Subscriber Line (DSL) are united with TCP/IP protocols to construct a network that interconnects objects and users across longer distances. Gateways are necessary to integrate LLN protocols of the collection phase with traditional Internet protocols employed in the transmission phase.
- *Processing, Management and Utilization Phase*: Applications process gather data to obtain useful data about the physical environment. These applications

may take decisions based on this data, controlling the physical objects to act on the physical environment. This phase also contains a middleware, which is responsible for facilitating the integration and communication between different physical objects and multi-platform applications.

2.3.2 Internet of Things Categorization

Internet of Things can be categorized into four categories such as Internet of Nano Things, Internet of WiFi-enabled Things, Internet of Things for Smart Society, and Global-scaled Internet of Things [7–13] as follows:

1. *Internet of Nano Things (IoNT)*: The IoNT consist of nano-devices that are communicating with each other over a nanonetwork. In the IoNT, it becomes possible to add a new dimension to the IoT by embedding nano-sensors to the numerous things and devices that surround us. Also, it can be used in different areas such as biology.
2. *Internet of WiFi-enabled Things*: Currently, the WiFi is a significant category of wireless networks for connecting several devices to the Internet. When the WiFi enabled devices are connected together over the Internet which it offered new kind of the IoT named Internet of WiFi-enabled Things.
3. *Internet of Things for Smart Society*: The idea of smart city become an attractive research topic for numerous scientists and researchers to introduce novel methods for connecting things or devices in the society in a smart manner to make a smart society through embedding sensors in all surrounding devices and things to allow them to interact and communicate together in an intelligent manner. This gets a new category of IoT known as the internet of things for smart society.
4. *Global-scaled Internet of Things*: It is utilizing in a global-scaled area such as unmanned aerial vehicle and satellite system. Using remote connections, these systems communicated and interacted with several devices which are connected to sensors to sense data in an effective way. The Tsunami Detection System is a real world example about the global-scaled internet of things [13].

2.3.3 Internet of Things Applications

Internet of Things is an imperative paradigm for providing smart applications which can improve and enhance the quality of our lives to the better level of life. Recently, the IoT has several applications in various areas such as; industrial control system smart society, smart manufacturing, smart agriculture, healthcare, military, and trade and logistics [7–13] as follows:

1. *Smart City:* The smart city is the idea of making smart cities which make people life more comfortable and easy. The innovative development of smart technologies assists the IoT in changing people life style. The IoT can be used in smart cities to provide many services as; intelligent highways with warning messages for unexpected actions such as accidents. Also, for monitoring of vehicles and pedestrian levels to optimize driving and walking routes, monitoring of parking spaces availability inside the city.
2. *Tracking Animals Movement*: recently, a large sensor network can be deployed to study the effect of micro climate issues in habitat choice of sea birds. Researchers located their sensors in burrows and used heat to detect the presence of nesting birds, providing invaluable data to biological researchers. The deployment is heterogeneous in that it employed burrow nodes and weather nodes.
3. *Industrial and Manufacturing Systems*: The IoT can be used in industrial control systems to enhance their performance through making them smarter with taking in consideration necessary factors like safety and availability to guarantee continues in business and save people life. The industrial control system can use the IoT for several purposes such as; auto-diagnosis of machines in control system, observing of toxic gas and oxygen levels inside chemical plants and monitoring of ozone levels during the drying meat process in factories of food engineering.
4. *Smart Agriculture*: Agriculture is a significant domain that provides people and society with food so that there are serious desires to improve the agriculture system through using smart technologies that are presented by the IoT. The IoT will improve operational efficiency and productivity in agriculture system. The benefits of using IoT in agriculture field are; monitoring soil moisture and trunk diameter in vineyards to control the amount of sugar in grapes and grapevine health, and control micro-climate conditions to maximize the production of fruits and vegetables and its quality.
5. *Healthcare*: The IoT can provide several benefits in healthcare domain. These benefits such as remote monitoring of patients, tracking of drugs, identification, and authentication of people. It also can use for the assistance of elderly or disabled people living independently and monitoring of conditions of patients inside hospitals.
6. *Trade and Logistics*: Innovative development in the IoT systems support to develop and manage products shopping from online websites. In addition to, using IoT in trade and logistics through embedding sensors and tags in roads and products for monitoring and tracking them. In Trade, IoT can be used for product tracking, monitoring of storage conditions and payment processing based on location. In logistics, the IoT can be used for observing of vibrations, strokes, container openings for insurance purposes, a search of individual items in big surfaces such warehouse.

2.4 Data Science

Data science, or more especially, big data analytics, become a hot and popular topic that has attracted attention among researchers in computer science and statistics. It concerns with a wide variety of data processing jobs, such as data analysis, data collection, data management, data visualization, and real-world applications. Today, the volume of data is increasing very quickly, the existing data processing tasks exceed the computing ability of classical computational models to store, validate, analyze, visualize, and extract knowledge. To analyze immense data, there are numerous complications, such as dynamical changes of data, a large volume of data, and data noise so that there is a serious need to develop novel and efficient methods to handle complex data analytics problems.

2.4.1 Data Science Definition

Data science is the interdisciplinary domain of computer science and statistics about scientific methods, processes and systems to extract knowledge from data in various forms, either structured or semi-structured and unstructured. The combination of computer science and statistics to take advantages of them to handle the massive amount of data in an efficient manner. The statistics are the science that concerns with collection, analysis, and organization of data. From the perspective of statistics, there are various objectives in data analyses such as predict the response/output of future input variables and deduce the association among response variables and input variables. While From the perspective of computer science, the data science is a process of data mining for converting raw data into useful knowledge and attempts to discover valuable patterns in large data storage.

2.4.2 Data Science Mission

The task of performing data science methods is to store, validate, analyze, visualize, and extract knowledge from the massive amount of data using computer science and statistical algorithms. Briefly, the data science area comprises of many sub areas, such as classification, clustering, and association analysis. The clustering and classification are two different types of basic problems, important methods in data mining research. Clustering is the process of grouping similar objects together. The data clustering analysis is a technique that divides data into several groups (i.e. clusters).The aim of clustering is to categorize objects being similar to one another in the similar cluster and place objects being distant from each other in dissimilar clusters. Data classification is a problem that finds the correct category(s) for data objects when a set of categories and a group of data set are given.

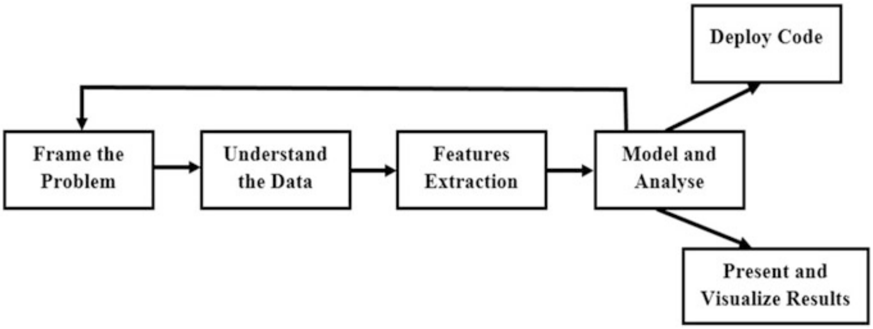

Fig. 2 Data science road map (DSRM)

2.4.3 Data Science Road Map

This part provides the idea of Data Science Road Map (DSRM) that refers to how to perform solving process for a data science problem as shown in Fig. 2. This DSRM consists of different stages as follows [14]:

1. *Frame the Problem*: It is a very important stage to understand the type of problem that will be solved using data science methods.
2. *Understand the Data*: Study and understand the data and the real-world things that it describes, which are related to the problem can help in choosing best methods to handle and manage this data in a given frame time.
3. *Feature Extraction*: It is the process of extracting features and hidden patterns from given data that will feed into the data science model for solving a certain problem.
4. *Model and Analyse*: In this stage, the data scientists are building a model which is suitable to a given problem as well as analyze the data sets related to the problem.
5. *Deploy Code*: Here, the data scientists write code, and they use several of the same tools as software engineers.
6. *Results Presentation and Visualization*: It is the final stage of designing and implementation of data science model.

2.4.4 Programming Languages for Data Science

There is various programming language which can be useful for data scientists. Here, will provide some of the most popular ones [14].

- *Python*: Python is a high-level scripting language, with functionality similar to Ruby and Perl and with an unusually clean and self-consistent syntax. Outside of the core language, Python has many open-source technical computing libraries that make it a powerful tool for analytics. Python is considered as one of the best programming language available for general-purpose use. It is also a very popular choice among data scientists, who feel like it balances the flexibility of a conventional scripting language with the numerical muscles of a good mathematics package.

- *R*: R is probably the most popular programming language among data scientists. Python is a scripting language designed for computer programmers, which has been augmented with libraries for technical computing. In contrast, R was designed by and for statisticians, and it is natively integrated with graphics capabilities and extensive statistical functions. One of the key reason to use R is just that there are so many special libraries that have been written for it over the years, and Python has not covered all the little special use cases yet.
- *MATLAB*: The data science community skews strongly toward open-source software, so good proprietary programs such as MATLAB often get less credit than they deserve. Developed and sold by the MathWorks Corporation, MATLAB is an excellent package for numerical computing. It has a more consistent syntax compared to R and more numerical muscle compared to Python. A lot of people coming from physics or mechanical/electrical engineering backgrounds are well-versed in MATLAB. It is not as well-suited to large software frameworks or string-based data munging, but it is best-in-class for numerical computing.

3 Cybercrimes Investigation in Internet of Things

Recently, The Internet of Things (IoT) has become an attractive research topic for academia and industry. It has several application domains such as medical, industry and military. The Internet of Things represents a network of connected devices or machines include mobile handsets, wireless sensors, refrigerators, cars, Radio Frequency Identification (RFID), fitness trackers, watches, eBooks, vending machines, and parking meters, and other types of devices are likely to grow exponentially over the next years. These devices are already generating, gathering and communicating enormous volumes of data about themselves, which is collated, curated, and harvested by a growing number of smart applications.

The IoT devices are considered as sources for massive volumes of data. The variety of these sources provides complex challenges to digital forensics community especially digital investigators who will be required to interact with this new technology to investigate IoT-based crimes. In the IoT environment, a lot of devices or machines are interconnecting together. This refers to the possibility of interconnecting various different threats and attacks such a malware can easily propagate through the IoT at an unprecedented rate. In the following design aspects of the IoT system, there may be various threats and attacks as follows [15]:

- *Data Perception and Collection*: In this part, typical attacks involve sovereignty and control, data leakage and authentication.
- *Data Storage*: Here many these attacks may happen such as denial-of-service attacks, data integrity, impersonation, and modification and tampering of sensitive data.
- *Data Processing*: In this stage, it may be computational attacks that have the objective of producing wrong data processing outcomes and results.
- *Data Transmission*: During the transmission process may occur severe type of attacks like session hijacks, routing attacks, flooding, and channel attacks. So, effective defense procedures and strategies are of the extreme significance to guarantee the security of the IoT infrastructure.

3.1 Digital Forensics in IoT Systems

In the last years, some researchers provide work related to the IoT Forensics area. Some of them explained the concept of the IoT Forensics while the others provided new methods for performing the digital investigation process in the IoT environment. Perumal et al. [16], proposed an integrated model which is planned based on triage model and 1-2-3 zone model for volatile based data preservation. This model started with the following authorization, planning and obtaining a warrant as fundamental steps in the digital forensic investigation process. Then starts to investigate the IoT infrastructure and finally after seizing the IoT device from the selected area or zone, the investigator completes the digital forensic method which includes a chain of custody, lab analysis, result and proof, and archive and storage.

Zawoad et al. [17], proposed a Forensics-Aware IoT (FAIoT) model for supporting digital forensics investigations in the IoT environment in a reliable manner. The FAIoT model provides secure evidence preservation module and secure province module as well as access to evidence using Application Programming Interface (API) that will reduce the challenge in performing investigation process. To facilitate the digital investigators a centralized trusted evidence repository in the FAIoT is used to ease the process of evidence collection and analysis. The IoT devices need to register this secure evidence repository service. The FAIoT architecture consists of three main parts as follows:

- *Secure Evidence Preservation Module*: This module can be used to monitor all the registered IoT devices and store evidence securely in the evidence repository. Also, segregating of the data according to the IoT devices and its owner will do in this module. Hadoop Distributed File System (HDFS) can be used to handle a large volume of data.
- *Secure Provenance Module*: This module guarantees the proper chain of custody of digital evidence by preserving the access history of the evidence.

- *Access to Evidence through API*: In this model, a secure read-only APIs to law enforcement agencies is proposed. Only digital investigators and the court member will have access to these APIs. Through these APIs, they can gather the preserved digital evidence and the provenance information.

Oriwoh et al. [18], they proposed two methods for digital investigation in IoT environment which are 1-2-3 Zones Digital Forensics and Next-Best-Thing Triage as follows:

1. ***1-2-3 Zones Digital Forensics***: This approach divides the IoT infrastructure into three areas or zones to help in performing digital investigation process. These zones are zone 1, zone 2 and zone 3 as follows:

    - *Zone 1*: This zone is called the internal zone that includes all IoT smart devices like a smart refrigerator and TV that can contain valuable data about committed crime in IoT infrastructure.
    - *Zone 2*: This zone includes all intermediate components between resides between the internal and external networks to support the communication process. These devices may be protection devices such as Intrusion Detection and Prevention Systems and Firewalls. The digital investigators can find evidential data that help them to extract facts about committed crime related to IoT.
    - *Zone 3*: This zone includes hardware and software components that reside in the external part of IoT infrastructures such as cloud services and other service providers that used to IoT devices and users. These components with hardware devices and software in zone 1 and zone 2 will help digital practitioners to perform their investigation mission in a timely fashion manner.

    This approach reduces the challenges that will be encountered in IoT environments and ensures that investigators can focus on clearly identified areas and objects in preparation for investigations.
2. ***Next-Best-Thing Triage***: The Next-Best-Thing Triage (NBT) can use in conjunction with the 1-2-3 Zones approach. This approach discusses to find an alternative source in the crime scene if it unavailable after a crime occurred in IoT environment. The NBT approach can be used to determine what devices were connected to the Objects of Forensic Interest (OOFI) and find anything which left behind the devices after they removed from the network. Direct access to the OOFI may not always be possible. Therefore, in such circumstances, the option of recognizing and considering the next best source of related evidence may have to be taken. The design of a technique of systematically deciding what this next best thing might be in different situations and scenarios can be the subject of further research.

4 Intrusion Detection in Internet of Things

Intrusion Detection System (IDS) is used to monitor network traffic, check for suspicious activities and notifies the network administrator or the system. In some instances, the IDS might also react to malicious or anomalous traffic and will take action such as barring the user or perhaps the IP address source from accessing the system. Detection and prevention malicious activities in Internet of Things environment becomes very important topic in the coming years.

A typical IDS is consist of sensors, an analysis engine, and a reporting system. Sensors are deployed at diverse network places or hosts [6, 19]. Their mission is to gather network or host data such as packet headers, traffic statistics, service requests, operating system calls, and file-system changes. The sensors send the gathered data to the analysis engine, which is responsible to investigate the gathered data and detect ongoing intrusions. When the analysis engine detects an intrusion, the reporting system generates an alert to the network administrator.

4.1 Attacks in Internet of Things

IoT infrastructure is exposed to various types of severe attacks both from internal and external so these attacks are mainly categorized by two types inside and outside attacks. In an inside attack, the attack can be originated by compromised or malicious nodes that are part of the infrastructure while in an outside attack, the attacker is not a part of the infrastructure. There are several types of attacks against IoT applications as follows [20]:

- *Sinkhole Attack*: In this attack, The criminal creates an attack by introducing false node inside IoT network where the malicious node attracts network traffic towards it. To launch these types of attack, a criminal node attracts all neighboring nodes to forward their packets through the malicious node by showing its routing cost minimum.
- *Wormhole Attack*: In this attack, the enemy node creates a virtual tunnel between two ends. An enemy node works as a forwarding node between two nodes. The two criminal nodes usually claim that they are one hop away from the base station. The wormhole attack can also be used to convince two different nodes that they are the neighbors by relaying packets between two of them.
- *Selective Forwarding Attack*: In this attack, criminal node works as a normal node but it selectively drops some packets. One of the simplest forms of selective forwarding attack is black hole attack where in it all packets are dropped by the criminal node.
- *Sybil Attack*: In this attack, the node has many identities. The routing protocol, detection algorithm, and cooperation processes can be attacked by a criminal node.

- *Hello Flood Attack*: In a network, the routing protocol broadcast hello message to announce its presence to its neighbors. A node which receives the hello message may assume that the source node is within its communication range and add this source node to its neighbor list.
- *Denial of Service (DOS) Attack*: This attack can damage the availability of resources to legitimate users. Such type of attacks, when launched by various criminal nodes is called Distributed Denial of Service (DDoS). This attack may affect the network resources, such as bandwidth and CPU time.

4.2 Categorization of IDS in Internet of Things

In [6], they classified Intrusion Detection in IoT regarding the following attributes: IDS placement strategy, detection method, security threat and validation strategy as shown in Fig. 3.

1. **IDS Placement Approaches**

In IoT infrastructure, the IDS can be located in the border router, in one or more dedicated hosts, or in every physical object. The advantage of placing the IDS in the edge router is the intrusion detection from the Internet against the devices in the physical domain. However, an IDS in the edge router might produce communication overhead between the LLN nodes and the edge router because of the IDS regular querying of the network state. There are three possible placement approaches for IDSs as follows:

1. *Distributed IDS Placement*: In this placement strategy, IDSs are employed in every single physical object of the LLN. The IDS deployed in each node must be optimized since these nodes are resource-constrained. In the distributed

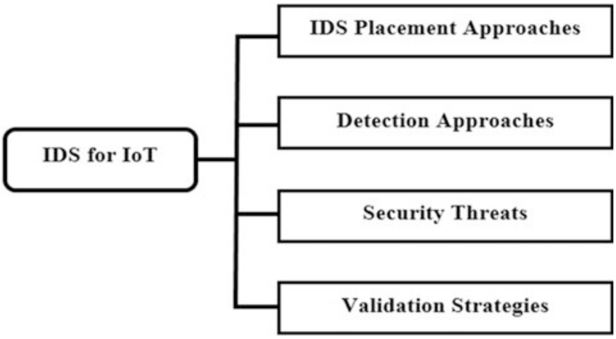

Fig. 3 Intrusion detection in internet of things

placement, the nodes may also be responsible for monitoring their neighbors. Nodes that audit their neighbors are called as watchdogs.

2. *Centralized IDS Placement*: In the centralized IDS placement, the IDS is located in a centralized component, for example, in the border router or a dedicated host. All the data that the LLN nodes collect and transmit to the Internet cross the border router along with the requests that Internet clients send to the LLN nodes. Consequently, the IDS placed in a border router can analyze all the traffic exchanged between the LLN and the Internet. However, analyzing the traffic that traverses the border router is not enough to detect attacks that involve only nodes within the LLN. Then, scientists must propose IDSs that can observe the traffic switched between LLN nodes, without ignoring the impact that this monitoring activity may have on low capacity nodes operation. Also, the centralized IDS may have trouble in auditing the nodes through an attack that compromises part of the network.

3. *Hybrid IDS Placement*: Hybrid IDS placement combines ideas of centralized and distributed placement to take benefit of their strong points and avoid their weaknesses. The first approach for hybrid placement organizes the network into clusters or regions, and only the main node in each host of cluster an IDS instance. Then, this node becomes responsible for auditing the other nodes of its cluster. In the second approach for hybrid placement, IDS modules are placed both in the edge router and in the remaining network nodes. The key difference of this approach to the first one is the presence of a central element. The IDS modules in the edge router are responsible for tasks that demand more resource capacity, while the IDS modules in regular nodes are usually lightweight.

2. Detection Approaches

Intrusion detection approaches in IoT are categorized into four types depending upon the detection mechanism which is anomaly-based, signature-based, specification-based and hybrid.

1. *Signature-based Approaches*: In signature-based approaches, IDSs detect attacks when the system or network behavior matches an attack signature warehoused in the IDS in-house databases. If any system or network activity matches with stored patterns/signatures, then an alert will be triggered. Signature-based IDSs are accurate at detecting known threats, and their technique is easy to understand. However, this approach is ineffective to detect new attacks and variants of known attacks, since a matching signature for these attacks is still unidentified.

2. *Anomaly-based Approaches*: Anomaly-based IDS compare the activities of a system at an instant against a normal behavior profile and produces the alarm whenever a deviation from normal behavior goes beyond a threshold. This

approach is effective to detect new attacks, in particular, those attacks related to misuse of resources. Nevertheless, whatever that does not match to a normal behavior is considered an intrusion and learning the entire scope of the normal behavior is not a simple task. Thereby, this method usually has high false positive rates. To construct the normal behavior profile, scientists usually employ statistical methods or machine learning algorithms that may be too heavy for low capacity nodes of IoT networks. Consequently, anomaly-based methods for IoT networks should take this particularity into account.

3. *Specification-based Approaches*: A specification is a group of rules and thresholds that express the expected behavior for network elements such as nodes, protocols, and routing tables. Specification-based approaches detect intrusions when network behavior deviates from specification definitions. Therefore, specification-based detection has the same purpose of anomaly-based detection: identifying deviations from normal behavior. However, there is one important difference between these methods: in specification-based approaches, a human expert should manually define the rules of each specification. Manually defined specifications usually provide lower false positive rates in comparison with the anomaly-based detection. Besides, specification-based detection systems do not need a training phase, since they can start working immediately after specification setup. However, manually defined specifications may not adapt to different environments and could be time-consuming and error-prone.
4. *Hybrid Approaches*: Hybrid approaches use ideas of signature-based, specification based and anomaly-based detection to maximize their advantages and minimize the impact of their disadvantages.

3. **Security Threats**

The goal of this part is to introduce how various types of attacks have been addressed in the IDS proposals for IoT. Enabling IoT solutions involves a composition of several technologies, services, and standards, each one with its security and privacy requirements. With this in mind, it is reasonable to assume that the IoT paradigm has at least the same security issues as mobile communication networks, cloud services, and the Internet. However, classical security countermeasures and privacy enforcement cannot be directly applied to IoT technologies due to three fundamental aspects: the limited computing power of IoT components, the high number of interconnected devices, and sharing of data among objects and users.

4. **Validation Strategy**

Validation consists of checking that the built model behaves with satisfactory accuracy within the study aims. There are several validation methods, and they may be distinguished by two sources of information: experts and data. While the use of experts provides a subjective and often qualitative model validation, the use of data may allow a quantitative and more objective validation. The objective of this part is to investigate the validation strategy employed in the intrusion detection methods for IoT. Such criteria could be a starting point for evaluating the maturity level of this domain. For this purpose, the classification of validation methods can be as follows:

- *Hypothetical*: theoretical examples, having unclear relation to actual phenomena and degree of realism.
- *Empirical*: empirical approaches, such as systematic experimental collecting of data from operational settings.
- *Simulation*: Simulation approaches of some IoT scenario.
- *Theoretical*: formal or precise theoretical arguments to support results.
- *None*: no validation approaches are employed.

5 Applying Data Science Methods for Cybercrimes Investigation and Intrusion Detection in Internet of Things

Data science and knowledge discovery methods become significant topics in security domain where they can assist security professionals and digital investigators to detect and investigate cybercrimes as well as introduce solutions to malware and threat prediction, detection, and prevention at an initial stage. Knowledge discovery is known as data mining which refers to the process, in which hidden, unknown and potentially valuable information are extracted from massive, noisy, incomplete, and random data. The extracted information will be used for deriving novel insights, promoting business and scientific events, and speeding up and advancing scientific innovation.

At the present time, the furthermost imperative data mining algorithms mainly cover clustering, classification, regression, association analysis, statistical learning and linking mining. The methods of data science can use in the areas of cybercrimes investigation and intrusion detection in IoT environment to provide effective performance in the investigation, detection, prevention and prediction of IoT-based crimes in a timely fashion manner.

5.1 Data Science Methods for Cybercrimes Investigation

Data science and big data analytics have become significant paradigms to investigate IoT-based cybercrimes. Data science methods can use to analysis generated data from Internet of Things to investigate the crimes as well as predict the new coming severe attacks and crimes in the future. Lately, there are some challenges in digital forensics such as [21]:

- *Visualization of large amounts of data to the tribunal*: Visualization of finding from the analysis of digital evidence is vital for presenting the results in a court of law. Presentation and visualization of large of data is a problem faced digital practitioners and examiners in the digital forensics area so that there is a need for novel methods to deal with the massive size of data that generated from the crime scene in forensically and timely fashion way.
- *Search in a large amount of data*: Search is a commonly used application in digital forensics for extract valuable proof from digital evidence. Classical data is a summarization of structured data, which is enhanced for fast access and well-defined queries. Standard search methods are good for classical data. However, big data is unstructured or semi-structured. Consequently, general search procedures are not applicable to big data, especially for text, images, and videos that are structured for storage and display but not structured according to the content. Big data search objectives to extract convenient evidence from enormous data, and to facilitate decision-making. How to get value out of big data is a big challenge.
- *Storing and rapid indexing of massive amounts of data*: Conventional data storage not suitable for a large amount of data that created as a result of big data idea. This large amount of data that generated from different data sources need high size storage capacity to store for the digital investigation purpose. In recent times, data become big so faster indexing of the data is a challenge for digital investigators. In order to rapid indexing to the analysis of the large size of data, there is a need for faster methods and devices that have the ability analysis data within a given time frame.

From the aforementioned challenges, there is a need for employing data science, data mining and big data analytics methods in cybercrime investigation area because they have many advantages to support the digital investigation. Current digital forensic methods cannot do the extraction and analysis activities for the massive amount of data in an efficient manner so that there is need to scale up these methods to be suitable to the huge size of data. Some advantages of data science methods for digital forensics can be as the following:

- Enhance the analysis of evidential data which extracting from the crime scene.
- Diminish processing time of huge data analysis.
- Improve information quality associated with data analysis.

- Better utilization of existing computing, processing and storage resources.
- Decrease costs and save the time of the digital investigation.

The combining data science and digital forensics is to solve the crucial challenge of analyzing immense amount of data in actionable time while at the same time preserving forensic principles in order for the results to be presented in a court. After introducing digital forensics and data science in the background section explores the challenges to propose how data science methods can be adapted to the unique context of cybercrime investigation, ranging from the evidence managing through Map-Reduce to machine learning approaches for triage and analysis of a large amount of forensic data.

Data science using machine learning techniques can use to handle several complications that currently exist in digital forensics such as extracting and analyzing digital evidence from the crime scene. Using techniques that can automatic extraction of complex data representations or features in digital forensics can enhance the process of analyzing large amount of digital evidence in short time with high quality and accuracy of results. These techniques motivated by digital investigators and examiners to use in the forensic analysis stage. There are a number of topics in cybercrimes investigation in Internet of Things where machine learning techniques can be used as follows:

- *Data Indexing*: Large-scale volume of data such as text, image, video, and audio are being extracted and collected from different sources that can make investigation process harder especially when crimes related to environment such as Internet of Things. These huge amounts of data need semantic indexing rather than being stored as data bit strings. Semantic indexing presents the data in a more efficient manner and makes it useful as a source for knowledge discovery and understanding.
- *Pattern Recognition*: Pattern recognition is an important area in machine learning that working on extracting patterns from input data. Supervised data that trained from labeled data and unsupervised learning that discover unknown patterns. Both of them can use in the pattern recognition. It is used to identify pattern or feature in data through determining and specify types or clusters of data. The pattern recognition can help in digital forensics\ for performing detecting a pattern in an e-mail message which indicates malicious code like spam or virus. Likewise, can be used to discover identities in digital evidence that is extracting from the crime scene.
- *Authorship Identification*: Criminals can use fake emails for performing activities without tracing them through hiding their identity. Authorship Identification is an important technique which used to solve this problem by identifying the authors of these fake e-mails that can help digital investigators and examiners to perform investigation process in a timely fashion manner.
- *Image Region Forgery Detection*: In recent time, the number of tampered images is increased incredible way due to the use of social networks like Facebook, Flicker, and Twitter. These tampered images can be shared easily by

the users that may lead serious consequences so the authenticity of digital images is urgently needed. The presence of tampered images is an important topic in digital forensics.

- *File Fragments*: Detection of data from disks is challenging faced by digital investigators to recover data from disk. The data when deleting from disk is not permanent where they simply mark each block of the file as unallocated and available for use. The process of recovering unallocated data called as 'file carving'. Machine learning introduces approaches to recognizing the file types of file fragments for the purpose of file carving for the reconstruction of partially erased files on disk into whole files.

5.2 Data Science Methods for Intrusion Detection

Intrusion detection refers to the procedure of auditing and analyzing the events occurring in a system to detect malicious behaviors. The intrusion detection process involves detecting a set of nasty actions that compromise available resources. In last years, there is a serious need for new data science methods for analysis of sophisticated attack in Internet of Things environment. Current methods suffer from evaluation, comparison, and deployment which originate from the scarcity of adequate publicly available network trace data sets. Also, publicly existing datasets are either out-of-date or generated in a controlled environment.

Data science involves various analytical techniques such as machine learning, artificial intelligence, and data mining that are useful for extracting features from data sets. There are many techniques which can use to detect unknown new attacks. These techniques such as prediction, classification, clustering and relation rule.

- *Prediction*: It is a technique that predicts the future possibility and trend. Regression analysis is a representative prediction technique. Researchers can predict attack possibilities using regressing analysis. Regressing analysis can predict similar behaviors from collected attack logs.
- *Classification*: It is a technique that predicts the group of a new attack from huge data. Classification helps security administrator to decide the direction of protection and analysis.
- *Clustering*: It is an unsupervised technique where the data set is divided into sub parts sharing same properties. The clustering process is used for finding similarities in data and putting similar data into sets. Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups. Clustering procedures are used extensively not only to organize and categorize data but are also suitable for data compression and model construction.

- *Relation Rule:* It is a technique that discovers hidden relations among data. The action of discovering relation rule is named association analysis or link analysis. The relation from time flow is named as sequence rule. This analysis technique can determine abnormal behavior by analyzing user or process behaviors.

Data scientists and researchers can make great achievements in this area through designing novel threat detection models that can combine data science, machine learning, and behavioral analysis. They can recognize the underlying purpose of traffic, detect attack behaviors in real time IoT applications. This model can be applied directly to network traffic to expose underlying attack features that unknown. Supervised and unsupervised machine learning algorithms can help in discovering uncover new attack behaviors.

In order to develop and propose new efficient intrusion detection systems based on data science methods, it is required to work on attacks datasets to test and evaluate their innovation detection and prevention models. One of the most common data sets for developing attacks and intrusions detection system is KDD CUP 99 dataset [22]. The KDD CUP 99 has been most commonly used in attacks detection using data mining techniques. The KDD data set contains 10% of original dataset that is approximately 494,020 single connection vectors each of which has 41 features and is labeled with exact one specific attack category. Every vector is labeled as either normal or an attack, with accurately one specific attack category. The simulated attack may be one of the following four categories [23]:

1. *Denial of Service (DOS) Attack*: In this attack, the attacker makes computing or memory resources busy to allow the legitimate request, or deny the access legitimate of users to the system. The DOS involves attacks such as ‘land’, ‘smurf’, ‘neptune’, ‘pod’, ‘back’ and ‘teardrop’.
2. *Users to Root (U2R) Attack*: In this type of attack, the attacker starts out with access to a normal user account on the system and is able to exploit some vulnerability to obtain root access to the system. The U2R involves attacks such as ‘loadmodule’, ‘rootkit’, ‘buffer_overflow’ and ‘perl’.
3. *Remote to Local (R2L) Attack*: In this attack, the attacker sends packets to the system over a network but who does not have an account on that system and exploits a vulnerability to gain local access as a user of that system. The R2L contains attacks such as ‘warezclient’, ‘imap’, ‘multihop’, ‘guess_passwd’, ‘warezmaster’, ‘spy’ ‘ftp_write’, and ‘phf’.
4. *Probing Attack*: In this category, the attacker attempt to gather information about the network of computers for the apparent purpose of circumventing its security. This attack covers the following attacks: ‘portsweep’, ‘satan’, ‘nmap’, and ‘ipsweep’.

6 Conclusions and Future Directions

In recent times, data science become a very significant topic that has attracted attention numerous researchers who interest in solving problems of intrusion detection and cybercrimes detection in Internet of Things environment. Therefore, this chapter introduced the principles of Digital Forensics, Intrusion Detection and Internet of Things as well as exploring data science concepts and methods that can help the digital investigators and security professionals to develop and propose new techniques and methods that can be adapted to the unique context of Internet of Things infrastructure for performing intrusion detection and cybercrimes investigation. As future research work, researchers may focus on some issues such as follows:

- To explore advantages and disadvantages of various current intrusion detection strategies.
- To improve the security of alert traffic, alert correlation, and autonomic management systems.
- To develop new/novel detection model for automated risk management through linking machine learning procedures, data science methods, and behavioral analysis.
- To propose invulnerable-based heuristic IDSs using neural and fuzzy methods to control the sensitivity of alerting malicious intrusions to decrease false alarm rate.
- To develop advanced feature extraction and selection algorithms for improving the performance of detection models will be positively affected. And also, help to construct strong and efficient classifier to detect new attacks and threats.
- To use deep learning methods for predictions and classification of attacks.
- To improve the performance of real-time intrusion detection systems.
- Use Big Data analytics tools and platforms such as Apache Hadoop ecosystems and Apache Spark to enhance and increase analysis performance in intrusions and attacks detection.

References

1. Palmer, G.: A road map for digital forensic research. First Digital Forensic Research Workshop, Utica, New York (2001)
2. McKemmish, Rodney: What is forensic computing?. Australian Institute of Criminology, Canberra (1999)
3. Khan, Minhaj Ahmad: A survey of security issues for cloud computing. J. Netw. Comput. Appl. **71**, 11–29 (2016)
4. Oscar Serrano, C.I.S.A.: CISSP CISM, and Luc Dandurand. Big Data Analytics for Sophisticated Attack Detection (2014)
5. Jabez, J., Muthukumar, B.: Intrusion detection system (IDS): anomaly detection using outlier detection approach. Procedia Comput. Sci. **48**, 338–346 (2015)

6. Zarpelão, B.B., et al.: A survey of intrusion detection in internet of things. J. Netw. Comput. Appl. (2017)
7. Kawamoto, Y., Nishiyama, H., Kato, N., Yoshimura, N., Yamamoto, S.: Internet of things (IoT): present state and future prospects. IEICE Trans. Inf. Syst. E **97**(10), 2568–2575 (2014)
8. Gubbi, J., Buyya, R., Marusic, S., Palaniswamia, M.: Internet of Things (IoT): a vision, architectural elements, and future directions. Futur. Gener. Comput. Syst. **29**(7), 1645–1660 (2013)
9. Atzori, L., Iera, T., Morabito, G.: The internet of things: a survey. Comput. Netw. **54**(15), 2787–2805 (2010)
10. Li, S., Xu, L.D., Zhao, S.: The internet of things: a survey. Inf. Syst. Front. **17**(2), 243–259 (2014)
11. Andrew, W., Agarwal, A., Xu, L.D.: The internet of things a survey of topics and trends. Inf. Syst. Front. **17**(2), 261–274 (2014)
12. Farooq, M.U., Waseem, M., Mazhar, S., Khairi, A., Kamal, T.: A review on internet of things (IoT). Int. J. Comput. Appl. **113**(1), 1–7 (2015)
13. El-Din, H.E., Manjaiah, D.H.: Internet of things in cloud computing. Internet of Things: Novel Advances and Envisioned Applications. Springer International Publishing, pp. 299–311 (2017)
14. Cady, F.: The Data Science Handbook. Wiley (2017)
15. Giuliano, R., et al.: Security and Privacy in Internet of Things (IoTs): Models, Algorithms, and Implementations (2015)
16. Perumal, S., Norwawi, N.M., Valliappan, R.: Internet of things (IoT) digital forensic investigation model: top-down forensic approach methodology. In: 2015 Fifth International Conference on Digital Information Processing and Communications (ICDIPC). IEEE (2015)
17. Zawoad, S., Hasan, R.: FAIoT: towards building a forensics aware eco system for the internet of things. In: 2015 IEEE International Conference on Services Computing (SCC). IEEE (2015)
18. Oriwoh, E., et al.: Internet of things forensics: challenges and approaches. In: 2013 9th International Conference Conference on Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom). IEEE (2013)
19. Qiu, T., Zhang, Y., Qiao, D., Zhang, X., Wymore, M.L., Sangaiah, A.K.: A robust time synchronization scheme for industrial internet of things. IEEE Trans. Ind. Inform. (2017). https://doi.org/10.1109/TII.2017.2738842
20. Sherasiya, T., Upadhyay, H.: Intrusion detection system for internet of things. Int. J. Adv. Res. Innov. Ideas Educ. (IJARIIE) **2**(3) (2016)
21. Uma, M., Salisu, S.: The use of big data in the field of digital forensics investigations (comparative study between digital forensics in uk and nigeria). Int. J. New Technol. Sci. Eng. **2**(4) (2015)
22. Tavallaee, M., et al.: A detailed analysis of the KDD CUP 99 data set. In: IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE (2009)
23. Siddiqui, M.K., Naahid, S.: Analysis of KDD CUP 99 dataset using clustering based Data Mining. Int. J. Database Theory Appl. **6**(5), pp. 23–34 (2013). https://doi.org/10.14257/ijdta.2013.6.5.03

http://www.springer.com/978-3-319-70687-0

Cognitive Computing for Big Data Systems Over IoT
Frameworks, Tools and Applications
Sangaiah, A.K.; Thangavelu, A.; Meenakshi Sundaram,
V. (Eds.)
2018, XVI, 375 p. 81 illus., 51 illus. in color., Softcover

ISBN: 978-3-319-70687-0