Chapter 7

Adaptive and Intelligent Data Collection and Analytics for Securing Critical Financial Infrastructure

By Habtamu Abie, Svetlana Boudko, Omri Soceanu, Lev Greenberg, Aidan Shribman, Beatriz Gallego-Nicasio, Enrico Cambiaso, Ivan Vaccari and Maurizio Aiello

DOI: 10.1561/9781680836875.ch7

Published in *Cyber-Physical Threat Intelligence for Critical Infrastructures Security: A Guide to Integrated Cyber-Physical Protection of Modern Critical Infrastructures* by John Soldatos, James Philpot and Gabriele Giunta (eds.). 2020. ISBN 978-1-68083-686-8. E-ISBN 978-1-68083-687-5.

Suggested citation: Habtamu Abie *et al.* 2020. "Adaptive and Intelligent Data Collection and Analytics for Securing Critical Financial Infrastructure" in *Cyber-Physical Threat Intelligence for Critical Infrastructures Security: A Guide to Integrated Cyber-Physical Protection of Modern Critical Infrastructures*. Edited by John Soldatos, James Philpot and Gabriele Giunta. pp. 104–140. Now Publishers. DOI: 10.1561/9781680836875.ch7.

now
the essence of knowledge

This chapter presents the FINSEC adaptive and intelligent data collection and analytics system for securing critical financial infrastructure. It enhances the intelligent, resilient, automated, efficient, secure, and timely manner the collection and analysis of security-related data for securing cyber-physical financial infrastructure and services. Making security data collection and analysis intelligent and capable of quickly spotting, learning from, and addressing zero-day threats is essential to economizing of resources and accessing the right information at the right time. This is achieved through the configuration of configurable collection probes and the adaptation of different collection strategies. The chapter further addresses how, inter alia, (i) the nature and quality of collected data affects the efficiency and accuracy of methods of attack detection and defense, (ii) the detection capability can be improved by correlating wide-ranging data sources and predictive analytics, (iii) the rate of the data collection at the various monitoring probes is tuned by managing the appropriate levels and types of intelligence and adaptability of security

monitoring, (iv) the optimization of bandwidth and storage of security information can be achieved by rendering adaptiveness and intelligence and by integrating smart security probes and a set of adaptive strategies and rules, and (v) the increased automation is achieved through a feedback loop of collection, detection, and prevention that allows the early detection and prevention of security compromises and consistently makes security analysis more effective.

7.1 Introduction

Cyber-physical attacks are growing rapidly and posing a substantial risk to the stability of the overall financial sector. Attacks are increasing in number, scope, and sophistication, making it difficult to predict their total impact. The nature and frequency of cyber risks have changed rapidly in the directions not anticipated before, and more risk-managers are becoming aware of the value of engaging with Fintech R&D to keep track of new types of attack surfaces and risk management options, such as FINSEC is addressing. Leading security researchers are coming to the same conclusions (e.g., the state-of-the-art 2019 Cyber Risk Outlook report [1]). In this chapter, we look further ahead than this Cambridge report, blending risk policy, risk technology, and risk-management best practice. Our findings include:

- It is essential to track and maintain the security of critical financial infrastructure and services through the collection and analysis of security-related data in an intelligent, resilient, efficient, secure, and timely manner.
- Making security data collection and analysis intelligent and capable of quickly spotting, learning from, and addressing zero-day threats is essential to economizing resources and accessing the right information at the right time through the configuration of data collection probes and the adaptation of different collection strategies.
- The nature and quality of collected data affects the efficiency and accuracy of methods of attack detection and defense.
- The detection and defense capability can be greatly improved by correlating wide-ranging data sources and by predictive analytics.

Adversaries may attack financial services, damage infrastructure, manipulate critical information, therefore causing serious financial loses. Considering the risks of a large-scale financial network system, it is important to calculate not only the risks of separate nodes but also the risks from connections. Furthermore, adaptive attackers will adapt their strategies to the current security situation and to newly deployed countermeasures. Such emerging attacks can become very sophisticated and can be coordinated, persistent, collaborative, or cooperative with specialized

attack expertise. Therefore, there is a need to implement adaptive and intelligent data collection and analytics to cope with a constant update of attack vectors.

The amount of data collected by financial organizations to maintain security is growing every day, and these huge amounts of data can no longer be stored efficiently or processed in real time. Therefore, due to a high dimensionality of data collected from cyber-physical systems, a constant growth of data due to improvements and exposure to new vulnerabilities, and a constant update of attack vectors, Deep Learning (DL)-based security models are essential for adaptability and extendibility with the data drift, continuous discovery of new system threats, and vulnerabilities [2]. In this chapter, we present a model for developing an adaptive and intelligent data collection and analytics that adapts the collection rate and storage state configuration to the analytical systems, threats detected by those systems over time, and economizing the cost of collection and storage resources.

The rest of the chapter is organized as follows: Section 7.2 briefly reviews related work. Section 7.3 sets the scene by describing data collection and analytics. Section 7.4 presents the architecture of adaptive and intelligent data collection and analytics and its implementation in the overall FINSEC Reference Architecture (RA) highlighting its peculiar characteristics. Section 7.5 presents the adaptive data collection strategies which are used to economize use of resources and optimize bandwidth and collection rate. Section 7.6 describes the implementation of different modules and the validation of the predictive analytics algorithms for intelligent processing. Finally, Section 7.7 concludes the chapter.

7.2 Related Work

Adaptive data collection refers to the collection of security-related data to improve collection efficiency, ensure collection accuracy, reduce the amount of collected data to minimize the effect of data collection, and automate the data collection by adjusting to different environmental contexts and situations. Several authors address adaptive data collection in different settings.

Lin *et al.* [3] present the design and implementation of an adaptive security-related data collector based on network context in heterogeneous networks, and they used adaptive sampling algorithm to reduce the amount of collected data. The authors argue that sampling methods to collect data and the collection frequency need to be determined according to specific conditions. For instance, if the data variation is large, the collection interval should be reduced, so as to reflect the variation trend of data; and if the data variation is small, the collection interval can be increased, so as to reduce the amount of data collected while ensuring the accuracy of data collection. They propose an adaptive collection frequency

adjustment strategy based on predicted variation ratio. They argue that regression algorithms can be used for prediction, such as linear regression, support vector regression (SVR), logistic regression, KNN regression, etc. They further argue that data variation can also be represented by calculating the ratio of predicted accuracy, which is the ratio of the predicted value of the data to the real value of the data. When the predicted value of the data is close to the real value, it indicates that the data variation is small, and when the predicted value of the data is very different from the real value, the data variation is large.

Habib *et al.* [4] investigated self-adaptive data collection and fusion for body sensor networks. Their approach uses an early warning score system to optimize data transmission and estimates in real time the sensing frequency, and uses a data fusion model using a decision matrix and fuzzy set theory. Their adaptive sampling algorithm adapts the sampling rates of sensors to the vital sign dynamic evolution. An adaptive data collection protocol was proposed in [5], which collects periodically sensor readings and prolongs the lifetime of a periodic sensor network. Authors' sampling rate adaptation is based on the similarity between periods of cycles using Euclidean distance measure to adapt its rate of sampling according to the dynamic modification of the monitored environment. An efficient adaptive sampling approach based on the dependence of conditional variance on measurements varies over time as proposed in [6], which adapts sampling rates to the physical changing dynamics and minimizes over-sampling, and improves resource efficiency of the overall network system. An adaptive sampling approach for energy-efficient periodic data collection in sensor networks is proposed [7]. The approach provides each sensor node the ability to identify redundancy between collected data over time, by using similarity functions and allowing adaptive sampling rate.

Ji and Ni [8] present an adaptive data collection method based on the network data correlation and variation routines. Their method selects the data collection in association with network data variation and adjusts collection frequency based on the ratio of the data variation amplitude. It can adjust data collection according to network load to reduce the burden on network bandwidth and processing resources. The frequency adjustment strategy can reduce data collection times when the data vary gently and increase data collection times when the data vary dramatically. Tang and Xu [9] investigate data collection strategies in lifetime-constrained wireless sensor networks. Their objective is to maximize the accuracy of data collected. They developed adaptive update strategies for both individual and aggregate data collections.

Lin *et al.* [10] highlight the challenges posed in collecting security-related data, which indicates relevance to security, safety, privacy, and trust, in the big data era. Their examples of making data collection difficult are due to its 5Vs (volume, variety, value, velocity, and veracity) characteristics and further the

5G networks' characteristics of being heterogeneous, supporting device-to-device, machine-to-machine and other communication technologies, and different networks such as Internet, Mobile Ad hoc Networks, mobile cellular networks and wireless sensor networks. Security-related data fundamentally affects the efficiency and accuracy of attack detection and defense methods. Jing *et al.* [11] survey existing studies about security-related data collection and analytics for measuring the Internet security. They argue that for measuring the security of the internet and detecting the Internet attacks, collecting different categories of data and employing methods of data analytics are essential. A number of surveys of data collection approaches exist [3, 10–15], addressing different settings.

As demonstrated above, there exist many adaptive data collection methods using different strategies. However, few of them are aimed at adaptive multi-layer data collection applying artificial intelligence and deep learning. This chapter addresses adaptive and intelligent multi-layer data collection through the correlation of wide-ranging data sources and predictive analytics to improve the detection capability, the improvement of the quality of collected data that affects the efficiency and accuracy of methods of attack detection and defense, the rendering of adaptiveness and intelligence, and the integration of smart security probes and a set of adaptive strategies and rules. It also addresses the different means for physical and cybersecurity as means of tuning the rate of the data collection at the various monitoring probes.

7.3 Data Collection and Analytics

7.3.1 Requirements

7.3.1.1 Data collection requirements

Before going through data collection in a physical system, one may verify a set of requirements aspects that are identified and summarized below, but more details can also be found in [15]:

- **Efficiency:** On one hand, the collected data should be compact, the unnecessary data that are useless in attack detection should not be collected. On the other hand, the needed data should be collected in a real-time and high-speed manner to decrease the time delay of attack detection.
- **Privacy:** In the data collection process, the sensitive information of some particular data should be protected.
- **Resource consumption:** The consumption of resources including power, memory, and network bandwidth in the process of data collection and data communication should be well considered.
- **Adaptability and Intelligence:** The data collection process should be adaptable to the context of the physical and cyber-world, as well as to the

security context. In particular, the rate of information acquisition/collections, along with the type of data collected, should be adaptable to changing security contexts. Adaptability should be performed in an intelligent way, i.e., towards optimizing the amount of information available for the security task at hand, while ensuring availability of the proper information.

- **Configurability:** To support adaptability and configurability in data collection, the data collection systems to the used in the project (e.g., probes) must be configurable.
- **Automation:** To automate the data collection and adaptation by adjusting to different environmental contexts and situations. Machine Learning (ML) techniques are helpful for implementing automatic adaptable solutions capable of adjusting to new situations and timely reacting in the face of threats and anomalies [16].

The authors [10] specify 13 functional requirements and 5 security requirements, and 9 functional objectives and 6 security objectives, and the relationship between these.

7.3.1.2 Quality attributes for data analytics

The authors in [17] present a systematic review aimed at identifying the most frequently reported quality attributes and architectural tactics for big data security analytic systems. Their findings are twofold: (i) identification of most frequently reported quality attributes and the justification for their significance for big data cybersecurity analytic systems; and (ii) identification and codification of architectural tactics for addressing the quality attributes that are commonly associated with big data cybersecurity analytic systems. The identified tactics include six performance tactics, four accuracy tactics, two scalability tactics, three reliability tactics, and one security and usability tactic each.

- ***Performance*** is a measure of how quickly a system responds to user inputs or other events.
- ***Accuracy*** is a measure to which a system provides the right results with the necessary degree of precision.
- ***Scalability*** is a measure of how easily a system can grow to handle more user requests, transactions, servers, or other extensions.
- ***Reliability*** is a measure of how long a system runs before experiencing a failure.
- ***Usability*** is a measure of how easy it is for people to learn, remember, and use a system.

- ***Interoperability*** is a measure of how easily a system can interconnect and exchange data with other systems or components.
- ***Adaptability*** is a measure of how easily a system adapts itself to different specified environments using only its own functionality.
- ***Modifiability*** is a measure of how easy it is to maintain, change, enhance, and restructure a system.
- ***Generality*** is a measure of the range of attacks covered by a security analytic system.
- ***Privacy assurance*** is the measure of the ability of a system to carry out its business according to defined privacy policies to help users trust the system.
- ***Security*** is the measure of how well a system protects itself and its data from unauthorized access.
- ***Stealthiness*** is the measure of the ability of a security analytic system to function without being detected by an attacker.

7.3.2 Data Sources

Data sources from which security event data are collected include, but are not limited to, network traffic data, firewall logs, web logs, system logs, router access logs, database access logs, and application logs, system statistics, etc. [11].

7.3.3 Data Collection Categories

The following categories of data collection can be distinguished [11].

Packet-level data: A packet consists of a packet header and a packet payload. They are generated when using protocols like TCP, UDP, ICMP, etc. Based on this definition, a classification of these data for detecting DDoS and Worm attacks can be as: Source/Destination IP address, Source/Destination port, Time to live, Timestamp, Packet payload, Packet size, and Number of packets.

Flow-level data: In high-speed networks with rates up to hundreds of Gigabit per second, collection of packet-level data requires expensive hardware. Thus, flow-level data was introduced and can be considered as a stream of packets. The flow-level data is classified into Flow count, Flow type, Flow size, Flow direction, Flow duration, and Flow rate.

Connection-level data: A connection is defined as the aggregated traffic between two IP addresses from the perspective of a specific network. A connection will contain many flows. Thus, a difference between a connection and a flow is the flow does not have size restriction, that is to say, the flow is generated even if a single

packet has been exchanged. But, a connection is generated by at least two packets. The connection level data can be divided into the following types: Connection size, Connection duration, Connection count, and Connection type.

Host-level data: This data is collected from a host. This data provide comprehensive knowledge of system events as it records host activities, changes, resource consumption, etc. These changes are widely used in Host-based IDS. We mention in the following two commonly used types of host-level data in attack detection: CPU and Memory usage and Operation log.

7.3.4 Security Probes

Security probes are created to capture and assess the overall security of servers, networks, databases, etc. and to generate events when they find problems, and have the following abilities:

Topology probes: Probes that have the ability to capture network topology, interface, bridge, namespace attributes. Examples include ethtool (a utility for Linux kernel-based operating system for displaying and modifying some parameters of network interface controllers and their device drivers), Network system simulation software (this includes Software-Defined Network or similar software to simulate the real network functions; An example is the Open vSwitch Database management protocol), Simple Network Management Protocol (an Internet Standard protocol for collecting and organizing information about managed devices on IP networks and for modifying that information to change device behavior), Telnet (protocol to provide a bidirectional interactive text-oriented communication facility, which can be used to connect to network equipment and extract management data), Network Interface Filtering Card (a hardware-based probe that can be remotely configured to focus in more detail on selected traffic and/or filter out malicious forms), etc.

Flow probes: Probes that have ability to follow a flow along a path in the topology. Examples include sFlow (sampled flow) (an industry standard for packet export at Layer 2 of the OSI model), Data Plane Development Kit (a set of data plane libraries and network interface controller drivers for fast packet processing, currently managed as an open-source project under the Linux Foundation), libpcap (commonly used packet capture library, which also defines the de facto external format for packets), sCap (a more efficient implementation of the standard libpcap, using shared memory and so-called subzero packet copy), Internet Protocol Flow Information Export (a protocol for exporting Internet Protocol flow information from routers, probes and other devices), NetFlow (a feature of Cisco routers that provides the ability to collect IP network traffic as it enters

or exits an interface), Flowmon probe (a hardware-based probe that uses IPFIX protocol), etc.

7.3.5 Predictive Security Analytics for Adaptive Data Collection

Predictive analytics are used to predict security attacks, threats, and anomalies. Based on the predicted security events, mitigation measures can be triggered, for example, to adapt the data collection rate, close a door, etc. It requires constant monitoring, capturing, and processing large amounts of various data. These data is often redundant. Thus, the storing and processing resources are used unnecessary, and the same prediction results can be achieved with significantly less data. It is, therefore, important to develop lightweight predictive data analytics that can give earlier indications about possible cyberattacks based on less data amount and processing. This will allow reducing the amount of collected and processed data while maintaining the required level of threat detection. We need to select the algorithms that give best prediction results and can, therefore, function as a base for the predictive analytics. For this purpose, several machine learning, deep learning (DL), artificial intelligence (AI) algorithms are to be selected and tested using different datasets available online.

DL or deep neural networks are especially relevant for scenarios where massive datasets are collected. One of the principal DL features is the ability of a DL model to adapt to the behavior of systems to previously unseen scenarios in cybersecurity, thus ensuring generalization of the models [2], which is one of the key goals of AI. Trust and explainability are two other important features to ensure trustworthiness of AI-based cyber systems. Recent research in Explainable AI (XAI) successfully showed how deep neural-network-based intrusion detection systems can help in improving user trust [18]. Adversarial learning offers an approach to increase our understanding of these models. Adversarial learning exploits how a DL system can be "fooled" to wrong conclusions. This knowledge strengthens the system against incorrect intrusion detection decisions. Hence, trust of the system is increased and explainability is improved [18].

Berman *et al.* [19] survey DL methods for cybersecurity applications covering a broad array of attack types including malware, spam, insider threats, network intrusions, false data injection, and malicious domain names used by botnets. They discussed the DL architecture and training process for popular and emerging methods ranging from RNNs to GANs and their application to a wide variety of these cybersecurity attack types.

In this Project, we among other things use AI-based (i.e., deep learning mechanisms) predictive analytics that enable us the identification of complex attack patterns.

7.4 Adaptive and Intelligent Data Collection and Analytics

This section first presents the architecture overview and its implementation in the overall FINSEC Reference Architecture (RA) followed by the descriptions of the various features and services. The security of critical financial infrastructure and services must be tracked and maintained through the collection and analysis of security-related data in an intelligent, efficient, secure, and timely manner. Making security data collection and analysis intelligent and capable of quickly spotting, learning from, and addressing zero-day threats is essential to economizing of resources and accessing the right information at the right time through the configuration of configurable data collection probes and the adaptation of different collection strategies.

The nature and quality of collected data affects the efficiency and accuracy of methods of attack detection and defense. The detection capability can thus greatly be improved by correlating wide-ranging data sources and by predictive analytics. Managing appropriate levels and types of intelligence and adaptability of security monitoring is achieved through different means for adaptive data collection and predictive analytics. This is important for physical and cybersecurity as a means of tuning the rate of the data collection at the various monitoring probes. The cyber and physical data need to be correlated taking the latency of communication into account.

7.4.1 Adaptive and Intelligent Data Collection and Analytics Architecture

Figure 7.1 shows the architecture of the multi-layer adaptive and intelligent data collection and analytics, which extends the classical data collection and analytics process that includes data collection, data parse, data analysis, and data processing. The approach makes this process adaptive by introducing feedback control loop and letting the data collection depends on the result of the last data processed. Adaptability refers to how a collection mechanism can adjust to different environmental contexts and situations.

In Figure 7.1, the process modules include **Monitor** (data collector), **Analyser** (data parser & analyser), **Adapter** (data processor), and **Multi-layer Probes** (Implemented FINSEC Probes). The arrow between modules is data flow and control direction.

The FINSEC project integrates smart security probes and a set of adaptive strategies for the multi-layer data collection functionality, which includes rendering adaptiveness and intelligence, optimizing bandwidth and storage of security

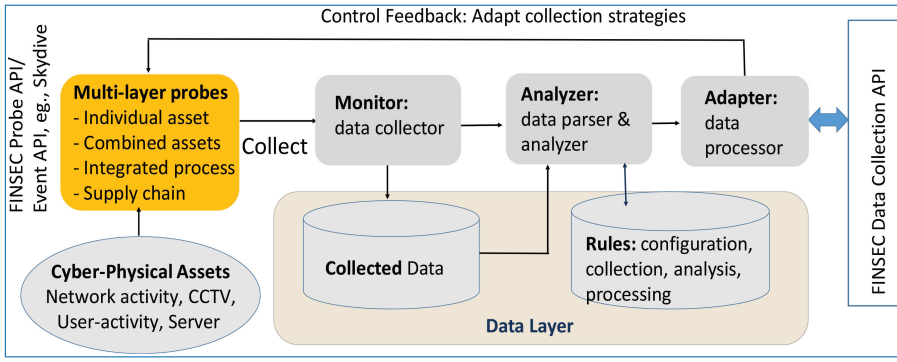

Figure 7.1. FINSEC adaptive and intelligent data collection and analytics architecture.

information, and boosting the intelligence of the probes. Security data analytics methods are integrated in the process at appropriate level-specific analytics. While predictive/regression algorithms such as linear regression, support vector machine (SVM), logistic regression, KNN regression, and Random Forest (classification & regression), K-nearest neighbors, and Decision Tree have been evaluated for the lightweight analysis of adaptive strategies with promising accuracy results of 93%–99%, deep learning mechanisms are under evaluation for the identification of complex risk and attack patterns. These will be described in a later section. A set of rules (both static and adaptive) will be defined for data processing and analysis, configuration, collection, and adaptation.

In the **Multi-layer Probes**, the FINSEC Data Collection API is called by the actual implemented probes, e.g., skydive, to collect data from cyber and physical assets at different levels (individual asset, combined assets, integrated process, and supply chain).

The **Monitor** collects the data using the FINSEC Data Collection API and stores it in the DB at the Data Layer. It analyzes and summarizes the probe data from some probe types and integrates the probes and the Data Layer. The Monitor notifies the Analyser module of collected data.

The **Analyser** module, such as anomaly detection and predictive analytics, analyzes the data and converts the standard data to service data (threats, anomalies, attacks, etc.). Further, it passes the service data to the Adapter module.

The **Adapter** module disposes the service data depending on its value such that it adapts collection strategies and controls the probes through the FINSEC Mitigation API and sends notification to external modules such as alarms and/or data visualization tool or database.

The combinations of Deep Learning algorithms and statistical approaches are utilized to deliver intelligence on anomalies and attacks with the sort of speed to maximize the value of that intelligence. This allows to (i) enhance components of

the FINSEC toolbox with more data and predictive security capabilities; (ii) train predictive models running different iterations of different algorithms; (iii) use different models on the same set of data, determine the one that best fits; (iv) establish predictive models to be used for wider use in the financial sector; (v) correlate cyber-physical events and detect cross domain anomalies through pattern detection engine; and (vi) learn typical behavior of the system and detect anomalies through machine learning engine.

The issue of false positives will be addressed to ensure reliability and accuracy. For the quality attribute performance, accuracy, and security & privacy [17, 20, 21], different measures are taken. Performance can be met through ML algorithm optimization, feature selection and extraction, data cutoff, etc. Accuracy can also be improved through alert correlation, combining signature-based and anomaly-based detection, etc. The Security and privacy of the collected and analyzed data is protected through encryption and cross-cutting security services of the FINSEC platform such as authenticity and integrity protection.

7.4.2 Implementation in the FINSEC Reference Architecture

The FINSEC Reference Architecture (RA) provides capability to foster new, intelligent, collaborative and more dynamic approaches to detect, prevent, and mitigate integrated (cyber & physical) security incidents, intelligent monitoring, and data collection of security-related information (the topic of this section); predictive analytics over the collected data; triggering of preventive and mitigation measures in advance of the occurrence of the attack; and allowing all stakeholders to collaborate in vulnerability assessment, risk analysis, threat identification, threat mitigation, and compliance.

Figure 7.2 depicts the implementation of the adaptive and intelligent data collection and analytics architecture with the process modules in the overall FINSEC RA, closing the adaptive loop of **Monitor**, **Analyse**, and **Adapt/Configure** through a feedback control loop. The **Monitor** module maps to the **Data Collection** module in the FINSEC RA, the **Analyser** module maps to the **Predictive Analytics**, **Anomaly Detection,** and **Risk Assessment** services in the FINSEC RA, and the **Adapter** module maps to the **Mitigation** service and **Mitigation Enabler** in the FINSEC RA.

Having data collected with flexible granularity on one hand and with high redundancy on the other allows the correlation of information between locations and layers and the use of various algorithms to produce insights. In this way, increased automation and optimization of bandwidth and storage of security information is achieved using the adaptive collection strategies such as security threats, content variation, collection/sampling rate, bandwidth variation/communication

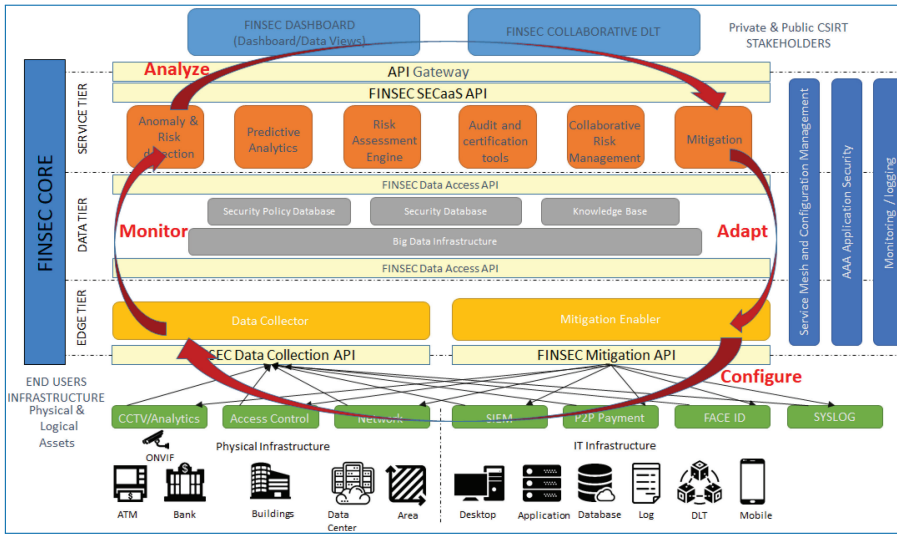

Figure 7.2. Implementation of the adaptive and intelligent data collection and analytics architecture in the FINSEC reference architecture.

dynamics, application needs, context changes, and storage needs. This automation can also be controlled through the FINSEC Dashboard user interface allowing the human in the loop.

7.4.3 Automation Through Predictive Analytics

As mentioned above, the increased automation and optimization of bandwidth and storage of security information is achieved using adaptive data collection strategies such as security threats, content variation, collection rate, bandwidth variation, communication dynamics, application needs, context changes, and storage needs. This, in turn, is achieved through predictive analytics and is achieved at different levels [16]:

- Automation of the data collection, which is inherently automatic in capturing and recording of data for later processing and analysis;
- Automation of the data pre-processing, normalization, and preparation to feed the inputs of the system;
- Automation of the analysis, training, and learning from the collected data, and the detection process;
- Automation of the mitigation process for taking mitigating actions to avoid escalation of the detected anomaly, intrusions, attacks through either passive reaction such as raising alarms or stopping of the system or active reaction such as avoiding system failure.

The FINSEC solution adds another level of automation by tying these automation levels to an overall adaptive and automation level through the feedback control loop (monitor, analyze, and adapt) increasing the automation of monitoring, analyzing, and adapting to the environmental context. This automation and adaptive nature of the FINSEC data collection and analytics allows us to meet quality attributes requirements described in [17, 20, 21] by adjusting its collection mechanism to different environmental contexts and situations, which are termed adaptive data collection strategies and will be described in Section 7.5.

7.4.4 FINSEC Multi-layer Security Probes

The FINSEC probes implemented for data collection and analytics are CCTV probe, Access Control probe, Network Skydive probe, SIEM probe, P2P Payment probe, FaceID probe, and Syslog/App Login as shown in Figure 7.2. This section describes these in brief.

7.4.4.1 CCTV probe

The CCTV probe monitors CCTV, analyzes movements, and detects physical events that may cause threats. The analytics service produces events coming from observations of physical interactions by CCTV.

7.4.4.2 Access control probe

The Access Control probe correlates cyber-physical events by checking the access to a secured area by both the use of a badge and a fingerprint and the state change signaled by movement sensors, vibration sensors, gas sensors, and temperature sensors. Data access events indicate legitimate authentication through HID (Human Interface Devices) readers and fingerprint readers.

7.4.4.3 Network Skydive probe

Skydive is an open source real-time network topology and protocols analyser. It provides real-time insights on network activity which can be used for anomaly detection. It provides agents that act as data collectors, employing efficient mechanisms to control the granularity of data collected and collection intrusiveness, which ensure minimal CPU, memory and network overheads on the monitored system. These mechanisms allow for extra flexibility in capturing network topology and network flow data, as compared to other existing tools. The challenge is to efficiently collect data with minimal disturbance to the production workloads. This includes memory and CPU but also the network itself that is shared in some level between

the monitoring and data acquisition tooling and the production workloads. In addition to the common methods, sFlow, netFlow, pcap, etc., a modern advanced networking infrastructure for Host level capturing known as bpf and eBPF is utilized. Those capturing methods make use of Linux Kernel and outperform legacy methods in a wide range of scenarios. With ebpf/bpf capturing, it is possible on one hand to limit and slice the networking data captured to some defined value, and even to change dynamically the capture to fit to on-going security demands and on the other hand allow much more efficient capturing that required significantly less CPU and Memory. All this optimization is achieved through configuring and re-configuring of the frequency of data collection based on different adaptive strategies. This is achieved using the probe configuration data model.

The Skydive probe is composed of Skydive Agents that collect topological information (the Hosts, Switches and NICs (Network Interface Controllers) in the system) and flow information (the L3 traffic streams; using powerful protocols analyzers to understand the traffic). This information is reported by the Skydive Agents to a Skydive Analyzer which aggregates the information at the cluster level and stores it in a time-series database. Figure 7.3 provides a multi-layer Skydive probe architecture.

The Skydive Analyzer exposes the real-time Flow information via a WebSocket which enables construction of Export pipelines. It processes these flows (transforming, encoding, compressing, and storing) and thus facilitates the construction of analytical tools that consume Skydive flow information.

The FINSEC Skydive Adapter (also implemented in Python) pushes network data as observed data to the data collector layer by performing the following steps:

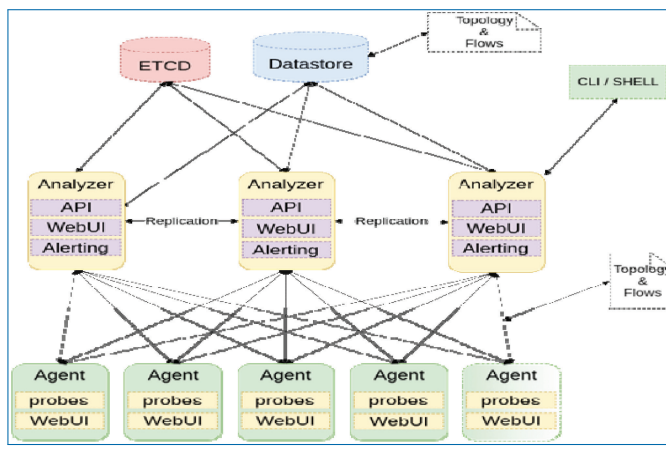

Figure 7.3. Multi-layer Skydive probe architecture.

- Classify flows according to traffic type (internal, ingress, egress, unknown)
- Reformat flows to FINSTIX (FINSEC Data Model)
- Submit flows to data-collector layer

7.4.4.4 SIEM probe

Security Information and Event Management (SIEM) systems have been used in IT since long ago to guarantee security in computer transactions and technological environments. SIEMs collect information about the monitored IT system by using agents deployed close to the infrastructure elements. This information is encapsulated in the form of events, stored and correlated to identify anomalous behaviors, discover possible threats, and detect security incidents. This way the SIEM offers a security administrator a view of the security status and of the activity that is going on in the monitored system.

In FINSEC, the SIEM probe is based on the XL-SIEM (Cross-Layer SIEM) tool developed by Atos [22], which produces alarms by correlating events received from different sources to offer extended information to other components. The event sources are typically application logs and sensors such as HIDS (Host Intrusion Detection Systems), NIDS (Network Intrusion Detection Systems), and AntiVirus.

7.4.4.5 P2P Payment probe

The P2P Payment Probe includes the following three modules that contribute the following features to the FINSEC platform: The P2P Pay module monitors and collects data of peer-to-peer payments sent on Blockchain infrastructure by end users via their commercial banks; The Block chain module monitors and collects Blockchain infrastructure parameters useful for anomaly detection on payments sent on Blockchain and Blockchain itself; and The Actuation module provides a web service interface to send specified events and commands to P2P Payment probe.

7.4.4.6 FaceID probe

The FaceID probe is two factors identification probe that combines physical level (face recognition) and credential entering to authenticate users.

7.4.4.7 Syslog/App Login

Syslog Probe analyzes the logs generated by the internal Bank monitoring infrastructure. It is installed inside the Bank premises in a virtual machine with access restrictions to users and software that can be added.

The responsibilities of the Syslog probe are to send initial information to the data collector with the FINSTIX x-assets, x-probes, x-probe-configurations, to monitor a local database which stores in near real time all the syslog events provided by the Bank's internal monitoring infrastructure, to filter and analyze records

received from the Syslog, and to generate corresponding x-event and observed-data FINSTIX objects based on a set of rules; and the events generated are related to a predefined threat providing the collaborative risk module the ability to perform risk calculations.

7.5 Adaptive Data Collection Strategies

The FINSEC data collection strategies are based on security threats, content variation, collection rate, bandwidth variation, application needs, communication dynamics, and environmental context change which all are addressed in the ensuing sections.

7.5.1 Content Variation and Security Threats

To adapt the collection rate to content variations, FINSEC will implement the adaptive sampling rate algorithm that is defined and presented in [7]. The algorithm uses a score for sets similarity, which is defined in this study. The algorithm computes the similarity between datasets collected during successive slots of monitoring. Further, the amount of the redundant data is determined based on the similarity score; thus, the size of the data sent for further processing is reduced.

To adapt the collection rate to security threats, the predictive analytics analyzes collected data and predicts security attack, threat, or anomaly. Then, predictive analytics initiates mitigation measure, in this case adaptive data collection strategy via the FINSEC mitigation service. The FINSEC mitigation service instructs the FINSEC Mitigation Enabler to adapt the collection rate. The FINSEC Mitigation Enabler instructs the Field tier probe to re-configure collection rate and the Field tier probe re-configures its collection rate and pushes data accordingly, thus adapting the rate of data collection based on the security context.

7.5.2 Anomaly Detection Driven Data Collection

Figure 7.4 shows a generic anomaly driven adaptive data acquisition approach proposed for the FINSEC platform. It is composed of three components: (1) **Mitigation rules** defined using FINSTIX and stored in the Data layer. These rules will define what events or attacks should trigger probe activations. **Mitigation service** will apply these rules to decide when and which Probe Mitigation API should be called; (2) **Probes Mitigation API** exposed by the probes to control what operations should be performed by probes for the mitigation; and (3) Analytics and probes produce event and attack **mitigation triggers** to trigger mitigation rules.

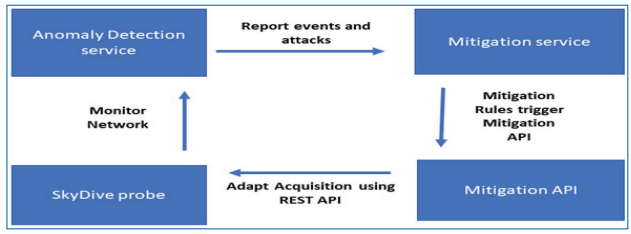

Figure 7.4. Anomaly driven adaptive data acquisition.

For Skydive probe the above components become: Mitigation rules specify which Skydive probe Actuation APIs should be called for anomalies detected on network data (Network events) or cyber-physical attacks as reported by Anomaly Detection service; Skydive's probe exposes an API to control what types of the netflows should be acquired; and Anomaly Detection service reports network anomalies and cyber-physical events to the Data Layer to trigger adaptive rules (e.g., start acquiring internal traffic).

The adaptive anomaly detection comprises Pattern Detection Engine (PDE), which correlates cyber-physical events and detects cross domain anomalies, and Machine Learning Engine (MLE), which learns typical behavior of the system and detects anomalies on Netflows. The online adaptive training updates models with the most recent observations and gradually "forgets" old behaviors. The Big data Spark-based process aggregates events over time periods and anomaly scores based on the deviation of the observed behavior from the learned models. The platform is modular that can be easily extended with new feature extractors, models, scorers, and pattern detection components.

The adaptive strategies for anomaly driven data collection include more historical data, physical measurement, change of acquisition, and outlier-driven rate of acquisition. The adaptive approach consists of adaptive rules defined using FINSTIX and stored in the Data layer; Adaptive service applies these rules to decide when and which Probe Activation API should be called, and Probes Activation API exposed by the probes to control how the data acquisition should be adapted.

7.5.3 Enhanced Security Analysis

The Atos XL-SIEM probe has been extended in FINSEC to support adaptive security data collection and this way, enhancing SIEM's security analytics capabilities. With this purpose, a new functional component has been designed, the SIEM Probe Analysis module, which is aimed to be deployed in the FINSEC platform. This module is in charge of analyzing the information received through the FINSEC Data Collector, from the SIEM Probe, and invokes the XL-SIEM Mitigation API

to take the necessary adaptive actions. Through this, the SIEM probe can reconfigure itself and the different sensors involved in the data collection, deployed at the target IT infrastructure of the organization, and thus adapts to a new cybersecurity context.

The SIEM Probe Analysis module analyzes FINSTIX data available in the FINSEC platform, together with other relevant Threat Intelligence retrieved from external sources. Two different strategies for FINSTIX data analysis are used:

- **Detection of noisy events to adapt the quantity of events received from the SIEM probe.** This is implemented by creating filtering rules in the SIEM, on-demand, to mute some specific kind of events. This improves the data collection rate in the SIEM probe by lowering down the frequency of periodic non-relevant events. Events are still collected in the SIEM but not reported to the FINSEC platform;
- **Exploitation of IoCs (indicator of compromises) to improve or extend SIEM capabilities to detect security incidents and thus enhance the quality of events received from the SIEM.** The SIEM Probe Analysis module will retrieve IoCs from external sources [e.g., OTX (Open Threat Exchange)], related to events or attacks reported to the FINSEC Platform. IoCs related to suspicious activity already detected in the FINSEC Platform contain valuable and high-quality information that, for instance, an IDS can use to improve or extend their detection capabilities.

7.5.4 Application-driven Innovative Attacks

In the context of the detection algorithms investigated, the focus is on the detection of application layer attacks. Threats like Slow DoS Attacks (SDA) [23], tunneling, and covert channels [24] belong to this category. In the anomaly based intrusion detection topic, after appropriate training on allowed scenarios, a characterization of legitimate conditions is accomplished and used for detection. Particularly, the aim of the algorithm is to monitor and analyze run-time traffic (through on-line or off-line techniques), hence flag as legitimate or anomalous the analyzed traffic.

In order to analyze a potentially anomalous situation, a capture of network traffic is needed to extrapolate predefined representative features able to characterize the considered scenario. If we consider, for instance, Slow DoS Attacks [23], such features may be related to the Delta parameters, extrapolated from network traffic and representing timings used during single connections lives [25]. By using such approach, by considering each Delta parameter, a proper threshold is defined as a consequence of the initial training [26]. The legitimate traffic is characterized to be included under the defined threshold, with a given confident interval. When

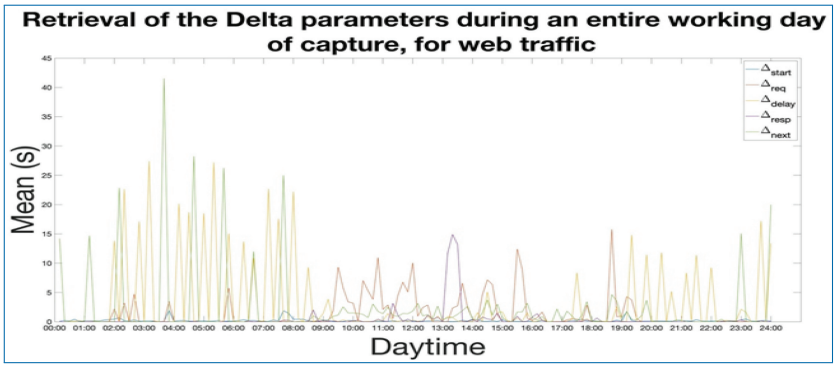

Figure 7.5. Overview of delta parameters dynamic behavior over time, for HTTP traffic between 00:00 and 24:00.

processing run-time traffic, each connection related to a Delta parameter exceeding the defined thresholds will be flagged as anomalous, hence, potentially legitimate.

Although such approach is potentially able to identify run-time threats, in particularly advanced scenarios, a malicious user may attempt to elude the detection system, by modifying the attack to make it behave like a legitimate condition. In this case, if the detection system is not able to refine the calculated Delta thresholds in real time, hence making the Delta threshold assume some sort of "dynamic" behavior, detection may fail, hence expose the system to the attack without triggering any detection.

Figure 7.5 reports the means of the Delta parameters over an entire day, from 00:00 to 24:00, when computing them on a network composed of around 50 nodes, in office environment, for HTTP network traffic.

As can be seen, their values are not static over time, but, instead, they assume some sort of "dynamic" behavior depending on the day time. Such behavior may depend, for instance, on scheduled backup activities executed overnight or on users browsing during office hours. Because of this, a first detection approach is based a dynamic adaptation of the Delta thresholds depending on the time of the day the (potentially anomalous) traffic is captured. By adapting the thresholds through such approach, it is possible to improve the detection of unknown threats, by contextualizing the detection algorithm on the time of the day considered. An extension of this approach may also monitor an entire week of traffic, to also extend the concept to non-working days like Saturday and Sunday, even though the run-time thresholds update activities.

By considering adaptive approaches, in conjunction with the approach described above, it is possible to dynamically enable and disable the network analysis process with a function of the network status. For instance, considering protection from slow DoS threats, it is possible to enable such analysis only when critical conditions

are measured. Hence, considering attacks targeting network services, it is possible to adaptively monitor traffic only when the service load exceeds a predefined threshold. This means that in case a network service is under loaded, or a partial DoS [27] is executed, protection may not be enabled, also in view of the application a green approach to cybersecurity [28]. Similarly, adaptive data collection and consequent analysis may therefore be enabled only for the features that characterize specific categories of attacks.

By considering a network platform like FINSEC, the detection algorithm may be represented as the execution of the following steps:

1. The network probe captures information from live traffic.
2. The data collector receives captured information for collection/storage.
3. The data monitor component extrapolates features from collected data.
4. The data analyzer component identifies anomalies/threats
5. The data adapter component re-configures the detection system, involving steps 3 and 4.

By adopting this approach is possible to build an adaptive detection system able to identify cyber threats.

7.6 Implementation and Validation

This section provides a prototype implementation of the adaptive and intelligent security monitoring infrastructure for the FINSEC project and its validation with the anomaly detection example.

In this first phase, a prototype implementation of the adaptive and intelligent security monitoring infrastructure is provided, which covers predictive analytics describing the most relevant approaches to analyze the collected data and detect attack patterns. In addition, the security threat and collection rate strategies are implemented. Various alternative adaptive strategies are also defined: (i) application layer adaptive collection strategies (Request start duration, Request duration, Request management duration, Response duration, and Next request start duration), (ii) adaptive techniques for data acquisition for anomaly detection (More historical data, Physical measurement, Change of acquisition, and Rate of acquisition), and (iii) Adaptive data collection for enhanced security analysis (Data Collection manager for reconfiguring the infrastructure of XL-SIEM agents, Threat intelligence update service, and Adaptive security module, which analyzes the events and alarms generated). The combination of these three architectural elements implements a feedback loop of collection, detection, and prevention that

allows for early detection of security compromises and consistently makes security analysis more effective.

7.6.1 Data Collector and Mitigation Enabler

The Data Collector (Monitor) conveys information from the probes to the Data Layer, and it may also perform additional functions for each probe. For the Skydive probe, it summarizes, at a regular interval, all "observed-data" objects seen during this interval and sends this summary to the Data Layer. The summary is created as an "x-collected-data" object, whose structure is fully described in The FINSEC Data Model (FINSTIX). It includes a list of IDs of the summarized objects, a sequence number and a time range bracketing the first and the last observed object. The Data Collector has three endpoints for the Skydive probe, supporting respectively ingress, egress, and internal traffic. Each of these traffic types is treated separately by the Data Collector, so that separate summaries are created for each traffic type, with separate sequence numbering.

In the prototype implementation, the Data Collector receives STIX objects of type "observed-data" from the probes. The Data Collector stores these objects in the Data Layer.

In the case of the Skydive probe in particular, the Data Collector also performs a summarization service of the "observed-data" objects received. Each "observed-data" object contains a set of "x-skydive-flow" objects representing native Skydive flow objects. At a regular, configurable interval (which is 10 minutes by default), the Data Collector sends an "x-collected-data" object to the Data Layer. This object contains a summary of all the "observed-data" objects received from the Skydive probe within the last interval. A separate series of 'x-collected-data' objects is created for every combination of network flow type (ingress, egress, and internal) and organization ID, and every object contains a sequence number within that series. These "x-collected-data" objects are intended to inform the analyzer that new data are available in the Data Layer.

7.6.1.1 Interface to Skydive

Skydive is a real-time network topology and protocol analyzer that can be used to capture network topology and data flows. The Skydive architecture consists of two types of software: agents and analyzers. The purpose of an agent is to collect topology and flow data various types of probes. Thus, an agent needs to be deployed on each computer to be monitored. The purpose of an analyzer is to consolidate the information collected from a set of agents. Only one analyzer is needed, although there may be more than one if redundancy is required.

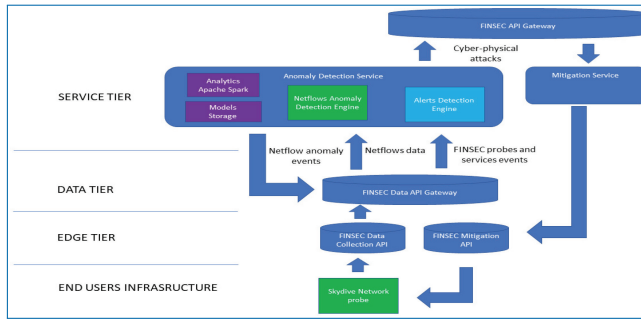

Figure 7.6. Anomaly detection with Skydive probe.

Each analyzer offers two types of interfaces for accessing its functions. The first is a graphical user interface for interactive use of management and monitoring functions. The second is an API that can be integrated with applications. This API is based on the JSON format for exchanging data and the Gremlin language for executing queries on the topology graph.

Figure 7.6 depicts the end-to-end data flow from Skydive probe to Anomaly Detection service. Here are the main steps of the dataflow:

1. The netflow collected by Skydive Network probe are pushed to FINSEC Data Layer through FINSEC Data Collection API as "observed-data" objects.
2. Data Collection service periodically produces "x-collected-data" object that references the "observed-data" objects.
3. Network Anomaly Detection Engine analyzes new "observed-data" objects and reports anomalies as to FINSEC Data Layer as "x-event" object.
4. Alerts Detection Engine correlates reported events according to "x-attack" models and report "x-attack" instances to FINSEC API Gateway
5. Mitigation Service (not implemented yes) will analyze produced "x-events" and "x-attacks" to activate adaptive Mitigation API of Skydive Network probe.

7.6.1.2 STIX and customizations

Structured Threat Information Expression (STIX^{TM}) is a JSON-based language for expressing cyber threat and observable information. A STIX [29] description consists of a set of STIX Domain Objects (SDOs) and a set of STIX Relationship Objects (SROs). The SROs describe relations between the SDOs, forming a graph. In addition to these types of objects, there are also STIX Cyber Observables, which are used by various SDOs to provide additional context to the data that they characterize.

The STIX language can be customized and remain compatible with STIX, as long as certain syntactic rules are observed. In the case of the Data Collector, two custom STIX object types, "x-skydive-flow" and "x-collected-data," were introduced. This was necessary, since Skydive delivers very detailed information on flows in its own JSON-based format, which is incompatible with STIX. Each of these flow descriptions is converted into an "x-skydive-flow" STIX object containing the same structure and the same properties as the Skydive flow object, except that the properties are converted to be compatible with STIX syntax, and "type" and "extensions" properties are added. The "x-collected-data" object type is used to summarize the aforementioned objects.

7.6.2 Predictive Analytics

The general goals of predictive analytics models are to reduce false-positive rates and to deal with a large amount of data for training and prediction, imbalanced datasets, a large number of features, and categorical and continuous features [30]. Random Forest models outperform in achieving these goals due to their advantages of low training time complexity, fast prediction, resilience to deal with imbalanced datasets, embedded feature selection method and intrinsic metrics to rank features by importance, and for their ability to deal natively with categorical and continuous features [30].

To evaluate approaches for the adaptation of the data collection strategies and intelligent processing, we have studied and tested predictive analytics based on machine learning algorithms. At this stage, the following machine learning algorithms have been selected for the predictive analytics toolkit: Support Vector Machine (SVM) using the RBF (Radial Basis Function) kernel method, K-nearest neighbors (KNN), Decision Tree using the Classification and Regression Tree (CART) algorithm, Random Forest, and Multilayer Perceptron (MLP). These algorithms are often applied to solve classification problems. We used the scikit-learn package, Python 3. The PyCharm Integrated Development Environment (IDE) was used for coding. pPckle files have been generated for each model and saved. Furthermore, we explored the possibility of using deep learning algorithms and tested a multi-layer perceptron neural network with 3 layers (on the CICIDS 2017 (Intrusion Detection Evaluation Dataset) dataset mentioned below).

The toolkit has been tested using the datasets KDDCup-99 [31], CICIDS 2017 [32], and UNSW-NB15 [33], which are described below.

The KDDCup99 is a relatively old dataset that was used for "The Third International Knowledge Discovery and Data Mining Tools Competition." The competition's task was to build a predictive network intrusion detector model capable of distinguishing between attacks and normal network traffic. This database contains

a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment. All features provided with this dataset have been applied.

The CIC IDS 2017 dataset was created by the Canadian Institute for Cybersecurity. It contains benign traffic and the most up-to-date common attacks. According to the authors, the network traffic analysis was performed using CICFlowMeter with labeled flows based on the time stamp, source, and destination IPs, source and destination ports, protocols and attack (CSV files). Generating realistic background traffic was prioritized. The authors used their B-Profile system (Sharafaldin *et al.* [34]) to profile the abstract behavior of human interactions and generate naturalistic benign background traffic. The dataset is built upon the abstract behavior of 25 users based on the HTTP, HTTPS, FTP, SSH, and email protocols. The CIC IDS 2017 dataset has over 2.83 M examples (2.27 M benign and 557,646 malicious ones) in contrast to KDDCup-99 dataset with 148,517 flows including 77,054 benign and 71,463 malicious ones. For prediction, we used all provided features.

The UNSW-NB15 dataset was created as an IoT dataset in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS). The authors aimed to generate a hybrid of real modern normal activities and synthetic contemporary attack behaviors. According to the authors, the raw network packets of the UNSW-NB 15 dataset was created by the IXIA PerfectStorm tool. This dataset has nine types of attacks: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms. The authors used the Argus and Bro-IDS tools to generate totally 49 features with the class label. All features that were included in the datasets have been used.

The test results for these three datasets are depicted in Tables 7.1–7.4, respectively. The tests were done using a laptop with Intel(R) Core(TM) i5-5300 U CPU, 2.30 GHz, RAM 16.0 GB, and 64-bit OP.

Table 7.1. Classification results for KDDCup-99 dataset.

	Training Set	Training Time	Testing Set	Testing Time	Accuracy
SVM	3428901	3721.03	1469529	648.56	0.922812634431
KNN	3428901	88.5	1469529	24.94	0.986108253728
Decision Tree	3428901	26.394	1469529	4.74	0.9983703703703
RF (100 estimators)	3428901	712.58	1469529	43.645	0.999948282784
RF (300 estimators)	3428901	2095.318	1469529	120.91274	0.99994896327
RF (500 estimators)	3428901	3424.962	1469529	229.4986	0.999949643763

Table 7.2. Classification results for CIC IDS 2017 dataset, the training set contains 3428901 records, and the testing set contains 1469529.

Algorithm	Training Time	Testing Time	Accuracy	Precision	Recall Score
SVM	2599.515625	226.453125	0.9352371389758432	0.9359399980539957	0.9319887947439143
KNN	3.28125	10.875	0.9977408304293899	0.997765750563484	0.9976316312033666
Decision Tree	2.546875	0.28125	0.9997932786013821	0.9997894281071482	0.9997894281071482
RF (100)	20.140625	0.6875	0.9998375760439431	0.9998122482419608	0.9998569235972009
RF (300)	60.59375	1.546875	0.9998375760439431	0.9998122482419608	0.9998569235972009
RF (500)	98.78125	2.296875	0.999822810229756	0.9997992315023175	0.9998398488439311
Light Gradient Boosting	69.625	1.25	0.9960132301695116	0.9964189831191492	0.9954752405759268
Logistic Regression	27.046875	0.359375	0.943860374461	0.9530305105666308	0.9359238242332494

	Training Set	Training Time	Testing Set	Testing Time	Accuracy
SVM	158021	2921.30	67724	267.26	0.935237138976
KNN	158021	14.32	67724	4.43	0.9977408
Decision Tree	158021	3.3072	67724	0.3744	0.9997637
RF (100 estimators)	158021	32.853	67724	0.9672	0.999948282784
RF (300 estimators)	158021	93.132596	67724	2.19961	0.999837576
RF (500 estimators)	158021	143.73932	67724	3.43202	0.9998228102
MLP (3 layers, 256 initial nodes, 4 epoch)	158021	41.21 (~10 sec. per epoch)	67724	3.11	0.9903431575217058 – I epoch 0.9986858425373575 – II epoch 0.9988335006792275 – III epoch 0.9990402220778454 – IV epoch

Table 7.3. Classification results for UNSW-NB15 dataset.

	Training Set	Training Time	Testing Set	Testing Time	Accuracy
SVM	82332	201.4	17534	37.2	0.720092009961
KNN	82332	14.32	17534	4.43	0.874246715967
Decision Tree	82332	0.1248	17534	0.0468	0.9474364579967
RF (100 estimators)	82332	1.1856	17534	0.234	0.958576507043
RF (300 estimators)	82332	3.2916	17534	0.5928	0.9625738272
RF (500 estimators)	82332	5.2884	17534	0.9984	0.9689476329

This preliminary study has shown that the SVM method has performed inadequately for training/testing time. It has also achieved lower accuracy for the UNSW-NB15 Dataset. We concluded that this method could be excluded from further work stage. The random forest method performs well while requiring slightly more time for training than the decision tree method; the deep learning MLP also performs well, with training time between random forest and the decision tree algorithms, and validation accuracy comparable with both, especially from the second epoch (after which the validation accuracy does not improve much).

In this preliminary study, we have used the datasets that were available online. All features supplied with these datasets have been applied. In the next stage, we plan to define how to select a feature set that produces acceptable results with predefined accuracy while reducing the volume of the collected and stored data. Further, we need to develop methods for predictive analytics that operate on real-time data collections and investigate new efficient predictive algorithms based on deep learning techniques. We, therefore, need to investigate how combining various deep learning approaches can improve the quality of the attack detection.

7.6.3 Anomaly Detector Service

7.6.3.1 Architecture overview

The Anomaly Detection service is composed of External and Internal Anomaly Detection services as depicted in Figure 7.7. The Internal Anomaly Detection service is part of the FINSEC infrastructure, and the External Anomaly Detection service is running outside of the FINSEC infrastructure on the IBM cloud. The External Anomaly Detection service is composed of two analytic engines: Network Anomaly Detection engine and Attack Detection engine.

Other related FINSEC components are the Dashboard, Data Layer, Data Collector, and the Skydive probe. Figure 7.8 shows the data flow between the

Table 7.4. Classification results for UNSW-NB15 dataset, the training set contains 490001 records, and the testing set contains 210000 records.

Algorithm	Training Time	Testing Time	Accuracy	Precision	Recall Score
SVM	2766.109375	346.53125	0.9687191965752544	0.7843639878854836	0.5002233948306303
KNN	17.9375	27.65625	0.9848238817910391	0.9268482036258789	0.802971164004187
Decision Tree	5.03125	0.515625	0.9981571516326113	0.984694679159947	0.9849088498677336
RF (100)	113.265625	3.28125	0.9986047685487212	0.9854760420933983	0.9916943506145978
RF (300)	329.59375	7.921875	0.9985809591382898	0.9849070673636564	0.9919029984302696
RF (500)	542.265625	12.828125	0.9985619116099447	0.9846162034317678	0.9918931670869641
Light Gradient Boosting	144.984375	2.296875	0.9687144346931681	0.48435721734658405	0.5
Logistic Regression	11.828125	0.578125	0.9685287212918033	0.7170305523649205	0.5185364976312337

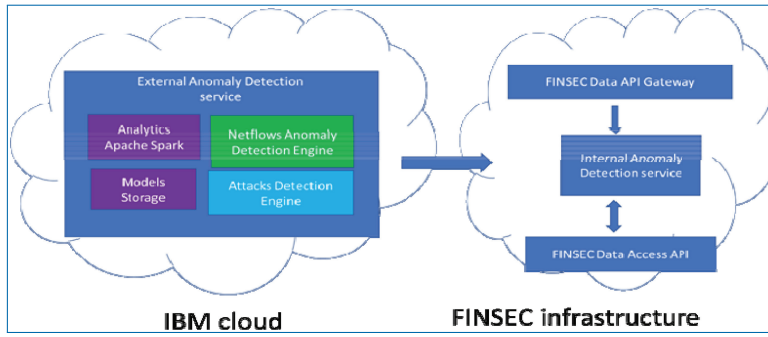

Figure 7.7. External and internal anomaly detection services.

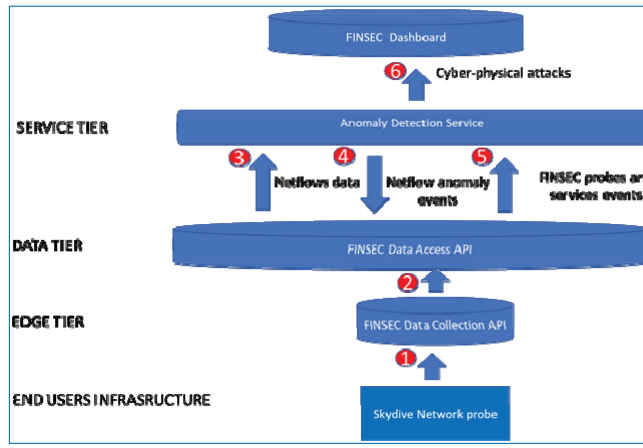

Figure 7.8. Anomaly detection data flow.

above-mentioned components and the Anomaly Detection service, starting from the probe data acquisition and culminating in attack detection and reporting to the dashboard. Described below are the main steps of the data flow:

1. The Netflow data is acquired by the Skydive Network probe and pushed into the Data Collector.
2. The Data Collector aggregates the data and pushes it to the Data Layer.
3. The Netflow data from the Data Layer is processed by the Netflow Anomaly Detection Engine of the Anomaly Detection Service.
4. The Netflow anomaly events detected in the previous step are reported to the Data Layer.
5. Netflow anomaly events along with events produced by other services are analyzed by the Attack Detection Engine.
6. The Cyber-physical attacks that are detected in the previous step are exposed to the FINSEC Dashboard.

7.6.4 SIEM Probe Analysis

As previously introduced in Section 5.3, the Atos XL-SIEM technology has been extended with a new module, the SIEM Probe Analysis module, that supports the implementation of adaptive data collection strategies with the ultimate goal of improving the quality of security events collected and controlling the data collection rate. This module, deployed as a service in the FINSEC platform, works in combination with other services and modules of the XL-SIEM probe running in the field. Figure 7.9 depicts all the elements that compose the XL-SIEM probe adaptive infrastructure and illustrates their intended deployment. The figure also shows the interaction of these elements with other services and components of the FINSEC platform, such as the Data Collector.

On the left hand side of Figure 7.9, Monitored Infrastructure is the target infrastructure under surveillance. This infrastructure is composed by different logical and physical assets such as laptops, servers, routers, printers, and the local area network. These elements are monitored by different typical security sensors or probes such as Host-based Intrusion Detection System (IDS), Network-based Intrusion Detection System (NIDS), or Antivirus (AV), all of them under the control of one or more XL-SIEM agents. XL-SIEM agents are in direct communication with the XL-SIEM probe to send security events or retrieve monitoring configuration updates. XL-SIEM Probe represents the core of the XL-SIEM technology. The Data Collection Manager module, the Data Collection Rules database, and the Configuration Update Service, which will manage the configuration of the remote monitoring components, deployed at the Monitored Infrastructure. This configuration can be updated as a result of an invocation of a specific adaptive action through the XL-SIEM Mitigation API. This API is used by the XL-SIEM to allow modifying the

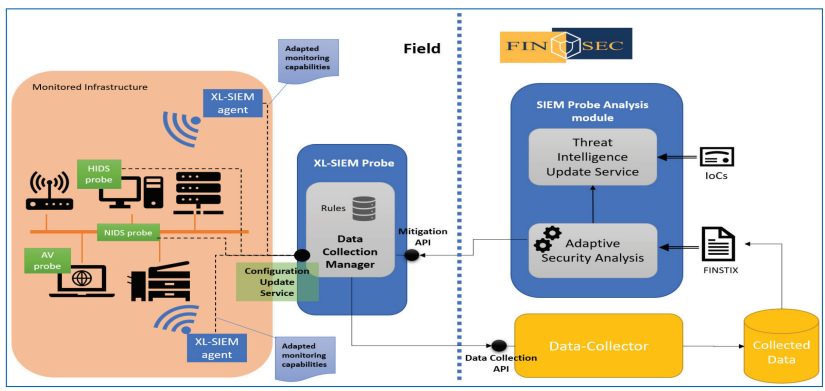

Figure 7.9. Overview of the XL-SIEM probe adaptive infrastructure.

configuration of the XL-SIEM, particularly the configuration of data collection process.

On the right hand side of Figure 7.9, the modules and services are represented which are running under the umbrella of the FINSEC platform. This is the case of the SIEM Probe Analysis module and the Data Collector with the corresponding database to store the collected data. As part of the SIEM Probe Analysis module, the Adaptive Security Analysis (ASA) service is in charge of, first, analyzing the information received in the data-collector from the SIEM Probe and, second, taking decisions on which adaptive strategy to invoke through the XL-SIEM Mitigation API. The Threat Intelligence Update Service (TIUS) supports the ASA service and is responsible for retrieving additional high-quality information about certain security events under analysis. This additional information can be obtained from another FINSEC source of security intelligence, such as the Knowledge Base, or from external sources of IoCs, e.g., Open Threat Exchange (OTX) [35].

Each FINSTIX instance received from the XL-SIEM probe at the Data-Collector is processed at the ASA to extract candidate IoCs from the list of attributes, e.g., URLs, IPs, domains, malware hashes, etc. If the IoC is a public IP, ASA uses the TIUS to consult in the OTX service and returns a list of related "pulses." The TIUS can subscribe the XL-SIEM probe to pulses in order to automatically get new relevant IoCs. The subscription is done only if it is a trusted pulse, i.e., if the number of subscriptions that this pulse already have is above a threshold. Through this process, also known as IoCs Expander, the ASA component can, for example, dynamically update the NIDS (e.g., Suricata [36]) with new rules retrieved from the official NIDS update service (Emerging Threats [37], in the case of Suricata). This way, the XL-SIEM probe adapts to collect additional relevant security events from the monitored infrastructure.

On the other hand, the decision of the ASA after the analysis of the FINSTIX collected data could be to reduce the quantity of events received from the XL-SIEM probe for various reasons, e.g., because the information about a specific IP address is considered not relevant (i.e., it is in a whitelist) or the probe can be instructed to send FINSTIX events wrapping XL-SIEM alarms (high-level correlated data) instead of XL-SIEM events (low level security information). The XL-SIEM probe can be instructed to mute a particular type of event through the invocation of the corresponding method of the Mitigation API. This results in one or more filtering rules created in the XL-SIEM probe. These rules do not prevent the XL-SIEM to generate the event and its corresponding FINSTIX instance but will not send it to the FINSEC Data Collector. This way, the muted events can be recovered upon request at a later point in time if necessary. Filtering rules can be retrieved and removed too, by using the corresponding methods of the Mitigation API.

7.6.5 Innovative Attacks

If we consider the last generation threats, it is important to consider that they may expose characteristics that make them improve their efficacy, compared to old-style threats. If we consider, for instance, the Slow DoS Attacks [23], the focus of our work, compared to old-style flooding threats, the quality of the attack is in this case enhanced, in terms of effects on the system and requirements to the attacker. This is due to the fact that during the execution of an "innovative attack" like the Slow DoS, almost all the packets composing the communication between the attacker and the victim contribute and are important for the success of the attack itself. This means that there is less waste of packets, from the attacker's perspective, compared to old-style flooding attacks, whose approach is to send a huge amount of packets to the victim to attempt to saturate its resources, in case of a slow DoS, a smarter approach is adopted. In virtue of this, reduced attack resources (CPU, memories, bandwidth, etc.) are required.

Considering innovative attacks we have investigated, it is important to consider that the Slow DoS category we have investigated is able to target application layer protocols based on TCP. Known attacks [23, 27] are found in literature for protocols like HTTP, HTTPS, or SMTP. Nevertheless, it is important to consider that the same concept can be adapted and ported to affect different protocols as well. In this case, it may be required to adapt the attack to make it able to target the considered protocol. If we consider, for instance, the MQTT protocol [38], widely used in the machine-to-machine (M2M) context, it may be required to send specific commands like CONNECT (with consequent reception of CONNECT+ACK) messages to perpetrate a long request DoS attack [23]. Preliminary tests executed [39] against a real MQTT service supporting secure communications are shown in Figure 7.10.

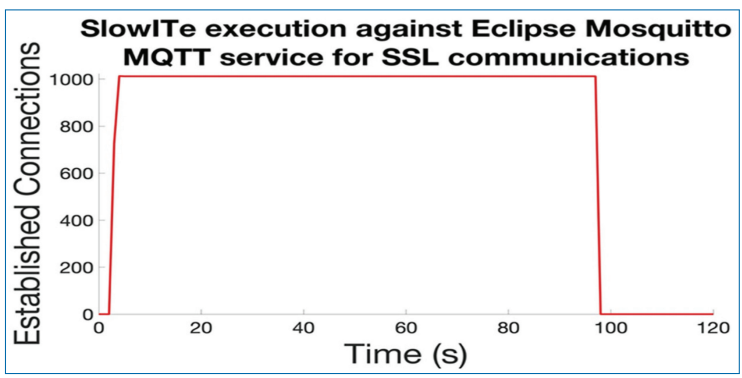

Figure 7.10. Results of tests of the SlowITe slow DoS attack against MQTT service [29].

Figure 7.10 shows that the attack is successful, even on encrypted communications, and it is able to initiate a large number of connections. Tests also reported that the denial of service is reached on the server just after the establishment of 1012 connections that are closed around 90 seconds after their establishment. Although such number of connections may appear high, in this case, since the application layer daemon is targeted, compared to the number of TCP connections, a network host is able to manage (in the order of tens of thousands), such number is considered low. In addition, it is important to consider that no application layer packets are exchanged after the establishment of a connection. Hence, required bandwidth is extremely low. Indeed, we measured that around 340 Kbps were used required for the attack.

In the cyber-security topic, it is therefore important to consider that innovative attacks may create serious damage to the network and its components. Therefore, it is extremely crucial to deploy appropriate monitoring and protection methods and, at the same time, investigate the cyber-security field to acquire knowledge on emerging threats.

Concerning detection from attacks that target specific protocols like MQTT, it is important to consider that efficient detection is still an open issue in research [39], since legitimate clients exploiting such protocols may be characterized by long times of inactivity. This can be also found on SSH protocol, for instance, where connected users may not exchange (at the application layer) any data with the server, even for hours, without experiencing any connection closure. In virtue of this, it is particularly important to investigate the topic and to adapt slow DoS detection algorithms to such kind of "silent" protocols.

7.7 Conclusions

This chapter presented the FINSEC adaptive and intelligent data collection and analytics system for securing critical financial infrastructure. Making the data collection intelligent, resilient, automated, efficient, secure, and timely is essential to economizing of resources, accessing the right information at the right time, and quickly spotting, learning from, and addressing zero-day threats. This is achieved through the configuration of configurable collection probes and the adaptation of different collection strategies. The chapter further addresses how, inter alia, (i) the nature and quality of collected data affects the efficiency and accuracy of methods of attack detection and defense, (ii) the detection capability can be improved by correlating wide-ranging data sources and predictive analytics, (iii) the rate of the data collection at the various monitoring probes is tuned by managing the appropriate levels and types of intelligence and adaptability of security monitoring, (iv) the

optimization of bandwidth and storage of security information can be achieved by rendering adaptiveness and intelligence and by integrating smart security probes and a set of adaptive strategies and rules, and (v) the increased automation is achieved through a feedback loop of collection, detection, and prevention that allows the early detection and prevention of security compromises and consistently makes security analysis more effective.

The chapter also presented the adaptive data collection strategies, implementation of the different components of the system, and validation of the predictive analytics algorithms for intelligent processing using publicly available and widely used datasets with promising results. In our future work, we plan to validate the efficiency of all components in real-life use-case scenarios of the FINSEC project.

Acknowledgments

Part of this work has been carried out in the scope of the FINSEC project (contract number 786727), which is co-funded by the European Commission in the scope of its H2020 program. The authors gratefully acknowledge the contributions of the funding agency and of all the project partners.

References

[1] Cyber Risk Outlook, Cambridge Centre for Risk Studies, May 2019, https://www.jbs.cam.ac.uk/fileadmin/user_upload/research/centres/risk/downloads/crs-cyber-risk-outlook-2019.pdf

[2] C. Wickramasinghe, D. Marino, K. Amarasinghe, and M. Manic, "Generalization of Deep Learning For Cyber-Physical System Security: A Survey" in Proc. 44th Annual Conference of the IEEE Industrial Electronics Society, IECON 2018, Washington DC, USA, Oct. 21–23, 2018. PDF, doi: 10.1109/IECON.2018.8591773

[3] H. Lin, Z. Yan, and Y. Fu, Adaptive security-related data collection with context awareness, Journal of Network and Computer Applications, Volume 126, 2019, Pages 88–103, ISSN 1084-8045, https://doi.org/10.1016/j.jnca.2018.11.002

[4] C. Habib, A. Makhoul, R. Darazi, and C. Salim, "Self-Adaptive Data Collection and Fusion for Health Monitoring Based on Body Sensor Networks," in IEEE Transactions on Industrial Informatics, vol. 12, no. 6, pp. 2342–2352, Dec. 2016. doi: 10.1109/TII.2016.2575800

[5] A. Al-Qurabat and A. Idrees (2017). Adaptive Data Collection protocol for Extending Lifetime of Periodic Sensor Networks. Qalaai Zanist Scientific Journal, 2. doi: 10.25212/lfu.qzj.2.2.11.

[6] D. Laiymani and A. Makhoul, "Adaptive data collection approach for periodic sensor networks," 2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC), Sardinia, 2013, pp. 1448–1453. doi: 10.1109/IWCMC.2013.6583769

[7] H. Harb, A. Makhoul, A. Jaber, R. Tawil, and O. Bazzi, (2016). Adaptive data collection approach based on sets similarity function for saving energy in periodic sensor networks. Int. J. Inf. Technol. Manage. 15, 4 (January 2016), 346–363. doi: https://doi.org/10.1504/IJITM.2016.079603

[8] Z. Ji, Z. Kuang and H. Ni, "A Novel Two-Dimension Adaptive Data Collection Method for Network Management," 2009 WRI International Conference on Communications and Mobile Computing, Yunnan, 2009, pp. 237–241. doi: 10.1109/CMC.2009.10

[9] X. Tang and J. Xu, "Adaptive Data Collection Strategies for Lifetime-Constrained Wireless Sensor Networks," in IEEE Transactions on Parallel and Distributed Systems, vol. 19, no. 6, pp. 721–734, June 2008. doi: 10.1109/TPDS.2008.27

[10] H.Q. Lin, Z. Yan, Y. Chen, and L.F. Zhang, "A Survey on Network Security-Related Data Collection Technologies," IEEE Access, vol. 6, issue 1, pp. 18345–8365, Dec. 2018. doi: 10.1109/ACCESS.2018.2817921 (IF: 3.224)

[11] X.Y. Jing, Z. Yan, and W. Pedrycz, "Security Data Collection and Data Analytics in the Internet: A Survey," IEEE Communications Surveys and Tutorials, 2018. doi: 10.1109/COMST.2018.2863942 (IF: 20.23)

[12] R. Erbacher (2008). Steps for Improving Data Comprehension for Digital Security and Forensics. Proceedings of the 2008 International Conference on Security and Management, SAM 2008. 318–326.

[13] D.H. Zhou, Z. Yan, Y.L. Fu, and Z. Yao, "A Survey on Network Data Collection," Journal of Network and Computer Applications, 2018. doi: 10.1016/j.jnca.2018.05.004 (IF: 3.5)

[14] G. Liu, Z. Yan, and W. Pedryczc, "Data Collection for Attack Detection and Security Measurement in Mobile Ad Hoc Networks: A Survey," Journal of Network and Computer Applications, vol. 105, pp. 105–122, March 2018. doi: https://doi.org/10.1016/j.jnca.2018.01.004 (IF: 3.500)

[15] L.M. He, Z. Yan, and M. Atiquzzaman, "LTE/LTE-A Network Security Data Collection and Analysis for Security Measurement: A Survey", IEEE Access, vol. 6, issue 1, pp. 4220–4242, 2018. doi: 10.1109/ACCESS.2018.2792534

[16] L. Cazorla, C. Alcaraz, and J. Lopez (2013). Towards Automatic Critical Infrastructure Protection through Machine Learning. In: Luiijf E., Hartel P. (eds.) Critical Information Infrastructures Security. CRITIS 2013. Lecture Notes in Computer Science, vol. 8328. Springer, Cham.

[17] F. Ullah and M.A. Babar, "An Architecture-Driven Adaptation Approach for Big Data Cyber Security Analytics," 2019 IEEE International Conference on Software Architecture (ICSA), Hamburg, Germany, 2019, pp. 41–50. doi: 10.1109/ICSA.2019.00013

[18] C. Rieger and M. Manic, On Critical Infrastructures, Their Security and Resilience – Trends and Vision, 2018, https://arxiv.org/pdf/1812.02710.pdf

[19] D. S. Berman, A. L. Buczak, J. S. Chavis and C. L. Corbett, A Survey of Deep Learning Methods for Cyber Security, Information 2019, 10, 122.

[20] F. Ullah and M.A. Babar, "QuickAdapt: Scalable Adaptation for Big Data Cyber Security Analytics," 2019 24th International Conference on Engineering of Complex Computer Systems (ICECCS), Guangzhou, China, 2019, pp. 81–86. doi: 10.1109/ICECCS.2019.00016

[21] F. Ullah and M.A. Babar (2019). "Architectural Tactics for Big Data Cybersecurity Analytic Systems: A Review," Journal of Systems and Software, vol. 151, May 2019, Pages 81–118.

[22] G. Gonzalez-Granadillo, S. Gonzalez-Zarzosa and M. Faiella, "Towards an Enhanced Security Data Analytic Platform". 15th International Conference on Security and Cryptography (SECRYPT), 2018.

[23] E. Cambiaso, G. Papaleo, G. Chiola, and M. Aiello, "Slow dos attacks: definition and categorisation," International Journal of Trust Management in Computing and Communications, vol. 1, no. 3–4, pp. 300–319, 2013.

[24] M. Aiello, M. Mongelli, E. Cambiaso, and G. Papaleo, (2016). Profiling DNS tunneling attacks with PCA and mutual information. Logic Journal of the IGPL, 24(6), 957–970.

[25] E. Cambiaso, G. Papaleo, G. Chiola, and M. Aiello, "A network traffic representation model for detecting application layer attacks," International Journal of Computing and Digital Systems, vol. 5, no. 01, 2016.

[26] M. Aiello, E. Cambiaso, S. Scaglione, and G. Papaleo, "Asimilarity based approach for application dos attacks detection," in 2013 IEEE Symposium on Computers and Communications (ISCC), pp. 000430–000435, IEEE, 2013.

[27] E. Cambiaso, G. Papaleo, G. Chiola, and M. Aiello, "Designing and modeling the slow next dos attack," in Computational Intelligence in Security for Information Systems Conference, pp. 249–259, Springer, 2015.

[28] L. Caviglione, M. Gaggero, E. Cambiaso, and M. Aiello (2017). Measuring the energy consumption of cyber security. IEEE Comm. Magazine, 55(7), 58–63.
[29] Introduction to STIX, https://oasis-open.github.io/cti-documentation/stix/intro.html
[30] P.A.A. Resende and A.C. Drummond, "A Survey of Random Forest Based Methods for Intrusion Detection Systems", ACM Computing Surveys (CSUR), vol. 51, no. 3, pp. 48, 2018
[31] KDD Cup 1999 Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[32] CICIDS 2017 (Intrusion Detection Evaluation Dataset), https://www.unb.ca/cic/datasets/ids-2017.html
[33] The UNSW-NB15 Dataset Description, https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/
[34] I. Sharafaldin, A. Gharib, A.H. Lashkari and A.A. Ghorbani, Towards a Reliable Intrusion Detection Benchmark Dataset, Journal of Software Networking, 2017, 177–200. doi: 10.13052/jsn2445-9739.2017.009
[35] [OTX] Open Threat Exchange database, https://otx.alienvault.com/
[36] Suricata website, https://suricata-ids.org/
[37] Emerging Threats rule server, https://rules.emergingthreats.net/
[38] R. Light (2017). Mosquito: server and client implementation of the MQTT protocol. Journal of Open Source Software, 2(13), 265.
[39] I. Vaccari, M. Aiello, and E. Cambiaso (2020). SlowITe, a novel denial of service attack affecting MQTT. 2nd Workshop on Attackers and Cyber-Crime Operations.