Audit Compliance and Forensics Frameworks for Complex, Large-Scale ICT Systems

João Henriques * † ‡, Filipe Caldeira * † ‡, Tiago Cruz*, Paulo Simões*
*University of Coimbra, DEI, CISUC, Coimbra, Portugal
†Informatics Department, Polytechnic of Viseu, Viseu, Portugal
‡ CISeD – Research Centre in Digital Services, Polytechnic of Viseu, Viseu, Portugal

***Abstract*—Current Information and Communication Technology (ICT) systems are rapidly increasing in scale and complexity, driven by the need to support sophisticated processes and the growing volume of heterogeneous data from interconnected services and devices. At the same time, rising concerns over cybersecurity and legal obligations demand greater attention to security, governance, and regulatory compliance. In this context, compliance management and cybersecurity forensics have become critical priorities, requiring innovative strategies and more effective tools, frameworks, and methodologies for audit and forensic analysis.**

This paper presents an integrated Forensics and Compliance Auditing framework, designed to support the identification and analysis of past security incidents and non-compliance events across both generic and domain-specific ICT systems, which stems from a PhD dissertation carried out in a mixed industrial/academic environment.

***Index Terms*—Analytics, Big Data, Cloud Computing, Compliance Auditing, Forensics, Cybersecurity.**

I. Introduction

Securing modern ICT systems is increasingly challenging due to their growing interconnectivity and distributed nature, incorporating many applications, services, and devices spread over large areas, leading to an expanding volume of data being exchanged among the system components. Although most of cybersecurity research focuses on threat prevention, detection, and mitigation, areas such as forensic support and compliance auditing often receive less attention. These capabilities provide the foundation to develop audit compliance processes As a result, current tools Forensics and Compliance Auditing (FCA) struggle to deal with modern ICT systems, calling for the development of FCA frameworks geared towards the specific needs of such large-scale, dynamic systems, taking into account the following requirements:

- To support the implementation of best practices for incident response, which include a "lessons learned" that relies on post-incident trace analysis, enabling experts to trace issues back to their source and gather meaningful evidence. This falls within the realm of forensics processes.
- To take advantage of existing knowledge to support establishing appropriate procedures and guidelines for the secure, reliable operation, management, and maintenance of ICT systems. These capabilities provide the foundation to develop audit compliance processes.

These requirements highlight the opportunity to integrate FCA approaches, consolidating diverse techniques and tools into a unified platform able to streamline the investigation process, reduce complexity and costs, and improve the ability to connect individual alerts to identify potential threats.

In this context, our work [1] set out to research and integrate FCA capabilities into a unified security framework applicable to both generic and domain-specific ICT systems, led by following research questions:

- How can forensic approaches for ICT-based scale with heterogeneous data?
- How can compliance auditing better protect ICT-based systems?
- Which analytical models best support anomaly detection, incident replay, and evidence collection in ICT-based systems?

The result was the design of an architecture for consolidated FCA operations [2], [3], developed with scalability, advanced inference analysis, and improved security policy lifecycle management in mind. Its main distinguishing features include:

- Mechanisms for anomaly detection using large log datasets [4].
- A federated approach for inference analysis using ontology data [5][6].
- An automated closed-loop framework for enforcing security policies based on anomaly detection [7].

II. A Framework for Forensics and Compliance Auditing

Figure 1 presents the proposed FCA framework, that complements typical security systems by offering compliance auditing and forensics tools. It facilitates forensic processes such as identifying, extracting, preserving, and analyzing digital evidence. Additionally, it supports compliance audits to assess adherence to relevant standards, business rules, and policies. The framework was designed to scale horizontally, dynamically adjusting to different deployment scenarios and workloads while maintaining the ability to collect, store, and correlate large volumes of data from diverse, geographically distributed sources.

For domain-specifc scenarios the FCA may work in a hybrid fashion: domain-specific anomalies are handled by corresponding domain-specific Intrusion Detection System (IDS), while

978-3-903176-75-1 ©2025 IFIP

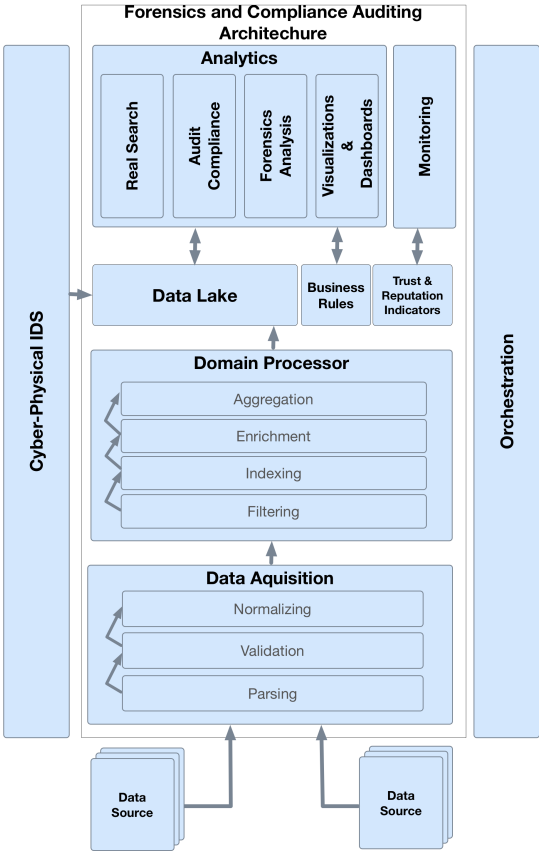

Fig. 1. Architecture of the Proposed FCA Framework [3]

more generic anomalies are handled by generic-purpose IDS. For instance, a Proof of Concept (PoC) implementation of this framework was deployed in the scope of the ATENA project [8] and integrated into the infrastructure of an electric utility, with cyber-physical system anomalies being handled by specialized IDS, while for Operational Technology (OT) systems above IEC62443 [9] layer 2 (e.g. components such as Supervisory Acquisition and Data Control (SCADA) servers or historian databases) and ICT systems the FCA is able to receive direct event feeds (from probes and log sources), to be normalised and preprocessed later. Anomaly detection in the scope of the FCA framework takes advantage of the post-processed CPS IDS data and the OT and IT feeds.

Besides functional validation, this PoC deployment was used to evaluate the performance of the framework [3], measuring throughput, scalability and latency during data ingestion and computing operations. Scalability and ingestion performance validation demonstrated the suitability of the proposed architecture to deal with considerable amounts of data (with some tests using batches of 10 million events), when adopting adequately dimensioned cluster setups, with net benefits in terms of throughput and computing latency.

A first use case showcased the potential of the proposed FCA framework, demonstrating how detection processes can be enhanced by recognizing different attack stages by analyzing data from various sources. Specifically, this is achieved by classifying heterogeneous data from multiple sources to identify multistage attacks and the connections between them. In a second use case, we emulated typical attack stages by injecting corresponding logs into the endpoint OS to determine if such attacks could be detected.

III. Anomaly Detection for Large Log Datasets

This section presents a parallel computing approach for identifying anomaly events in massive log files, to be applied in out-of-core datasets where log sources are so massive that it becomes impossible to use more traditional approaches.

To address this, we devised an integrated detection method (see Algorithm 1) that uses filters to improve accuracy and efficiency in large log data scenarios, supported by k-means and XGBoost [10] classification.

Algorithm 1 Unsupervised Learning Model [4]

INPUT: D_S, Data

$clusters \leftarrow 2$
$K \leftarrow \text{KMEANS}(clusters)$
$Y \leftarrow K.\text{TRAIN}(D_S)$
$X \leftarrow \text{XGBOOST}(D_S, Y)$
$X.\text{TRAIN}()$
$ypred \leftarrow X.\text{PREDICT}(D_S)$
$R_1, R_2 \leftarrow X.\text{DECISIONTREES}()$
for all $i \in ypred$ **do**
  if $ypred_i > 0.5$ **then**
    $ypred_i^1 \leftarrow 1$
    $k2 \leftarrow k2 + 1$
  else
    $ypred_i^1 \leftarrow 0$
    $k1 \leftarrow k1 + 1$
  end if
end for
if $k1 > k2$ **then**
  $R_a \leftarrow R_2$
  $R_n \leftarrow R_1$
else
  $R_a \leftarrow R_1$
  $R_n \leftarrow R_2$
end if

OUTPUT: R_a, Anomaly Decision Trees
OUTPUT: R_n, Non-Anomaly Decision Trees

For evaluation purposes, we used the NASA datasets [11] to extract features and clustering log events into normal and anomalous groups, with sparse clusters indicating potentially anomalous activities. Events identified in the smaller, sparser cluster are flagged for further forensic and compliance analysis. The clustered data provided labeled input for training a gradient tree boosting model using the XGBoost algorithm, creating rules to generalize anomaly classification for new events in a distributed computing environment. Together with Dask [12], k-means and XGBoost form a robust toolset for building scalable clustering and classification solutions to identify events warranting forensic and compliance analysis.

The obtained results [4] highlight the obviously normal events in highly coherent clusters, with a minor subset of events being classified as anomalies for further analysis. The interpretability of the model is indirectly validated by the produced decision rule set, which implicitly shows how the model identifies classes.

IV. A Federated Ontology for Handling Heterogeneous Data Sources

This section proposes a federated ontology-based approach to manage and integrate heterogeneous data sources within an existing Security Information and Event Management (SIEM) framework [6]. This integration is often costly and complex, requiring significant adaptation efforts and frequent maintenance, as corporate systems evolve. Moreover, Most organizational databases are not natively formatted for the Semantic Web, making integration difficult. Relational databases (Relational Databases (RDBs)), for example, do not store data in Semantic Web-friendly formats such as Resource Description Framework (RDF).

The proposed approach makes organizational data, typically stored in RDBs, accessible through simplified interfaces. These interfaces aggregate data queried from multiple and heterogeneous sources within an organization. Each RDB instance contains specific datasets structured in unique schemas and technologies. Our ontology-based approach enables seamless integration of data from RDBs and Semantic Web sources (stored in RDF), using an interface layer to handle retrieval from those repositories. This framework uses a federated layer, mapping brokers and databases. Data consumers submit SPARQL queries to the federated interface layer, which forwards them to brokers. Brokers transform SPARQL queries into native queries for each underlying RDB. Data retrieved from each database is compiled into a single result set at the federated layer before being returned to the client.

In our PoC, a simplified ontology was developed for FCA processes, capturing norms, policies, and regulatory requirements. The ontology enables organizations to infer new insights, such as unauthorized access patterns or access conflicts. A federated query web service was used to assess whether employees have the necessary permissions to access specific assets. The service's intermediary layer translates user queries into database-specific queries, facilitating consistent and comprehensive access evaluations across heterogeneous data sources. This allows to gather and combine – without requiring the end user to be aware of low-level details – information dispersed across tables and databases built using different schema.

For evaluation purposes, a practical compliance audit scenario was implemented to demonstrate the ability of the proposed approach to handle access rights assessment across the resources of an international company. This scenario, described in [6], highlights the challenge of integrating data on human and asset resources managed under different national regulations.

V. Automated Closed-Loop Enforcement of Security Policies from Anomaly Detection

Due to the growing complexity and scale of ICT systems, there is an increasing need to automate and streamline routine maintenance and security management procedures to reduce costs and improve productivity. As a result, approaches such as ETSI Zero-touch Network & Service Management (ZSM) [13] are becoming increasingly popular. Such approaches enable greater consistency and uniformity and significantly improve operations and maintenance activities. Moreover, they lead to cost savings and a significant reduction in human errors.

Once security issues are found, the design, implementation and application of specific security policies require significant efforts from operators and developers. They need to design the policies and translate them to rules, code or other artifacts. This burden further increases in large organizations. Also, the verification of policies by humans is time-consuming. This is aggravated by the fact that rules may not exist a priori, being created and evolved as data becomes available.

Frequently, policies are enforced by directly embedding them in the the source code. Many existing policies or Access Control Lists (ACLs) are set by the use of options in user interfaces, which is not an easily repeatable or versionable task. This is inefficient, makes it difficult to keep up-to-date inventories, and hampers automated testing. Moreover, ACLs usually lack support for auditing or checking policy violations.

To address these issues, we proposed a continuous automated closed-loop process based on three stages: first, to extract the Decision Tree (DT) from Machine Learning (ML) models to identify the anomalies; next, translating them to policies; finally, enforcing them along with the different system components. This continuous closed-loop makes it possible update policies along time with the most recent data. The ML model produces DTs that identify the anomalies to be translated to PaC in a language recognized by the PE.

In this framework, described in detail in [7], policies specify the conditions under which particular activities should be allowed to enable logic-based enforcement decisions. The proposed approach uses a continuous closed-loop model S^n which applies a three-stage S_1^n, S_2^n, S_3^n loop over n iterations, as illustrated in Figure 2. The first stage (S_1) automatically considers new incoming data to classify security anomalies. A DT model fits the data to classify the anomalies. At the second stage (S_2), the previously generated DTs are translated into Policy as Code (PaC) rules in a format recognized by the Policy Engine (PE). These rules bring together conditional logic and granular controls. Finally, at the third stage (S_3), the produced PaC is enforced by PE. The next cycle may be triggered periodically or based on specific events, which, by their nature, might require rule adjustments.

Figure 3 presents the measured accuracy in each of these steps. For the baseline system, the accuracy remained stable (with slight natural fluctuations). With our approach the system kept improving over time, since the data from the previous period was used to further refine the models.

VI. Conclusions

After surveying the field [2], we started by presenting an architecture [3] that converges FCA activities. First, we combined K-Means and XGBoost models for anomaly detection over large datasets [4]. Next, we applied ontology-based federations to provide Semantic Web inference capabilities over data stored in heterogeneous sources [5][6]. Third, we

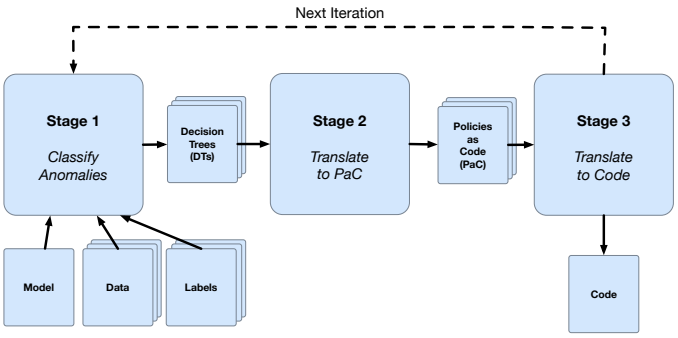

Fig. 2. Proposed Closed-Loop Approach [7]

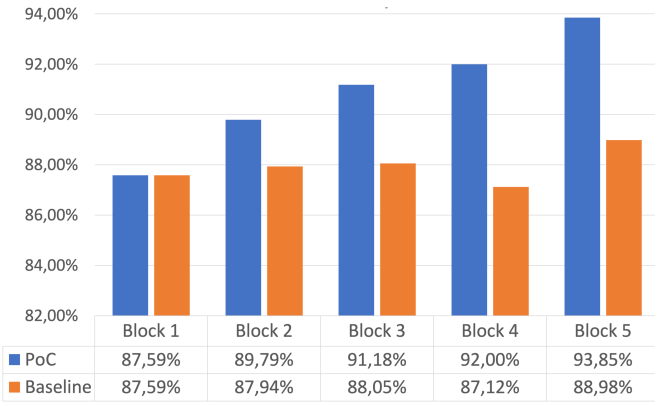

Fig. 3. Measured accuracy over time for PoC and baseline systems [7]

created an automated closed-loop to enforce security policies from anomaly detection models [7].

Areas of improvement of the FCA framework include the exploration of Large Language Models [14] and the evaluation of adaptive automation approaches to continuously update compliance and auditing policies in cloud-native environments.

Reference Material

The work supporting the presented FCA framework, which was undertaken in the scope of a PhD thesis [1], led to several refereed publications [2][3][4][5][6][7], also contributing to three co-authored papers [15][16][8].

Acknowledgements

This work was partially funded by the FOCUS-PA project (2024.07695.IACDC/2024), through the FCT—Foundation for Science and Technology under call FCT PRR RE-C05-i08. Furthermore, we thank the Research Center in Digital Services (CISeD) and the Polytechnic of Viseu for their support.

References

[1] J. Henriques, "Audit compliance and forensics frameworks for improved critical infrastructure protection," Ph.D. dissertation, Universidade de Coimbra, 2024.

[2] J. Henriques, F. Caldeira, T. Cruz, and P. Simões, "A survey on forensics and compliance auditing for critical infrastructure protection," *International Journal of Critical Infrastructure Protection*, 2022.

[3] ——, "A forensics and compliance auditing framework for critical infrastructure protection," *International Journal of Critical Infrastructure Protection*, vol. 42, p. 100613, 2023.

[4] ——, "Combining K-Means and XGBoost models for anomaly detection using log datasets," *Electronics*, vol. 9, no. 7, 2020, doi:10.3390/electronics9071164.

[5] ——, "On the use of ontology data for protecting critical infrastructures," in *Proceedings of the 17th European Conference on Cyber Warfare and Security*, 2018, pp. 208–216.

[6] ——, "On the use of ontology data for protecting critical infrastructures," *Journal of Information Warfare*, vol. 17, no. 4, pp. 38–55, 2018.

[7] ——, "An automated closed-loop framework to enforce security policies from anomaly detection," *Computers & Security*, p. 102949, 2022.

[8] L. Rosa, T. Cruz, M. B. de Freitas, P. Quitério, J. Henriques, F. Caldeira, E. Monteiro, and P. Simões, "Intrusion and anomaly detection for the next-generation of industrial automation and control systems," *Future Generation Computer Systems*, vol. 119, pp. 50–67, 2021.

[9] IEC, "IEC 62443 - IEC Tech Specification - Industrial communication networks - Network and system security - Part 1-1: Terminology, concepts and models," 2017.

[10] T. Chen and C. Guestrin, "XGboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[11] NASA, "NASA HTTP," http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html, 2018.

[12] M. Rocklin *et al.*, "Dask: Parallel computation with blocked algorithms and task scheduling," in *Proc of the 14th python in science conference*, 2015, pp. 126–132.

[13] ETSI, "Zero-touch network and service management (ZSM); reference architecture," ETSI, Tech. Rep., 2019.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[15] L. Rosa, M. B. de Freitas, J. Henriques, P. Quitério, F. Caldeira, T. Cruz, and P. Simões, "Evolving the security paradigm for industrial IoT environments," in *Cyber Security of Industrial Control Systems in the Future Internet Environment*. IGI Global, 2020, pp. 69–90.

[16] L. Rosa, J. Proença, J. Henriques, V. Graveto, T. Cruz, P. Simões, F. Caldeira, and E. Monteiro, "An evolved security architecture for distributed industrial automation and control systems," in *European Conference on Cyber Warfare and Security*. Academic Conferences International Limited, 2017, pp. 380–390.