Forensic Science International: Reports 2 (2020) 100122

Contents lists available at ScienceDirect

Forensic Science International: Reports

journal homepage: www.elsevier.com/locate/fsir

Quantifying the need for supervised machine learning in conducting live forensic analysis of emergent configurations (ECO) in IoT environments

Victor R. Kebande [a,*], Richard A. Ikuesan [b], Nickson M. Karie [c], Sadi Alawadi [a], Kim-Kwang Raymond Choo [d], Arafat Al-Dhaqm [e]

[a] Department of Computer Science, Malmö University, Sweden
[b] Cyber and Network Security Department, Science and Technology Division, Community College of Qatar, Qatar
[c] School of Science, Edith Cowan University, Australia
[d] Department of Information Systems and Cyber Security, University of Texas at San Antonio, San Antonio, TX 78249-0631, USA
[e] School of Computing, Faculty of Engineering, Universiti Teknologi Malysia, Johor, Malaysia

ARTICLE INFO

Keywords:
Supervised machine
Learning
Live forensics
Emergent configurations
IoT

ABSTRACT

Machine learning has been shown as a promising approach to mine larger datasets, such as those that comprise data from a broad range of Internet of Things devices, across complex environment(s) to solve different problems. This paper surveys existing literature on the potential of using supervised classical machine learning techniques, such as K-Nearest Neigbour, Support Vector Machines, Naive Bayes and Random Forest algorithms, in performing live digital forensics for different IoT configurations. There are also a number of challenges associated with the use of machine learning techniques, as discussed in this paper.

1. Introduction

As Internet of Things (IoT) devices become the norm, so does the need for IoT forensics. The latter is a branch of digital forensics, which involves the investigation of IoT devices as well as the supporting infrastructure. Unlike conventional digital forensics, collecting or acquiring evidence from IoT devices can be challenging due to the diversity of IoT devices and the underpinning operating and file systems.

It is also noted that in an IoT system, especially in the case of emergent configurations (ECOs), data can be dynamic and consequently challenging to label datasets during live forensics. Live forensics in this context refers to a forensic investigation conducted in near real-time. ECOs, as defined by existing studies [1–4], are systems formed by a set of things, with their services, functionalities, and applications, that cooperate temporarily to achieve some user goals. ECOs adapt in response to (unforeseen) contextual changes, such as changes in available things or user goals. Given the heterogeneity and increased connectivity of emerging configurations, ECOs platforms can be challenging to perform live forensics, given that such systems may comprise one or more dynamic and heterogeneous (IoT) systems, which may also be distributed [5].

In recent times, there have been attempts to utilize machine learning (ML) techniques to facilitate digital forensics, including IoT forensics. However, this inclusion has largely been within the scope of static IoT platforms such as Smart Homes where the 'context of things' are largely unchanged. Hence, in this manuscript, the authors survey existing literature on the use of supervised ML techniques (e.g., K-Nearest Neigbour, Support Vector Machines (SVM), Naive Bayes and Random Forest) in conducting live forensics across dynamic and context-changing IoT systems, typical of ECOs. At the time of our study, this is the first study to explore the feasibility of integrating ML into an ECO platform to facilitate digital forensics. Therefore, the contributions of this paper are as follows:

- explore the feasibility of integrating supervised ML techniques to perform live forensic analysis in a dynamic (ECO) IoT platform;
- demonstrate how forensic activities could dynamically be conducted in an ECO environment; and
- provide a contextual evaluation that shows that the forensic challenges in an IoT environment and how automation for incident identification may occur.

In Section 2, a review of the related literature and the research gap from existing studies are presented. Then, in Sections III and IV, we present our proposed conceptual framework and how it can be deployed. Discussions and conclusion are presented in the last two sections of this manuscript.

* Corresponding author.
E-mail addresses: victor.kebande@mau.se (V.R. Kebande), richard.ikuesan@ccq.edu.qa (R.A. Ikuesan), n.karie@ecu.edu.au (N.M. Karie), sadi.alawadi@mau.se (S. Alawadi), raymond.choo@fulbrightmail.org (K.-K.R. Choo), mrarafat@utm.my (A. Al-Dhaqm).

http://doi.org/10.1016/j.fsir.2020.100122
Received 21 April 2020; Received in revised form 12 June 2020; Accepted 8 July 2020
Available online 15 July 2020

2. Related literature

2.1. Existing literature

Machine Learning (and deep learning) approaches have gained renewed interest in recent years, such as the example approaches presented in Table 1.

There have also been attempts to utilize ML and deep learning for digital (forensic) investigations. For example, a generic framework that allows the application of deep learning cognitive computing techniques in cyber forensics (CF) was presented in [16]. However, this framework is not designed to facilitate live forensics in IoT environment.

Another research by [17] explored the effectiveness of employing machine learning methodologies for computer forensic analysis by tracing past file system activities and preparing a timeline to facilitate the identification of incriminating evidence. Their approach is, however, not designed to facilitate IoT forensics. Costantini et al. [18] explored the applicability of artificial intelligence (AI) along with computational logic tools to automate evidence analysis, while Mitchell [19] discussed the potential usefulness of AI in digital forensics.

However, we observe that the potential role of supervised ML techniques in live forensics on ECOs across IoT ecosystems is not well-understood or fully explored in the literature. The concept of ECO is not relatively new, and has been extensively studied [1–4]. Generally, ECOs are formed by a set of things, with their services, functionalities, and applications, which cooperate on an ad-hoc basis to achieve some user goal [2,4]. ECOs are adapted in response to (unforeseen) context changes, such as changes in the resources available or changing/evolving user goals. Given the heterogeneity and increased connectivity of emerging configurations, it can be challenging to identify malicious activities in ECOs.

The connection between IoT and ECOs can be broadly explained by the widespread adoption of IoT in different sectors (e.g., smart health, smart transport, smart cities, automation, agriculture, and manufacturing). IoT is also regarded as a disruptive technology, including by the US National intelligence council [20,21]. Therefore, in the context of IoT ECOs, we need to consider emergent behavior, connectivity (exchange of information), localization and tracking, how distributed components are, ubiquity and device heterogeneity. Implications for forensic investigation are existence of interaction, coordination and interoperability, which mainly encompasses events, context, environment and actions [22].

2.2. Research gaps

Based on a review of the existing literature, we identify the following research gaps.

- The shift in conventional digital forensics to cloud forensics, network forensics, device-level forensics and live forensics across the IoT ecosystems has compounded the challenges in performing digital investigations, for example in terms of data size and the rapidly changing technological landscape [23–26]. Hence, there is a need to ensure that digital forensic capabilities keep pace with emerging technologies [27], as well as designing AI-based approaches to facilitate digital forensics and real-time incident detection and incident response for ECOs [28,29]. This necessitates the understanding of the composition of ECOs, for example in terms of process and architecture [30].
- Conventional labeled datasets and extracted features may not necessarily be useful to facilitate live forensic across emerging IoT configurations, due to the dynamic nature of the system interactions and threat landscape [31,32].

3. Proposed framework for adopting supervised machine learning approaches

We will now present our proposed conceptual framework, as shown in Fig. 1. The three key building blocks are discussed next.

3.1. Emerging IoT configurations

ECOs can be broadly defined to be a dynamic collection of 'things' with functionalities seeking to achieve a given goal [1], and a concrete

Table 1
Snapshot of existing ML approaches in security incidents.

Reference	Objective	Machine learning approaches	Algorithm used	Application
[6]	Bot detection using unsupervised learning	(Unsupervised Machine Learning)	Flow clustering, and simple K-means clustering	Based on the flows generated by bots based on the destination port number, largest size of packet, smallest size of packet, the time the packet is flagged.
[7]	Digital forensic text string searching	(Unsupervised Machine Learning)	Clustering digital forensic text string..	Uses Self-organizing maps to tests the feasibility and utility of post-retrieval clustering of digital forensic textstring search results
[8]	Classification model for anomaly-based intrusion detection	(supervised Machine Learning)	Naïve Bayes classification, K-nearest Neigbor	Used NSL-KDD dataset to detect User to Root (U2R) and remote to Local (R2L).
[9]	forensics data task for multi-class classification	-(Supervised and neural networks)	Decision trees, Bayes classifiers, ANN and Nearest neighbor	Classifiers have been evaluated based performance measures and Cohen's kappa. A statistical analysis has been conducted in order to compute each of algorithms based on accuracy
[10]	Digital forensic readiness	Supervised Learning Approach	Bayes, Neural Net, SVM, C4.5, HMM, Nearest Neighbor,Logistic Model tree	Implemented C4.5 decision tree on Keystroke dataset for live user identification
[11]	User Identification	Supervised Learning Approach	Rule based machine learning, Decision Tree classifier	Used labeled data to perform user identification
[12]	Network forensic analysis.	Feature engineering at Analysis layer.	Analysed KDD Cup99 Dataset by applying a reputation value in data analysis method	The author used KDD Cup'99 collection of 9 week TCPdump datasets which has shown real time performance of the network based on the reputation value
[13]	Passive audio bootleg detector	(Deep learning and supervised)	Deep learning, Deep Belief Network (DBN), classification-SVM.	Implemented three class SVM and applied feature learning to detect whether music audio track relates to unauthorized recording
[14]	Intelligent Self-learning system for home automation in IoT	(Guided Learning classification)	Naive Bayes Algorithm	Automatic fault detection in connected devices
[15]	SVM-based malware detection for IoT services	(Guided Learning classification)	Linear SVM	Detecting malware that targets android-based platforms

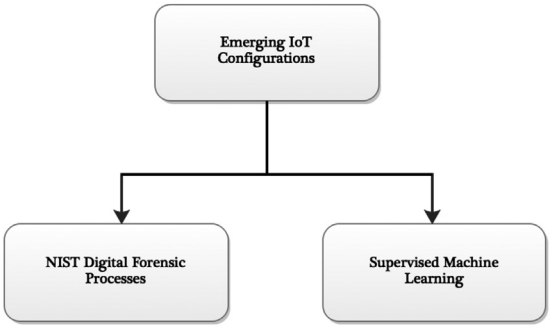

Fig. 1. High-level overview of the approach.

implementation scenario will be presented in Section 4. The ECOs are designed to achieve their goals over heterogeneous environment, and facilitate real time interactions of scenarios through successful executions while also ensuring interoperability. In IoT settings, such interactions normally require a number of actions to be executed, which implies massive amount of data that can be exploited by cyber attackers (e.g., as the proverbial phrase, 'needle in a haystack'). Hence, an in-depth understanding of the configurations and the potential data types and sources will significantly reduce the amount of time required in forensic investigations.

3.2. NIST digital forensic process

While there are a number of existing digital forensic process, we use NIST Special Publication 800-86 as the guiding process due to its widespread adoption and that it allows the integration of forensic techniques into incident response. Similar to other digital forensic domains, IoT forensics may cross jurisdictions and hence involve different laws and requirements, for example in terms of evidence collection and admissibility. As IoT systems may be deployed in critical infrastructure sectors, where taking it offline for forensic investigations is impractical, we posit the importance of live forensic-readiness too. The role of each of these processes is outlined below in the context of IoT.

- Collection: Timely identification of potential evidence sources in (interconnected) IoT ecosystems is crucial, particularly to live forensics. However, it can be challenging to do so manually due to of the dynamic nature of data interactions in IoT systems. Hence, we could explore using ML techniques, such as classification algorithms (e.g., Naive Bayes Classifier, Nearest Neighbor, and Support Vector Machines) to automate the collection process. Care should, however, be taken to ensure that one strikes a balance between false-negative and false-positive.
- Examination: This process may include pre-processing of digital data collected from emerging configuration devices/applications, the selection of suitable tools (e.g., encryption algorithm and hashing algorithm to be used), and the selection of appropriate techniques (e.g., logistic regression, to statistically analyze data collected in the previous process, and identify any information useful to the investigation such as existing relationships between objects of interest as well as variables).
- Analysis: Successful completion of examination will help us to make an informed decision on the tools and approaches to be adopted. For example, should we use K Nearest Neighbors or Decision Trees? Using the geometric distance, it may be possible to use the k-nearest-neighbors to decide which is the nearest object in the ecosystem. On the other hand, decision trees may be used to break down any collected dataset into smaller subsets while at the same time incrementally developing an associated decision tree. Then, live forensics and/or in-depth analysis of the data will be undertaken.
- Reporting: Findings from the analysis process will then be included in the report, which should also include the tools, techniques and approaches used, their rationale and the limitations (if any). For example, by using classification algorithms such as Neural Network, what is the limitation? Will any data be missed out during live forensics due to the use of such classification algorithms?

3.3. Supervised machine learning approaches

One of the benefits of using supervised ML approaches in live forensics is the potential for such techniques to give a prediction on possible events based on past occurrences. We will now discuss a few potential supervised ML algorithms that can be used in this context: Support Vector Machines (SVM), k-Nearest Neighbors (kNN), Naive Bayes and Random Forests.

- kNN: kNN can facilitate the identification of existing relationships based on the forensically acquired digital data. Specifically, due to its non-parametric learning techniques, it can be used to classify samples from a dataset on the principle of similarity. Generally, kNN's output primarily depends on the instances that emanate or are stored in the memory. Also, a majority of the kNN neighbors are tasked with giving a decision on the continuous variables that are used [33]. The KNN adopts three distinct distance metrics, namely: euclidean; Manhattan and Minkowski distance functions. The algorithm in this context adopts a K to be equal to the square root of the tuple numbers and then the distance that exists between the samples is calculated. After this, it is sorted in ascending order and thereafter, the nearest neighbor are easily selected. The distance metric is represented as follows:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{1}$$

which is the Euclidean distance function from the nearest existing points

$$\sum_{i=1}^{k}|x_i - y_i| \tag{2}$$

which is the Manhattan distance function

$$[\sum_{i=1}^{k}(|x_i - y_i|)^q]^1/q \tag{3}$$

which is the Minkowski distance function
- Support Vector Machine (SVM): A support Vector Machine (SVM) is able to learn by way of assigning labels to objects. Basically SVM which is based on statistical learning can easily be applied in forensic analysis of the collected digital data because SVM is able to generate a hyper-plane that has a capability of maximizing a margin that exists between classes [34]. SVM adopts a technique that allows a whole training set to be considered as the main root node of a given tree [35], which thereafter may be split to various sub nodes based on the existing useful information. It is represented as follows:
A training set may S may be represented as:

$$S = [(a_1 - b_1), (a_2 - b_2), \ldots, (a_n - b_n)] \tag{4}$$

a hyper-plane for the training set is represented as $F(x) = 0$ where $a_i \in R, b_i \in (-1, 1)$ then the sum attempts to find the weight vector and the bias [36,37]. This makes it more suitable to categorize different aspects and dimensions of data that is collected for purposes of forensic analysis.
- Naive Bayes Algorithm: The Naive Bayes which is also a classification algorithm could be employed to predict the probabilistic of occurrence of events from a given class. The authors still emphasize on the fact that Naive Bayes technique is independent and do not need

to depend on other existing attributes. Generally, Naive Bayes classification is based on extracting the standard deviation and the mean during classification [38,39]. Furthermore, it allows input data to be grouped based on the training and tests data. This allows Naive Bayes to work on isolated data with outlier characteristics while facing irrelevant attributes as shown in equation 5.

$$g(x,\mu,a)=\frac{1}{\sqrt{2m}}e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad (5)$$

- Random Forests: The nature of voluminous data that is acquired from connected environments has a significance of adopting random forest classifier as a supervised learning technique in conducting live forensic analysis. Basically random forest allows a single classifier to be able to provide a machine learning model that is aimed at achieving different reasons like parameterization and over-fitting. This is based on ensemble decision trees, where each tree is tested independently [40]. This allows a dataset to be split into random samples. For example, given a class c, the random forest can be used in the estimation of the probability that predicts c for a sample as follows:

$$P(c|X)\sum_{i=1}^{N}p_n(c|X) \quad (6)$$

where $P(c|X)$ becomes the estimated density of the class labels.

Given that live forensics consists of data with continuous features, the learning methods would be more suitable in this context.

3.4. Stages of preparing live forensic data

This section gives general insights that are used while preparing live forensic data for investigation using machine learning approaches.

- Feature Engineering: Feature engineering is a process in this framework that distinctively allow for the selection of various subsets of exclusive features from a set of collected live data coming from ECO. This allows one to be able to obtain important or selective features that

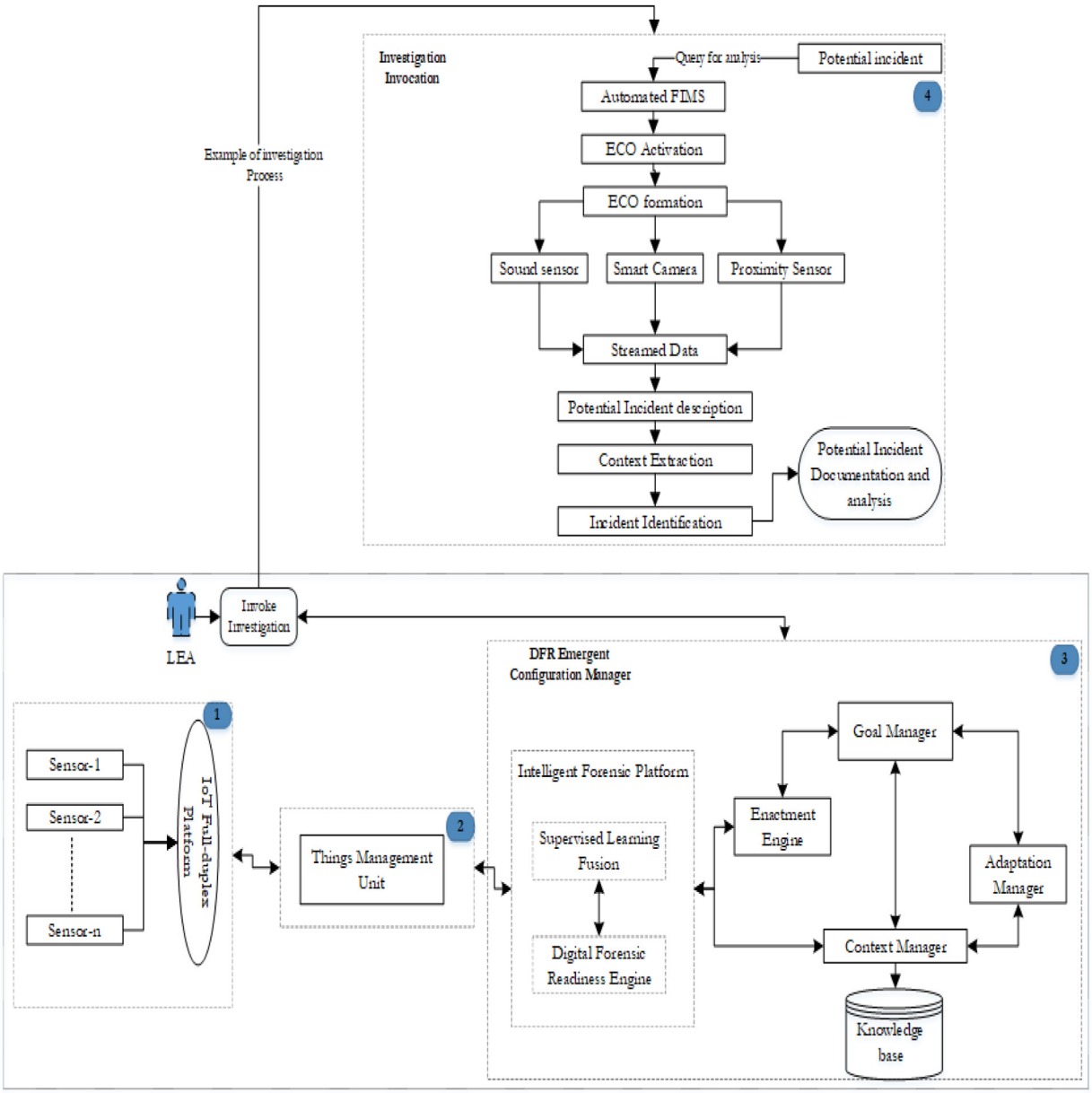

Fig. 2. Scenario representation based on ECO (adapted from [1]).

allow proper classification. Using feature engineering also allows for the identification of a multitude of features that can assist during classification [41]. The suggested framework when fully implemented in an IoT environment to realise a digital forensic tool would aim to utilize Machine learning algorithms such as those discussed earlier to automated feature identification in order to avoid unnecessary redundancies during feature selection and elimination. While this study does not employ a direct scenario with respective implementation, it sensitizes on the need for employing supervised technique during live forensics.

- Feature Selection: Feature selection basically attempts to distinctively select from a subset of features X, a set of Y features and through this it minimally identifies the sufficient features that are necessary to improve how a given classified model can accurately predict the outcome of live analysis process [42]. Consequently, the ultimate goal of employing feature selection in this context is to allow, during forensic analysis, the learning model to identify useful and key subsets from a distribution of features and be able to map them to the original class distribution based on the identified features [43].
- Feature Elimination: Feature elimination emphasizes that a given live forensic dataset can forensically be used to judge the exclusive features present in the dataset as being useful/ relevant or not. In this context relevant/useful has been used to show whether those features are in a position of being eliminated or not. It is imperative top note that a number of factors may contribute to this elimination, for example, in ECO there exist rampant dynamic configuration and reconfiguration of devices of emerging devices that provide massive data. Also, over time some of the changes in the technical and technological aspects may hinder identification of features that needs to be eliminated and this may be a bottleneck when it comes to digital forensics. In this process, care should be taken because it is possible to remove what may be useful in the process [44,45]. Also, [44,45] has identified different strategies for forensic profiling adversaries in the wake of a forensic investigation while doing feature elimination. This is owing to the fact that behavioral changes is a common aspect.
- Feature Normalization: The importance of normalizing the features of a given data during live forensic would be to independently give room to normalize each of the feature based on some given range given that extracted data from different sources normally consist of a variety of features [46]. Also using different feature distances measures like the Euclidean distances, Manhattan distance etc may assign different weights to these extracted data. Based on that feature normalization becomes important because it can balance the range of these features based on the computing similarity. Generally, the procedure involves transforming the feature components statistically such that the values are able to give correct or better estimates of the features. Based on the collected data from IoT environments, the features of the collected data could be transformed based on a uniform random variable, based on ranks or based on some scaling approaches [46].
- Feature Representation: One core characteristics of a dynamic environment, such as the ECO, is the integration of multiple sources of information into a centralized process. Therefore, when several features from multiple sources are aggregated over a given spectrum of analysis, there will be a need to define a unique format for the instances in each feature vector. Furthermore, a feature space data format can be defined to accommodate the potential heterogeneity of data. As a way to ensure such process, the feature representation phase will define the data format.

3.5. Implementation Feasibility of Machine Learning in ECO in IoT Platform

The generic framework given in Fig. 1, is further designed using the architectural model for ECO in IoT platform developed by [1], as shown in Fig. 2. This consideration is then used to develop an hypothetical investigative scenario, through which the implementation feasibility of supervised machine learning approach and digital forensic readiness (DFR) can be evaluated. The integrated ECM proposed in this scenario consist of a goal manager, adaptation manager, context manager, enactment engine, knowledge base, and digital forensic readiness engine. By function, the goal manager interprets the goal of the user (a forensic investigator in our case) to coordinate ECOs that can be used to achieve the goal. The Adaptation manager attempts to align the ECOs to the dynamism of the goal and the environment. The context manager on the other hand, attempts to maintain the contextual dynamism of the ECOs, while the enactment engine is responsible for enacting ECOs by ensuring that ECOs constituents perform functionalities in specific sequence. The knowledge base serves as the systems container for the ECO. Refer to [1] for details of these components. The DFR engine is a mechanism that identifies, captures and stores potential digital content from the IoT platform based on pre-defined rules (adaptive rule table). Such pre-defined rules are aligns with the context maintained by the context manager, and the specific sequence of functionalities ensured by the enactment engine.

Thus, the DFR engine provide a preemptive and proactive approach for IoT information collection, in manner that can be used during a forensic investigation. Furthermore, the notion of DFR posit that the forensic soundness of the information collected is ensured, suitable for litigation. Input from the machine learning process plays a critical role in this regard. To correctly identify the composition of potentially viable digital evidence, rules based on decision trees (Random Forest and C4.5 decision trees for instance), and even Naive Bayes algorithm can be leveraged to identify and extract potential digital evidence from a given context within a given the sequence of functionalities of each ECOs. However, to ensure the degree of accuracy of such rules, distance measures such as Manhattan and Minkowski distance functions, and other dissimilarity metrics as depicted in [47] can be leveraged. Furthermore, the process of classifying potential digital information would require an algorithm that is robust to noise. Also, such an algorithm would be fairly robust to dimensionality challenges often associated with such as exploratory process. In this regard, classifiers such as multi-class support vector machine, and the Neural Network families can be considered.

4. Hypothetical investigative scenario

We present a hypothetical scenario that dynamically conducts forensic activities that aid in potential incident identification by focusing on three main aspects: Collecting streamed sensor data, analysis of the state of the collected evidence through dynamic discovery and reporting the findings. In response to reports about a potential incident in the harbor area (that produced a large bang), a Law Enforcement Agent (LEA) requests a dynamically Automated Forensic Incident Management System (AFSM) to “analyze the potential incident”. An ECO is dynamically formed from the dynamically discovered “things” located around the crime scene, e.g., sound sensors and a camera controlled by the CMAs. The sensors and the camera stream sound and video to the law enforcement server, respectively. Consequently, the camera is also able to track suspects activities spontaneously. The server can process the streamed data and classify the incident (e.g., shooting, accident etc) and this can be used to draw conclusions that helps in the formation of an objective forensic hypothesis. As is shown in Fig. 2, the hypothetical scenario comprise an emergent configuration manager and an investigation invocation modules that plays a significant role in potential incident identification. The concept behind this approach is that, like a broker, a requester can query for the discovery of things and once the things are discovered the response is relayed to the requester in order to draw conclusion on the potential incident as is illustrated by Fig. 3.

4.1. Investigation invocation

From the aforementioned scenario, the investigation is invoked by way of querying the AFIMS, that has a knowledge base to assist in forensic

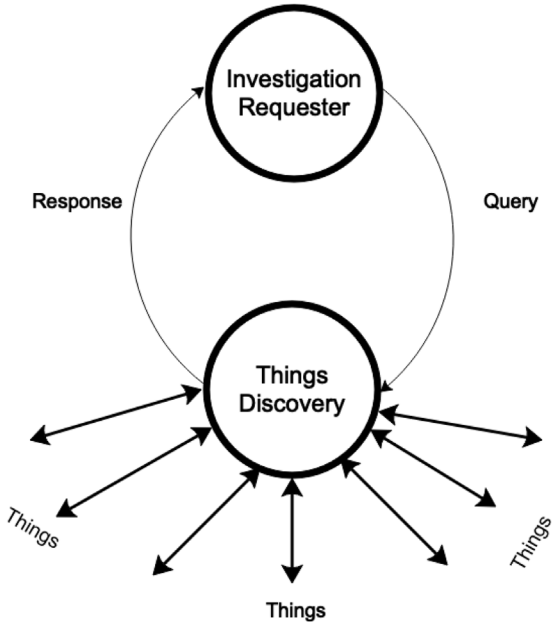

Fig. 3. Phases of things discovery.

incident identification approaches in IoT environments. This allows analysis to be conducted effectively based on rapid dynamic discovery of things that surround the crime scene. Consequently, among the important aspect of the investigation invocation is the activation of ECO by the AFMIS that makes it possible for streamed sensor data to be analysed using supervised machine learning approaches.

4.2. DFR emergent configuration manager

Leveraging the architecture on emergent configuration manager (ECM) developed in [1] to formulate the scenario, an integrated a digital forensic readiness (DFR) engine as a major component is further introduced to address forensic investigation process requirement. As stated in [1,3], an ECM is responsible for the management of the emergent configuration in ways that addresses emergent user needs, adaptive capabilities included. Furthermore, the emergent configuration is referred to as a collaborative environment of diverse 'Things' designed towards a common goal as well as to address a potentially unforeseen contextual dynamism. The integrated DFR emergent configuration manager is therefore a mechanism that is capable of identifying and store potential digital evidence that would otherwise not be available when the investigation process is invoked. The adapted approach further combines the DFR engine, and the supervised machine learning components into the existing ECO architecture, to form an intelligent engine which can be leveraged for investigation. This Idea is further depicted in the example scenario presented in Part-4 of Fig. 2. Specifically, the intelligent unit is responsible for the extraction of context, incident identification, and incident analysis. Whilst the machine learning component of the intelligent forensic platform can be used to extract meaningful pattern from the context extracted from the ECO streamed data, the DFR engine can be used to ascertain and document the potential incident which the investigator would then use to conduct investigation. Suffice it to note that the DFR engine presents a proactive mechanism for a potential investigation. This notion is based on the assertion that a properly developed DFR engine will have the capacity to query and be queried, as well as a storage potential. However, this could further open up a potential incident categorization and identification challenge, as extensively highlighted in [48].

5. Discussions

The heterogeneity and the dynamic composition of an ECO represents a classical feature engineering problem which undermines the reliability (particularly the area under the receiver operating characteristics curve -AUC) of any machine learning (supervised, reinforced, semi-supervised, or unsupervised) approach. Fundamental to this problem is the probability of extracting relevant and useful configuration data that can be leveraged to conduct a live forensic analysis. The proposed ECO framework (as depicted in the High-level approach in Fig. 1) presents baseline for the realization of a live forensic analysis in any given IoT environment. However, the challenge of forensic analysis, specifically in a dynamic environment, contains myriad of challenges which should be addressed going forward. One of such challenges include the potential of large feature space which is typically termed "curse of dimensionality" in the soft computing discipline. Additionally, the dynamic discovery of things within the proximity of the crime scene as is shown in the hypothetical scenario (see Section 4) shows that fundamentally streamed sensor data could easily be used to conduct live forensic analysis for purposes of incident identification given that the AFIMS could only triggered when a potential incident is thought to have occurred. Based on this, the authors have been able to put across far-reaching propositions that have a focus on how ECO can be utilised in IoT environment to achieve this objective. Consequently, several feature engineering approaches and supervised machine learning approaches have been developed in the soft computing domain to attempt to address such challenge. However, such solution would further require a context dependent approach to better engineer and contextualize the solution to achieve a reliable outcome. Expectantly, the induction of dimension reduction algorithms would generate a context-dependent weight for features within the feature space. Consequently, the weight of a given feature within the feature space can be used to redesign the forensic analysis process. Studies in [49,50] have explored diverse supervised learning algorithms that can be applied to augment such a live (near-real time) analysis process. Besides, another potentially fundamental challenge is the process of ascertaining the relevance of each feature in the feature space, beyond the semantic weight of the feature. Whilst dimensionality reduction algorithms, such as principal component analysis, are suitable and fundamentally required in any dynamic data classification tasks, the degree of (forensic) evidential usefulness of the feature presents a logical challenge towards the reliability of the forensic process. This is essentially important in a live forensic analysis process where a low computational time is required. In addition, the degree of accuracy is required to be very high, as false error rate is expected to the minimized to 0.001 [51].

The definition of appropriate metrics of evaluation would be another area of interest in this proposed approach. Existing soft computing metrics such as accuracy, specificity, equal error rate, AUC, and F-measure are often suggested to be effective. However, given the contextual and live-nature of the proposed analysis approach, the need to develop a context-based evaluation metrics could arise. This is could be essential when no apriori information or database might be available. The lack of apriori information would evidently suggest that an unsupervised machine learning approach would be considered, or a reinforced learning approach. This can, however, be extensively explored in the experimentation phase of the proposed approach.

Peculiar to the proposed analysis process is the potential of a supervised ML approach to data analysis. Whilst the unsupervised approach could provide a direct approach to analysis through clusterization, the induction of a supervised approach can be used to finetune the degree of accuracy of the analysis process. A supervised approach posits that the input data stream is parsed into classes which are then fed into the learning algorithm(s). The class-formation would, therefore, be a potential challenge in an ECO in IoT, which is characterized by heterogeneous streams of data sources. However, given

that ECOs are systems formed by a set of things, with their services, functionalities, and applications, that cooperate temporarily to achieve some user goal, an investigator could leverage the commonality to define classes. For instance, input streams from applications and services from different 'Things' can be classified distinctly using identifiers from such sources. Consequently, this can provide a baseline for extracting classes for the supervised machine learning process. However, there exist the potential of miss-classification except when a fundamental framework is defined as a baseline for class formation. Therefore, a forensic analysis process in an emergent configuration in IoT environment would require the definition of such class identification and extraction process.

6. Conclusion and future works

We explained the importance of a context-dependent on-the-fly forensic analysis process to facilitate live forensic analysis on emergent configuration in IoT environment. Specifically, our conceptual framework leverages NIST SP 800-86 standard and supervised ML approaches.

Such a proposed approach has the potential to be a game changer in IoT forensics, although extensive evaluations on different datasets from a broad range of applications are required. However, careful planning on the evaluation scenarios is required. Hence, one potential research agenda is to collaborate closely with relevant stakeholder groups to design and develop different evaluation scenarios.

Once these evaluation scenarios have been developed, we will also evaluate a prototype of our proposed framework in the different scenarios. This will allow us to identify any limitations, for example in the ML techniques, scenarios, or configurations.

Declaration of Competing Interests

The authors have no competing interests to declare.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsir.2020.100122.

References

[1] F. Alkhabbas, R. Spalazzese, P. Davidsson, Eco-iot: An architectural approach for realizing emergent configurations in the internet of things, European Conference on Software Architecture (2018) 86–102.
[2] F. Alkhabbas, R. Spalazzese, P. Davidsson, Architecting emergent configurations in the internet of things, 2017 IEEE International Conference on Software Architecture (ICSA) (2017) 221–224.
[3] F. Alkhabbas, R. Spalazzese, P. Davidsson, Emergent configurations in the internet of things as system of systems, 2017 IEEE/ACM Joint 5th International Workshop on Software Engineering for Systems-of-Systems and 11th Workshop on Distributed Software Development, Software Ecosystems and Systems-of-Systems. (JSOS) (2017) 70–71.
[4] F. Alkhabbas, M. Ayyad, R.-C. Mihailescu, P. Davidsson, A commitment-based approach to realize emergent configurations in the internet of things, 2017 IEEE International Conference on Software Architecture Workshops (ICSAW) (2017) 88–91.
[5] F. Alkhabbas, R. Spalazzese, P. Davidsson, Iot-based systems of systems, Proceedings of the 2nd edition of Swedish Workshop on the Engineering of Systems of Systems (SWESOS 2016) (2016) .
[6] W. Wu, J. Alvarez, C. Liu, H.-M. Sun, Bot detection using unsupervised machine learning, Microsystem Technologies 24 (2018) 209.
[7] N.L. Beebe, J.G. Clark, Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results, Digital Invest. 4 (2007) 49.
[8] H.H. Pajouh, R. Javidan, R. Khayami, D. Ali, K.-K.R. Choo, A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in iot backbone networks, IEEE Transactions on Emerging Topics in Computing (2016) .
[9] A.J. Tall'on-Ballesteros, J.C. Riquelme, Data mining methods applied to a digital forensics task for supervised machine learning, Computational Intelligence in Digital Forensics: Forensic Investigation and Applications (2014) 413–428.
[10] M. Mohlala, A.R. Ikuesan, H.S. Venter, User attribution based on keystroke dynamics in digital forensic readiness process, 2017 IEEE Conference on Application Information Network Security (AINS) (2017) 124–129.
[11] I.R. Adeyemi, S. Abd Razak, M. Salleh, Understanding online behavior: exploring the probability of online personality trait using supervised machine-learning approach, Front. ICT 3 (2016) 8.
[12] N. Huang, J. He, B. Zhao, G. Liu, Forensic analysis of distributed computing network based on decision values, 2016 International Symposium on Computer Consumer Control (IS3C) (2016) 423–427.
[13] M. Buccoli, P. Bestagini, M. Zanoni, A. Sarti, S. Tubaro, Unsupervised feature learning for bootleg detection using deep learning architectures, 2014 IEEE International Workshop on Information Forensics Security (WIFS) (2014) 131–136.
[14] V.H. Bhide, S. Wagh, i-learning iot: An intelligent self learning system for home automation using iot, 2015 International Conference on Communications and Signal Processing (ICCSP) (2015) 1763–1767.
[15] H.-S. Ham, H.-H. Kim, M.-S. Kim, M.-J. Choi, Linear svm-based android malware detection for reliable iot services, J. Appl. Math. 2014 (2014) .
[16] N.M. Karie, V.R. Kebande, H. Venter, Diverging deep learning cognitive computing techniques into cyber forensics, Forensic Sci. Int.: Synergy 1 (2019) 61.
[17] M.N.A. Khan, Digital Forensics using Machine Learning Methods Ph.D. thesis, school University of Sussex, 2008.
[18] S. Costantini, G. De Gasperis, R. Olivieri, Digital forensics and investigations meet artificial intelligence, Ann. Math. Artif. Intel. (2019) .
[19] F. Mitchell, The use of artificial intelligence in digital forensics: An introduction, Digital Evid. Elec. Signature L. Rev. (2010) .
[20] P.P. Ray, A survey on internet of things architectures, J. King Saud Univ.-Comput. Inform. Sci. (2018) .
[21] S. Khorashadizadeh, A.R. Ikuesan, V.R. Kebande, Generic 5g infrastructure for iot ecosystem, International Conference of Reliable Information and Communication Technology (2019) 451–462.
[22] R.-C. Mihailescu, R. Spalazzese, C. Heyer, and P. Davidsson, A role-based approach for orchestrating emergent configurations in the internet of things, arXiv preprint arXiv:1809.09870 (2018).
[23] V.R. Kebande, I. Ray, A generic digital forensic investigation framework for internet of things (iot), 2016 IEEE 4th International Conference on Future Internet of Things Cloud (FiCloud) (2016) 356–362.
[24] S. Li, K.-K.R. Choo, Q. Sun, W.J. Buchanan, J. Cao, Iot forensics: Amazon echo as a use case, IEEE Internet Things J. 6 (2019) 6487.
[25] X. Zhang, O. Upton, N.L. Beebe, K.-K.R. Choo, Iot botnet forensics: A comprehensive digital forensic case study on mirai botnet servers, Forensic Sci. Int.: Digital Invest. 32 (2020) 300926.
[26] X. Zhang, K.-K.R. Choo, Digital Forensic Education: An Experiential Learning Approach, Vol. 61, Springer, 2019.
[27] X. Zhang, K.-K.R. Choo, N.L. Beebe, How do i share my iot forensic experience with the broader community?. an automated knowledge sharing iot forensic platform, IEEE Internet of Things J. 6 (2019) 6850.
[28] O. Alkadi, N. Moustafa, B. Turnbull, K.-K.R. Choo, A deep blockchain framework-enabled collaborative intrusion detection for protecting iot and cloud networks, IEEE Internet Things J. (2020) .
[29] M. Saharkhizan, A. Azmoodeh, A. Dehghantanha, K.-K.R. Choo, R.M. Parizi, An ensemble of deep recurrent neural networks for detecting iot cyber attacks using network traffic, IEEE Internet Things J. (2020) .
[30] F. Alkhabbas, M. De Sanctis, R. Spalazzese, A. Bucchiarone, P. Davidsson, A. Marconi, Enacting emergent configurations in the iot through domain objects, International Conference on Service-Oriented Computing (2018) 279–294.
[31] R.-C. Mihailescu, J. Persson, P. Davidsson, U. Eklund, Towards collaborative sensing using dynamic intelligent virtual sensors, International Symposium on Intelligent and Distributed Computing, Springer, 2016, pp. 217–226.
[32] A. Tegen, P. Davidsson, R.-C. Mihailescu, J.A. Persson, Collaborative sensing with interactive learning using dynamic intelligent virtual sensors, Sensors 19 (2019) 477.
[33] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, U. Abbasi, Improved churn prediction in telecommunication industry using data mining techniques, Appl. Soft Comput. 24 (2014) 994.
[34] C. Cortes, V. Vapnik, Support-vector networks, Machine Learn. 20 (1995) 273.
[35] Q. He, J.-F. Chen, The inverse problem of support vector machines and its solution, 2005 International Conference on Machine Learning and Cybernetics, Vol. 7, IEEE, 2005, pp. 4322–4327.
[36] Z. Liu, L. Bai, Evaluating the supplier cooperative design ability using a novel support vector machine algorithm, 2008 12th International Conference on Computer Supported Cooperative Work in Design, IEEE, 2008, pp. 986–989.
[37] L.-m. He, X.-b. Yang, F.-s. Kong, 2006 International Conference on Machine Learning Cybernetics, IEEE, 2006, pp. 3503–3507 Support vector machines ensemble with optimizing weights by genetic algorithm.
[38] Y.N. Dewi, D. Riana, T. Mantoro, Improving na"ive bayes performance in single image pap smear using weighted principal component analysis (wpca), 2017 International Conference on Computing, Engineering, and Design (ICCED), 1 (2017) .
[39] S.N.N. Alfisahrin, T. Mantoro, Data mining techniques for optimization of liver disease classification, 2013 International Conference on Advanced Computer Science Applications Technologies, IEEE, 2013, pp. 379–384.
[40] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5.
[41] V.N. Garla, C. Brandt, Ontology-guided feature engineering for clinical text classification, J. Biomed. Inform. 45 (2012) 992.
[42] M. Dash, H. Liu, Feature selection for classification, Intelligent Data Anal. 1 (1997) 131.
[43] P.M. Narendra, K. Fukunaga, A branch and bound algorithm for feature subset selection, IEEE Trans. Comput. 917 (1977) .
[44] Z.M. Hira, D.F. Gillies, A review of feature selection and feature extraction methods applied on microarray data, Adv. Bioinform. (2015) .
[45] K. Kira, L.A. Rendell, et al., The feature selection problem: Traditional methods and a new algorithm, Aaai, Vol. 2 (1992) 129–134.

[46] S. Aksoy, R.M. Haralick, Feature normalization and likelihood-based similarity measures for image retrieval, Pattern Recognit. Lett. 22 (2001) 563.
[47] A.R. Ikuesan, M. Salleh, H.S. Venter, S.A. Razak, S.M. Furnell, A heuristics for http traffic identification in measuring user dissimilarity, Human-Intelligent Syst. Integration 1 (2020) .
[48] A. Al-Dhaqm, S. Razak, D.A. Dampier, K.R. Choo, K. Siddique, R.A. Ikuesan, A. Alqarni, V.R. Kebande, Categorization and organization of database forensic investigation processes, IEEE Access 1 (2020) .
[49] A.R. Ikuesan, S.A. Razak, H.S. Venter, M. Salleh, Polychronicity tendency-based online behavioral signature, Int. J. Machine Learn. Cybernet. 10 (2019) 2103.
[50] I.R. Adeyemi, S.A. Razak, M. Salleh, H.S. Venter, Observing consistency in online communication patterns for user re-identification, PLOS ONE 11 (2016) e0166930.
[51] A.R. Ikuesan, H.S. Venter, Digital behavioral-fingerprint for user attribution in digital forensics: Are we there yet? Digital Invest. 30 (2019) 73.