applied sciences

MDPI

Review

Machine-Learning Forensics: State of the Art in the Use of Machine-Learning Techniques for Digital Forensic Investigations within Smart Environments

Laila Tageldin [1,*] and Hein Venter [2]

[1] Department of Computer Science, Sudan University of Science and Technology, Khartoum 11111, Sudan
[2] Department of Computer Science, University of Pretoria, Pretoria 0002, South Africa; hventer@cs.up.ac.za
* Correspondence: laylataj@hotmail.co.uk

Abstract: Recently, a world-wide trend has been observed that there is widespread adoption across all fields to embrace smart environments and automation. Smart environments include a wide variety of Internet-of-Things (IoT) devices, so many challenges face conventional digital forensic investigation (DFI) in such environments. These challenges include data heterogeneity, data distribution, and massive amounts of data, which exceed digital forensic (DF) investigators' human capabilities to deal with all of these challenges within a short period of time. Furthermore, they significantly slow down or even incapacitate the conventional DFI process. With the increasing frequency of digital crimes, better and more sophisticated DFI procedures are desperately needed, particularly in such environments. Since machine-learning (ML) techniques might be a viable option in smart environments, this paper presents the integration of ML into DF, through reviewing the most recent papers concerned with the applications of ML in DF, specifically within smart environments. It also explores the potential further use of ML techniques in DF in smart environments to reduce the hard work of human beings, as well what to expect from future ML applications to the conventional DFI process.

Keywords: IoT devices; smart environments; digital forensics; machine-learning techniques

check for updates

Citation: Tageldin, L.; Venter, H. Machine-Learning Forensics: State of the Art in the Use of Machine-Learning Techniques for Digital Forensic Investigations within Smart Environments. *Appl. Sci.* **2023**, *13*, 10169. https://doi.org/10.3390/app131810169

Academic Editors: Dimitris Mourtzis and Gino Iannace

Received: 21 June 2023
Revised: 16 August 2023
Accepted: 7 September 2023
Published: 10 September 2023

1. Introduction

Currently, smart environments offer various technologies and services, such as smart transport systems, smart vehicles, smart homes, smart urban lighting, integrated travel ticketing, smart energy grids, and smart sensors [1]. These technologies strongly depend on the use of small electronic chips and electromechanical devices (i.e., IoT devices), such as sensors, wireless technologies, radio-frequency identification (RFID) devices, localisation technologies, and near-field communication devices [1].

The wide variety of IoT devices used within smart environments makes it very difficult to perform digital forensics (DF) in this environment. The challenge for DF professionals and practitioners is that standard industrial DF equipment and its capabilities concerning conventional computing operating systems are not coping with the smart environment due to its complex, heterogeneous, and distributed nature [2].

The problem raised in this paper is that little to no reliable DF applications or DF directives currently exist to retrieve data from Internet-of-Things (IoT) devices in the event of a digital attack, an active investigation, or a litigation request within a smart environment [3]. Thus, researchers and practitioners in the DF field are working hard to define new techniques and tools to improve DF capabilities for coping with this problem. For example, it is currently possible to gather evidential data from a computer hard drive or even a mobile phone. However, when it comes to smart devices like smart watches or smart switches, there is no standard interface to connect to in order to reach their storage components. In yet another example, many such devices do not host large amounts of storage space, but rather communicate their data to other devices. On the other hand, some of these devices generate such vast amounts of data that, should an investigator not act fast enough, evidential data might be lost forever. The vast volume of data as well as the short-lived data created by these smart devices become humanly impossible to sift through. ML techniques may potentially be employed to assist with this dilemma in order to find evidence much more effectively in a much shorter time span.

The numerous challenges that face traditional digital forensic investigation (DFI) in smart environments result from the heterogeneity of, distribution of, and huge amounts of data involved. This exceeds the capabilities of human DF investigators to cope with all these challenges in a short time. It severely slows down or even incapacitates the conventional DFI process. Due to the rapid pace at which digital crimes are committed, better and more intelligent DFI techniques are sorely needed, especially in smart environments. Machine-learning (ML) techniques might offer a solution to these challenges [4].

ML has recently been applied in DFI and is still evolving; for example, Ref. [5] designed a new framework known as IoTDots to help protect the data collected by various smart devices and applications. This features two main components: the IoTDots analyser and the IoTDots modifier. The former scans the source code of the applications and detects forensic information. The latter automatically inserts tracking logs and reports the results.

In an IoT system, particularly in the case of emergent configurations, data might also be dynamic, making it difficult to classify information during live forensics. In this sense, live forensics refers to a forensic investigation that is done in near-real time. Hence, ref. [6] proposed a conceptual framework based on supervised machine-learning techniques. One of the advantages of using supervised ML techniques in live forensics is the ability of such techniques to predict possible events based on past occurrences. In addition, automated feature identification was used to prevent redundancy throughout feature selection and elimination.

The importance of ML in DFIs should not be underestimated, since such intelligent technologies have the potential to support and significantly enhance the conventional DFI process. ML technologies can potentially assist in the automation of manual DFI processes when significant volumes and a large variety of data must be analysed. Using more intelligent techniques will increase the chances of identifying and successfully investigating cybercrimes in modern smart environments. This will help DF specialists get to the root cause much faster and more efficiently [6].

For all the reasons mentioned above, ML holds great potential for DFIs. However, it is a foreign field to most DF investigators, and the scope for new research is vast. That being said, there exists a small corpus of research where ML technology was used to investigate digital crimes [4].

ML techniques, which are often used to predict behaviour, make use of pattern recognition software for investigators to analyse huge amounts of data. ML techniques seek to learn from historical perspectives so as to predict future behaviour. Therefore, by using ML techniques, investigators may gain the capability to recognise patterns of criminal activity and learn from the historical data when, where, and how the cybercrime probably took place.

The remainder of this paper is structured as follows. Section 2 provides some background on digital forensics, the ISO/IEC 27043 international standard on the DFI process, smart environments, and ML. Section 3 presents state-of-the-art ML techniques used in digital forensics. Section 4 discusses the role of ML techniques in the DFI process and future directions in the use of ML in this process. The paper is concluded in Section 5.

2. Background

This section deals with digital forensics, the internationally standardised DFI process, smart environments, and machine learning—all the important concepts of which the reader needs to take cognisance in this paper.

2.1. Digital Forensics (DF)

DF forms part of the greater field of forensic science. DF investigators are responsible for retrieving and investigating data on digital devices. As these new and updated platforms work with IoT and cloud technologies in smart environments, industry and practitioners are struggling to develop DF tools and procedures to keep up with the challenges involved. These technologies may even be embedded electronics or computing systems with specific functionalities that may exist as part of a larger platform [3].

A systematic and standardised process that has been developed to perform DFI is captured in the international standard "ISO/IEC 27043:2015 Ref. [7]–Incident investigation principles and processes", which is briefly elaborated on in the next section.

2.2. ISO/IEC 27043

The ISO/IEC 27043 international standard was initially proposed by Valjarevic and Venter [8] to handle DF incident investigation principles and processes [7]. Figure 1 shows a high-level overview of this ISO/IEC 27043 international standard.

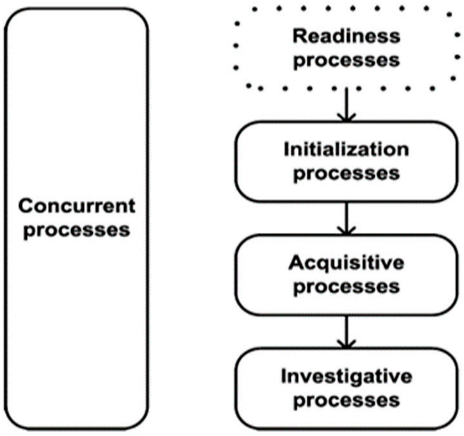

Figure 1. High-level overview of the 27043 international standard [9].

The conventional DF process (i.e., the process that had been followed before ISO/IEC 27043 was imposed) was only concerned with initialisation, acquisitive, and investigative processes. However, the conventional DF process consisted of various disparate process models that were not harmonised. Therefore, Valjarevic and Venter [8] considered all relevant models and other standards so as to address the disparities and harmonise them into a single standardised model, known as ISO/IEC 27043. In addition to the harmonisation effort, Valjarevic and Venter [8] added the readiness and concurrent process classes.

However, since it has not been tailored for IoT and smart environments, using the ISO/IEC 27043 DFI process within the smart environment is still challenging, due to the wide variety of IoT devices that exist within this environment. The next section briefly describes smart environments to allow the reader to understand the solutions proposed by recent research.

2.3. Smart Environments

The smart environment comprises various types of smart devices, sensors, and computers that are connected to the internet and embedded in numerous objects within this environment. Smart environments have fast grown into a network of internet-enabled devices, also known as IoT devices [8]. Currently, IoT devices are adopted in almost all parts of our lives, for instance, home temperature management, smart lighting, smart appliances, smart sensors, and smart cities [8].

Although a smart environment may improve our quality of life, it also provides a new set of previously untapped data with tremendous forensic value, due to the huge amount of data generated in this environment [4]. The rapid pace at which digital crimes are conceptualised and committed makes it essential to develop better and more intelligent DFI techniques, especially in smart environments. ML techniques might offer a solution to these challenges [4,10,11].

However, researchers argue that smart-environment DF is still at a progress level where an international standard implementation of infrastructure for smart cities has not been completed yet. Meanwhile, this provides an opportunity for law-enforcement organizations and investigators to swiftly expand their DF solutions and capabilities [10,11].

The following section presents a brief background on ML techniques.

2.4. Machine Learning (ML)

The application of ML in the field of DF has given rise to a new discipline known as machine-learning forensics (MLF), which has the capacity to detect criminal patterns, anticipate criminal activities (e.g., where and when crimes are likely to occur), and automate DF investigative procedures. To conduct MLF, an adapted DF framework is required, which must be capable of capturing and analysing data in smart environments—regardless of whether devices in this smart environment are connected to the internet via wired or wireless networking interfaces [12].

Furthermore, ML is an approach to artificial intelligence (AI) that allows a system to learn on its own from experience and example, rather than from programming. In other words, ML is used to describe a system that continually learns and makes decisions based on data rather than programming [4]. ML is not only utilised for AI goals such as simulating human behaviour but also to minimise human effort and time spent on complex and time-consuming jobs. ML techniques include supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning is a method of developing AI by training a computer program on labelled input data for a certain output. The model is trained until it recognises the underlying patterns and correlations between the input data and the output labels, allowing it to produce appropriate labelling results when given previously unseen data. Supervised learning excels in classification and regression issues. The goal of supervised learning is to make meaning out of data in the context of a given topic [13]. Supervised learning was proposed by some researchers to improve DFIs in smart environments, as mentioned in Section 3.

In contrast to supervised learning, unsupervised learning is presented with unlabelled data and is designed to detect patterns or similarities on its own. In other words, unsupervised learning techniques include two types: clustering and association, which find all kinds of unknown patterns in data and help to find features that can be useful for categorisation [13].

Reinforcement learning is totally different from both supervised and unsupervised ML techniques. The relationship between supervised and unsupervised techniques is the presence or absence of data labelling. However, reinforcement learning is a subfield of machine learning concerned with how intelligent agents should behave in a given environment. When the system being represented is independent and not affected by an external actor, Markov models are utilised [14]. Markov chains are the simplest type of Markov model and are used to represent systems where all states are observable. Markov chains show all possible states. Applications of this type of model include prediction, which is a probabilistic technique that uses Markov models to predict the future behaviour of some variable based on its current state and can be used in many domains.

However, ML has a substantial influence on DF and has various applications in this sector. These applications can improve the overall efficiency of DFIs by finding trends and patterns, similarities, anomalies, and other characteristics inside digital evidence. Therefore, forensic professionals can produce leads and solve crimes in less time and with fewer resources. These advancements lead to the second major contribution of ML applications, which is a reduction in the cost of a DFI [15].

Section 3 presents an overview of research papers using ML techniques in DF as proposed by different researchers between 2018 and 2023. Section 4 then presents (in table form) the contribution of ML to the DFI process, in order to identify gaps in the reviewed papers and suggest high-level solutions.

3. State-of-the-Art Use of Machine-Learning Techniques in Digital Forensics

Due to the challenges that traditional DFIs face in smart environments (i.e., the heterogeneity, distribution, and huge amount of data, managing which in a short time exceeds human capabilities), ML seems to be the best solution for these environments [4]. These technologies can automate the laborious DFI operations of analysing huge amounts and wide ranges of data to increase the likelihood of successfully detecting and investigating cybercrime. This would greatly aid DF professionals in rapidly and effectively determining the fundamental causes of incidents [6].

This section presents how ML techniques can be further used to support DF in smart environments and to reduce hard work and time spent through reviewing MLF research papers between 2018 and 2023.

As mentioned before, the amount of data collected by IoT devices and sensors is immense and contains valuable forensic evidence. This data can help identify and prevent unauthorised access within smart environments. The authors of [5] designed a new framework known as IoTDots to help protect the data collected by various smart devices and applications. This features two main components: the IoTDots analyser and the IoTDots modifier. The former scans the source code of the applications and detects forensic information. The latter automatically inserts tracking logs and reports the results. However, to reduce the amount of manual analysis required in DFI, ref. [16] proposed a methodology for the automatic prioritising of suspicious file artefacts. Rather than providing the final analysis results, this methodology aims to predict and recommend the artefacts that are likely to be suspicious. A supervised machine-learning approach is used, which makes use of previously processed case results. One of the most discussed challenges in DFI is the growing volume of data. Since the majority of file artefacts on seized devices are usually irrelevant to the investigation, manually retrieving suspicious files relevant to the investigation is very difficult. In support of DF, “intelligent methods” are proposed, which include the ability of computers to learn a specific task from data, data mining, machine learning, soft computing, and traditional artificial intelligence. This term is commonly used to express ways to automate problem solving in DF, and two main intelligent approaches are utilised, namely rule-based and anomaly-based [17]. The authors of [18] introduced a novel and practical DF capability for smart environments, since current smart platforms lack any digital forensic capability for identifying, tracing, storing, or analysing data generated in these environments. The collector and the analyser are the two main components of VERITAS. The collector employs mechanisms to automatically collect forensically relevant data from the smart environment. The analyser then uses a first-order Markov chain model to extract valuable and usable forensic evidence from the collected data for the purposes of a forensic investigation. Therefore, to discover and declare the presence of adversaries, DF necessitates intensive data analysis, such as retrieving and confirming system logs, blockchain information evaluation, and so on. Hence, ref. [19] proposed a blockchain-assisted shared audit framework to analyse DF data in an IoT environment. This was created to identify the sources and causes of data scavenging attacks in virtualised resources. It uses blockchain technology to manage access logs and controls. Using logistic regression ML and cross-validation, access-log data is examined for the consistency of adversary event detection. The number of cases needing DF competence and the volume of data to be processed have overburdened digital forensic investigators. Automated evidence processing based on artificial intelligence techniques holds considerable potential for speeding up the digital forensic analysis process while improving case-processing capacity [4]. In DFI, automation uses ML techniques for classification. ML techniques can obtain important information for investigations more efficiently by exploiting existing digital evidence-processing knowledge. Additionally, digital-evidence triage was developed for the prompt detection, processing, and interpretation of digital evidence. Currently, with AI techniques, the investigator determines the priority of device gathering and processing at a crime scene [4]. Furthermore, ref. [20] proposed an intelligent framework based on clustering and classification. The model learns from past crimes, and, when a new crime is registered, some of the crime information needs to be inserted by the investigator, such as the crime type, location, and time. The clustering process then automatically groups the new crime with previous similar crimes in the system using the k-nearest neighbour and crime-matching classification algorithms. In this way, the investigator can gain insights into the pre-investigation process by exploring the new crime, which is then clustered with previous similar crimes. Moreover, with the growth of cybercrime that targets minors, chat logs can be examined to detect and report harmful behaviour to law authorities. This can make a significant difference in protecting youngsters on social media platforms from being abused by cyber predators. Since DFI is done primarily by hand, the enormous volume and variety of data cause DF investigators to have a tough assignment; Ref. [21] suggested an approach using a DF process model backed by ML methodologies, to enable the automatic finding of hazardous talks in chat logs. One of the most fundamental characteristics of any smart device in an IoT network is its ability to acquire a bigger set of data than has been produced and then send the obtained data to the destination/receiver server through the internet. Thus, IoT-based networks are particularly vulnerable to simple or sophisticated assaults, which must be discovered early in the data transmission process in order to protect the network against these hostile attacks. The authors of [22] developed and built an intelligent intrusion detection system utilising machine-learning models so that assaults in the IoT network may be discovered. The adaptability of IoT devices raises the probability of continual attacks on them. Due to the low processing power and memory of IoT devices, security researchers have found it challenging to preserve records of diverse attacks performed on these devices during a DFI. The authors of [23] proposed an intelligent forensic analysis mechanism, to automate the detection of attacks on IoT devices based on the machine-to-machine framework. However, the proposed mechanism combines several ML techniques and different forensic analysis tools to detect different types of attacks. Furthermore, by providing a third-party logging server, the problem of evidence gathering has been overcome. To assess the effects and types of attacks and violations, forensic analysis is done on logs utilising a forensic server. In addition, ref. [24] indicated that the use of ML and deep-learning algorithms is effective for cyber-attack discovery, identification, and tracing by proposing a framework of cyber-attacks against smart satellite networks. In addition, IoT forensics and smart environments, with their recognised challenges, provide a great opportunity to develop new forensic tools to make the task of forensic investigators easier, which can be used for acquiring, preserving, and also analysing such forensic data. The authors of [25] proposed a user-friendly tool for smart devices that support WiFi and used smart-environment scenarios to allow forensic investigators, network administrators, and data scientists access to various features of network traffic with simple steps. The proposed tool allows network traffic features to be computed in real time on any WiFi access point running the OpenWrt firmware, avoiding the time-consuming tasks of dumping network traffic and implementing the procedures needed to analyse the captured traffic. On the other hand, due to the lack of examination and available data, ref. [26] selected a smart fridge as an IoT device to be examined and investigated. The dataset was examined using two ML algorithms, Bayes net and decision stump. Each algorithm represents a distinct idea. A stump tree is a simple version of the decision-tree ML technique. The Bayes net is useful for estimating the likelihood of numerous recognised causes, one of which is the occurrence of an event. The validation results indicate that the Bayes net algorithm is more accurate than the decision stump tree.

Research shows that the main issues that face DF investigators in the smart environment are the large volume of data and attack and violation detection. The proposed solutions are summarised in Figures 2 and 3. The authors decided to split the summary into two separate figures, since there were two main themes detected in all existing solutions: the first theme involved MLF solutions for large amounts of data, while the second theme involved MLF solutions for attack and violation detection.

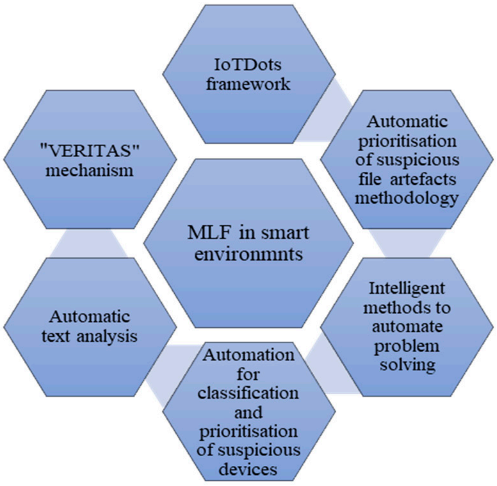
IoTDots framework
"VERITAS" mechanism
Automatic prioritisation of suspicious file artefacts methodology
MLF in smart environmnts
Automatic text analysis
Intelligent methods to automate problem solving
Automation for classification and prioritisation of suspicious devices

Figure 2. MLF solutions for large amounts of data in smart environments.

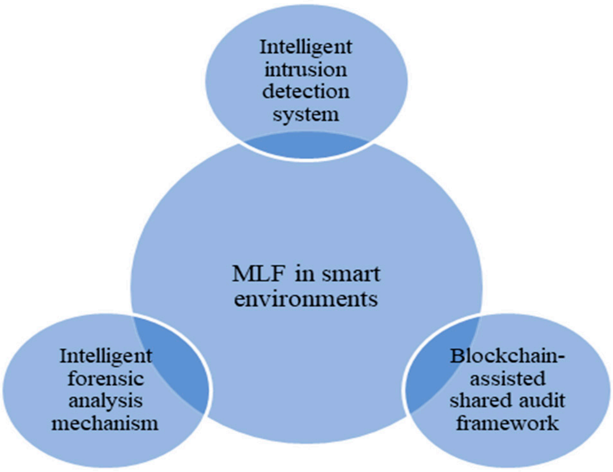
Intelligent intrusion detection system
MLF in smart environments
Intelligent forensic analysis mechanism
Blockchain-assisted shared audit framework

Figure 3. MLF solutions for attack and violation detection in smart environments.

Figure 2 summarises the applications of MLF that were reported in research papers from 2018 to 2023 to serve as proposed solutions for dealing with the large amounts of data generated in smart environments. The following list explains the elements of Figure 2 in more detail:

- The *IoTDots framework* was proposed as a solution to deal with the large amounts of data collected by IoT devices and sensors.
- *Automatic prioritisation of suspicious file artefacts* was proposed as a solution to deal with the growing volume of data and manual retrieval of suspicious files.
- *Intelligent methods to automate problem-solving* were proposed as a solution to deal with the massive amounts of data that must be analysed for digital evidence.
- *Automation using ML techniques for classification and AI techniques for prioritising suspicious devices* was proposed as a solution to deal with the growing number of cases needing DF competence and the large volumes of data to be processed.
- *Automatic text analysis to detect online sexual predatory talks* was proposed as a solution to deal with the growth of cybercrime targeting minors, the large volume of data, and the DFI process, which is done primarily by hand.

- *The "VERITAS" mechanism* to automatically collect and extract forensic evidence from smart environments was proposed as a solution to deal with the large amounts of data that is generated in smart environments.

Figure 3 summarises the applications of ML in DF as proposed in research published between 2018 and 2023 for detecting data attacks and violations in smart environments. The following list explains Figure 3 in more detail:

- *An intelligent intrusion detection system* to detect regular and malicious attacks on data created in smart environments was proposed as a solution to deal with the simple and complex attacks that face IoT networks in particular.
- *A blockchain-assisted shared audit framework* for identifying data-scavenging attacks in virtualised resources was proposed as a solution to deal with attack and violation detection in smart environments.
- *An intelligent forensic analysis mechanism* was proposed as a solution to deal with the probability of continual attacks on IoT devices and the low processing power and memory of these devices.

The following section discusses the impact of MLF on the DFI process.

4. The Impact of MLF on the DFI Process

As can be seen in Section 3, a review of research papers examines the contributions of ML techniques to DF in smart environments. It also identifies digital forensic issues that each of the reviewed papers addresses and proposes solutions that are based on machine learning to improve the DFI process.

Table 1 summarises the role of ML techniques in the DFI process. The column headers present the paper reference number, the ML techniques used, and the main ISO/IEC 27043:2015 process class headings.

Table 1. The role of ML techniques in the DFI process ISO/IEC 27043:2015.

Reference No.	Used ML Technique	Readiness Processes	Initialisation Processes	Acquisitive Processes	Investigative Processes	Concurrent Processes
[5]	Markov chain model				X	
[6]	Supervised machine learning	X			X	
[16]	Supervised machine learning				X	
[17]	Unsupervised identification		X		X	
[18]	Markov chain model				X	
[19]	Logical regression		X		X	
[21]	Logistic regression				X	
[22]	Markov chain model		X			
[23]	Decision-tree algorithm				X	

Table 1 presents the role of ML techniques in ISO/IEC 27043:2015 DFI processes and highlights gaps where ML techniques may be used to improve the processes. An "X" in a particular cell indicates that the specific ML technique was applied in the processes indicated. These techniques contributed mainly to the initialisation and investigative processes of the ISO/IEC 27043:2015 set of standards, and there was a lack of application of ML techniques to the other process areas of this standard.

Applying ML and AI techniques in the areas of ISO/IEC 27043:2015 can automate and improve the DFI process, since the uncovered areas are currently mostly processed manually. For example, the Markov chain model (see Table 1 [5]) already automates the analysis process through two main components, referred to as the 'modifier' and 'analyser' components. The 'modifier' component examines smart applications in search of forensically significant information then modifies the smart application by introducing specialised logs and sending them to a specialised logs database. The 'analyser' component uses data processing and Markov chain models on the logs database to learn the status of the smart environment and the users' activity during the time of the forensic analysis so as to identify possible security violations from people, devices, or smart apps. However, the technique presented in [5] does not focus on the automation of any of the other ISO/IEC 27043:2015 processes. The remainder of Table 1 can be interpreted in a similar fashion.

The authors furthermore propose the integration of ML techniques into ISO/IEC 27043:2015 processes that are not currently covered by ML techniques. While there currently exist several digital forensic process models, Table 1 explores the integration of ML techniques into the ISO/IEC 27043:2015 set of standards, since this standard represents the de facto DFI process owing to its widespread acceptance and ability to integrate new digital forensic methods into its existing processes. Such integration can improve efficiency and reduce time and human effort by automating the manual tasks of the DFI process. For example, it was proposed by [17] that intelligent methods for intrusion detection and real-time intrusion prevention be used with two main techniques—rule-based and anomaly-based—to support DFI. Rule-based techniques mostly utilise databases that include predefined rules to detect known intrusions. The most widespread use of intelligent techniques in this field is connected to the creation of new rules or the optimisation of an enabled set of rules. Anomaly detection may be thought of as a conventional clustering and outlier identification problem in terms of intelligent approaches. Because a detected anomaly is not always proof of intrusion and might be attributed to odd, but proper, user behaviour, this strategy typically has a high false alarm rate.

Furthermore, ref. [27] proposed a DF analysis system based on natural language processing (NLP) techniques and the blockchain for social media data as a significant source of digital evidence that can support various DFIs; NLP is used for data collection, text analysis, and evaluation, and blockchain is used for securing the analysed data and avoiding any other attacks.

Therefore, ref. [28] used a well-structured and realistic dataset to test ML and deep-learning techniques that can be used in the DF analysis process to detect multimedia content manipulations. The dataset was technically validated by convolutional neural networks (CNN) and support vector machine (SVM) algorithms, which concluded that SVM had less processing time than CNN, as one of the goals of incorporating ML techniques into DFI is automating processes and reducing time-consuming work. ML techniques also aim to overcome complexity, consistency, correlation, and data volume, as the evidence is gathered from several sources [29].

Moreover, recent research indicates that ML algorithms are also useful for drone data analysis, by generating judgments and predictions about the likelihood of an event occurring by examining varied datasets with varying volumes. As commonalities between pieces of data may be discovered by clustering common data samples into a single cluster and then visualizing the data clusters, which can then be labelled, it becomes feasible to forecast the trajectories of drones in flight and cluster these as either legitimate flight pathways or compromised ones [30].

Thus, according to the state of the art in applying ML techniques in DFI within smart environments (which mainly involves the initialisation and investigative processes (see Table 1)), ML techniques should be applied more prominently in the readiness processes, acquisitive processes, and concurrent processes of DFI.

Readiness processes in smart environments will be improved by applying ML techniques to enable automation and pre-incident prediction and detection. A thorough awareness of the setups and the various data sources and types will also greatly reduce the time necessary for DFIs. The automation of DFR processes will enable the automatic capturing and saving of digital evidence from smart environments, based on pre-defined rules, using a rule-based classifier and association rules. This should provide proactive and preventive methods for automation in such an environment.

Furthermore, DFR principles already assure the forensic soundness of the information gathered, making it appropriate for litigation. Therefore, ML techniques can be applied in the DFR process, making use of techniques such as noise-resistant algorithms, support vector machines, and neural networks. By making use of such ML techniques, investigators could deduct the rules of classification from existing and historical datasets and scenarios to learn and train the readiness model. Clustering could then be applied to improve the accuracy of classification and allow the model to make decisions by itself.

ML techniques are mainly used for prediction and classification. Therefore, the acquisitive and concurrent processes in ISO/IEC 27043:2015 can be automated using ML techniques. This will benefit the readiness and initialisation processes in this set of standards to predict and detect incidents using decision trees and neural networks.

On the other hand, the incorporation of ML techniques can be powerful for DFI, but there is also a lack of interpretability and inadequate training data, which may lead to powerless and improperly comprehended models [29].

5. Conclusions

By presenting an overview of MLF research papers from 2018 until 2023, this paper shows how ML techniques have recently been used across different areas of the DFI process in smart environments. Common challenges for DF in these environments were also highlighted. Although intelligent technologies such as ML have the potential to aid in DFI, these technologies mainly facilitate the automation of manual DFI processes. However, this paper reports that numerous research papers found that ML techniques are applied in DF in a bid to improve the efficiency of the DFI process by means of automation, which decreases investigators' manual effort and hard work. It also investigated numerous ways to highlight what to expect in the future from MLF applications. Finally, it discussed the role of ML techniques in the DFI process as advocated in ISO/IEC 27043:2015. This was done to highlight gaps that need more attention and where ML techniques can also be applied to improve the current DFI process.

In other words, the main contribution of this paper is to let the reader know what has been done in this area and where current gaps are still evident. This will help researchers not to do excessive research themselves to learn what the current gaps are. All the current gaps can then be easily identified by a researcher, and the researcher can decide which of those gaps to solve in their own future research.

Therefore, as mentioned in Section 4, the main limitations of ML are that there is a lack of interpretability and inadequate training data, which may lead to powerless and poorly comprehended models. Thus, it would be powerful to simulate results and then compare the performance of different ML techniques in a future work. In addition, according to the state of the art in applying ML techniques in DFI within smart environments (which mainly involves the initialisation and investigative processes, as shown in Table 1), ML techniques should be applied more prominently in the readiness processes, acquisitive processes, and concurrent processes of DFI.

Author Contributions: Conceptualization, L.T.; methodology, L.T. and H.V.; validation, H.V. and L.T.; formal analysis, L.T.; investiga-tion, L.T.; resources, L.T.; data curation, L.T.; writing—original draft preparation, L.T.; writing—review and editing, H.V. and L.T.; visualization, L.T.; supervision, H.V.; project administration, H.V.; funding acquisition, H.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Popescul, D.; Radu, L.D. Data Security in Smart Cities: Challenges and Solutions. *Inform. Econ.* **2016**, *20*, 29–38. [CrossRef]
2. Quick, D.; Choo, K.-K.R. Big forensic data management in heterogeneous distributed systems: Quick analysis of multimedia forensic data. *Software Pract. Exp.* **2016**, *47*, 1095–1109. [CrossRef]
3. Watson, S.; Dehghantanha, A. Digital forensics: The missing piece of the Internet of Things promise. *Comput. Fraud. Secur.* **2016**, *2016*, 5–8. [CrossRef]
4. Du, X.; Hargreaves, C.; Sheppard, J.; Anda, F.; Sayakkara, A.; Le-Khac, N.A.; Scanlon, M. SoK. In Proceedings of the 15th International Conference on Availability, Reliability and Security, Virtual, 25–28 August 2020. [CrossRef]
5. Babun, L.; Sikder, A.; Acar, A.; Uluagac, A. IoTDots: A Digital Forensics Framework for Smart Environments. *arXiv* **2022**, arXiv:1809.00745. Available online: https://arxiv.org/abs/1809.00745 (accessed on 1 March 2023).
6. Kebande, V.R.; Ikuesan, R.A.; Karie, N.M.; Alawadi, S.; Choo, K.-K.R.; Al-Dhaqm, A. Quantifying the need for supervised machine learning in conducting live forensic analysis of emergent configurations (ECO) in IoT environments. *Forensic Sci. Int. Rep.* **2020**, *2*, 100122. [CrossRef]
7. Valjarevic, A.; Venter, H.S. A Comprehensive and Harmonized Digital Forensic Investigation Process Model. *J. Forensic Sci.* **2015**, *60*, 1467–1483. [CrossRef]
8. Conti, M.; Dehghantanha, A.; Franke, K.; Watson, S. Internet of Things security and forensics: Challenges and opportunities. *Futur. Gener. Comput. Syst.* **2018**, *78*, 544–546. [CrossRef]
9. Valjarevic, A.; Venter, H.; Petrovic, R. ISO/IEC 27043:2015—Role and application. In Proceedings of the 2016 IEEE 24th Telecommunications Forum (TELFOR), Belgrade, Serbia, 22–23 November 2016; pp. 1–4. [CrossRef]
10. Tok, Y.C.; Chattopadhyay, S. Identifying threats, cybercrime and digital forensic opportunities in Smart City Infrastructure via threat modeling. *Forensic Sci. Int. Digit. Investig.* **2023**, *45*, 301540. [CrossRef]
11. Sahib, H.I.; AlSudani, M.Q.; Ali, M.H.; Abbas, H.Q.; Moorthy, K.; Adnan, M.M. Proposed intelligence systems based on digital Forensics: Review paper. *Mater. Today Proc.* **2023**, *80*, 2647–2651. [CrossRef]
12. Qadir, A.M.; Varol, A. The role of machine learning in Digital Forensics. In Proceedings of the 2020 8th International Symposium on Digital Forensics and Security (ISDFS), Beirut, Lebanon, 1–2 June 2020. [CrossRef]
13. Goni, I.; Gumpy, J.M.; Maigari, T.U.; Muhammad, M.; Saidu, A. Cybersecurity and Cyber Forensics: Machine Learning Approach. *Mach. Learn. Res.* **2020**, *5*, 46. [CrossRef]
14. Iqbal, S.; Alharbi, S.A. Advancing Automation in Digital Forensic Investigations Using Machine Learning Forensics. *Digit. Forensic Sci.* **2020**. [CrossRef]
15. Jarrett, A.; Choo, K.R. The impact of automation and artificial intelligence on digital forensics. *WIREs Forensic Sci.* **2021**, *3*, e1418. [CrossRef]
16. Du, X.; Scanlon, M. Methodology for the automated metadata-based classification of incriminating digital forensic artefacts. In Proceedings of the 14th International Conference on Availability, Reliability and Security, Canterbury, UK, 26–29 August 2019; pp. 1–8. Available online: https://bit.ly/2Oqh6u6 (accessed on 9 March 2023).
17. Krivchenkov, A.; Misnevs, B.; Pavlyuk, D. Intelligent Methods in Digital Forensics: State of the Art. In *Lecture Notes in Networks and Systems*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 274–284. [CrossRef]
18. Babun, L.; Sikder, A.; Acar, A.; Uluagac, S. The Truth Shall Set Thee Free: Enabling Practical Forensic Capabilities in Smart Environments. In Proceedings of the 2022 Network and Distributed System Security Symposium, San Diego, CA, USA, 24–28 April 2022. [CrossRef]
19. Shakeel, P.M.; Baskar, S.; Fouad, H.; Manogaran, G.; Saravanan, V.; Montenegro-Marin, C.E. Internet of things forensic data analysis using machine learning to identify roots of data scavenging. *Futur. Gener. Comput. Syst.* **2021**, *115*, 756–768. [CrossRef]
20. Adam, I.Y.; Varol, C. Intelligence in digital forensics process. In Proceedings of the 2020 8th International Symposium on Digital Forensics and Security (ISDFS), Beirut, Lebanon, 1–2 June 2020. [CrossRef]
21. Ngejane, C.; Eloff, J.; Sefara, T.; Marivate, V. Digital forensics supported by machine learning for the detection of online sexual predatory chats. *Forensic Sci. Int. Digit. Investig.* **2021**, *36*, 301109. [CrossRef]
22. Kalnoor, G.; Gowrishankar, S. IoT-based smart environment using intelligent intrusion detection system. *Soft Comput.* **2021**, *25*, 11573–11588. [CrossRef]
23. Mazhar, M.S.; Saleem, Y.; Almogren, A.; Arshad, J.; Jaffery, M.H.; Rehman, A.U.; Shafiq, M.; Hamam, H. Forensic Analysis on Internet of Things (IoT) Device Using Machine-to-Machine (M2M) Framework. *Electronics* **2022**, *11*, 1126. [CrossRef]
24. Koroniotis, N.; Moustafa, N.; Slay, J. A new Intelligent Satellite Deep Learning Network Forensic framework for smart satellite networks. *Comput. Electr. Eng.* **2022**, *99*, 107745. [CrossRef]
25. Palmese, F.; Redondi, A.E.; Cesana, M. Feature-Sniffer: Enabling IoT Forensics in OpenWrt based Wi-Fi Access Points. *arXiv* **2023**, arXiv:2302.06991. Available online: https://arxiv.org/abs/2302.06991 (accessed on 5 June 2023).
26. Salih, K.M.M.; Dabagh, N.B.I. Digital Forensic Tools: A Literature Review. *J. Educ. Sci.* **2023**, *32*, 109–124. [CrossRef]
27. Shahbazi, Z.; Byun, Y.-C. NLP-Based Digital Forensic Analysis for Online Social Network Based on System Security. *Int. J. Environ. Res. Public Health* **2022**, *19*, 7027. [CrossRef] [PubMed]
28. Ferreira, S.; Antunes, M.; Correia, M.E. A Dataset of Photos and Videos for Digital Forensics Analysis Using Machine Learning Processing. *Data* **2021**, *6*, 87. [CrossRef]

29. Balushi, Y.A.; Shaker, H.; Kumar, B. The use of&nbsp;machine learning in&nbsp;digital forensics: Review paper. In *Proceedings of the 1st International Conference on Innovation in Information Technology and Business (ICIITB 2022)*; Atlantis Press: Amsterdam, The Netherlands, 2023; pp. 96–113. [CrossRef]
30. Baig, Z.; Khan, M.A.; Mohammad, N.; Ben Brahim, G. Drone Forensics and Machine Learning: Sustaining the Investigation Process. *Sustainability* **2022**, *14*, 4861. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.