CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Digital Forensic Investigation of Automotive Systems: Requirements and Challenges

Master's thesis in Computer science and engineering

Yitao Dong
Jun Zhang

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2023

Master's thesis 2023

Digital Forensic Investigation of Automotive Systems: Requirements and Challenges

Yitao Dong
Jun Zhang

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2023

Digital Forensic Investigation of Automotive Systems: Requirements and Challenges
Yitao Dong and Jun Zhang

Supervisor: Kim Strandberg, Volvo Cars
Examiner: Tomas Olovsson, Chalmers

Master's Thesis 2023
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2023

Digital Forensic Investigation of Automotive Systems: Requirements and Challenges
Yitao Dong and Jun Zhang
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

With the increasing complexity of vehicle architecture and interconnection between the vehicle and other entities, many problems arise, such as cyber attacks. Accidents can be caused by hardware or software failures, intentional or unintentional incidents. Automotive Digital Forensics (ADF) is used for determine the cause of accidents and usually includes processes like data collection, data analysis, data preservation and documentation. ADF is a relatively new field and is not well-researched. The lack of common and unified industry guidelines and standards makes ADF challenging. We have investigated previous work within the automotive and similar areas, with the aim of identifying parts applicable to ADF, such as forensic mechanisms, guidelines, standards, fulfilment of security properties, and data extraction and verification. Furthermore, we propose a framework that considers the entire life cycle of ADF.

Keywords: Automotive digital forensics, ADF, IoT, cyber security, V2X communication, forensics guidelines, forensics model.

Acknowledgements

We would like to express the most sincere gratitude to our supervisor, Kim Strandberg. During we writing the thesis, Kim helped us a lot on our work - answering our questions, meeting with us regularly and revising the thesis. With his assistance, we were able to finish our thesis smoothly. We are also grateful to our examiner Tomas Olovsson, who has given us invaluable inspiration and encouragement to complete this meaningful work. Furthermore, many thanks to the Cyber Security team at Volvo Cars for their kindness and friendly support during the writing of the thesis.

Yitao Dong and Jun Zhang, Gothenburg, 2023-06-30

Contents

List of Figures

List of Tables

List of Acronyms

DF	Digital Forensics
ADF	Automotive Digital Forensics
V2X	Vehicle-to-Everything
IVN	In-Vehicle Network
ECU	Electronic Control Unit
EDR	Event Data Recorder
OBD-II	On-Board Diagnostics-II
JTAG	Joint Test Action Group
CAN	Controller Area Network
LIN	Local Interconnect Network
MOST	Media Oriented System Transport
GDPR	General Data Protection Regulation
VANETs	Vehicular Ad-Hoc Networks
DLC	Diagnostic Link Connector
UAVs	Unmanned Aerial Vehicles
JSON-LD	JavaScript Object Notation for Linked Data
RSU	Roadside Units
CASE	Cyber-investigation Analysis Standard Expression
TRLs	Technology Readiness Levels
UUID	Universally Unique Identifier
AL	Applicability Level
VIN	Vehicle Identification Number
VPKI	Vehicular Public Key Infrastructure

1

Introduction

According to Cambridge dictionary, the word "forensics" refers to the scientific methods of solving crimes, that involve examining objects or substances related to a crime [1]. From this definition, we conclude that the most important step of forensics is examining objects, i.e., collecting evidences, and finally use them to establish the crime timeline. During the traditional forensic process, investigators collect physical or biological evidences (e.g., fingerprints and DNA) with the aim to identify the suspect. Unlike traditional forensics, Digital Forensics (DF) emphasizes on collecting digital evidences such as the data in memory, hard drives or cloud. We mainly discuss Automotive Digital Forensics (ADF) in this thesis, which is a branch of digital forensics. This chapter introduces the background of our work, the challenges that currently exist and our purpose.

1.1 Background

In recent years, the complexity of vehicle architecture has significantly increased, as well as the interconnection between vehicles and other entities, i.e., Vehicle-to-Everything (V2X) communication. The rapid development of automotive technology increases safety and makes autonomous driving come true, but it also raises many issues, such as cyber attacks. Accidents can be caused by hardware or software failures, intentional or non-intentional incidents such as caused by a distracted driver. To determine the cause, we need to collect and analyse relevant data generated by the vehicle.

A vehicle has a lot of data originating from the In-Vehicle Network (IVN) and its communication with the outside world via V2X. Currently a very limited amount of data is stored in the memory of Electronic Control Unit (ECU)s and the cloud. For a vehicle, to support forensic investigation, much more data need to be stored securely. However, ADF is a relatively new area and is not well-researched. In a systematic literature review over the area of ADF [2], Strandberg et al. mention various challenges. For instance, the lack of common and necessary industry guidelines and standards has made ADF difficult. Furthermore, no standardized data formats or interfaces exist in vehicles, which makes data collection and extraction difficult.

1.2 Concepts

Digital forensics has been investigated by many scholars [3], [4], [5]. Still, ADF is relatively a new area compared to DF because of its characteristics and distinct challenges. To familiarize readers, we are going to introduce some concepts below.

1.2.1 Automotive digital forensics

ADF is defined as a branch of digital forensics relating to recovery of potential evidence stored in automotive modules, networks and messages sent across operating systems [6]. There is a large volume of data exchange while the vehicle is running. Today's vehicles are equipped with multiple sensors, such as GPS, cameras and radar. The ECUs handle the input from sensors, whereafter signals are sent to actuators to respond with different actions like braking, steering or acceleration.

It is crucial that ADF evidences should not be altered. Thus, integrity is imperative. In general, a "CIANP" model is presented in [2] to meet the security requirement for forensics evidence data, where the capital letter are abbreviations for Confidentiality, Integrity, Availability, Non-repudiation and Privacy. In ADF, specific steps of carrying out forensics may vary between different models, but it typically contains four steps: data collection, data analysis, data preservation and documentation. In the data collection phase, data is collected from various data sources existing in, e.g., IVN, V2X and the cloud, whereafter the data is filtered based on specific criteria concerning the investigated crime. The next step is to analyse the data to determine if it can be potential evidence. Automated machine-learning based approaches can be used to filter out forensically relevant information from the collected data [7]. Once the evidences are identified, the next phase aims at preserving evidences in a forensically sound manner, i.e., preserving the integrity of evidence [8]. In the last phase, a final report is generated from the results of the previous phases. Finally, the documentation can be used and presented in relation to a crime in the court of law.

From a forensic perspective, the "CIANP" model is interpreted as follows. Confidentiality means that data shall only be accessed by authorized entities. Integrity ensures that the data is not tampered within the forensics process. Availability means that data remains available even in the event of a crash or other unexpected situations. Non-repudiation refers to the property that the occurrence of a certain event and its origin cannot be denied [2]. User's personal data such as GPS location, call logs and contacts are highly sensitive from a privacy perspective. The privacy property ensures that the data is well protected and not disclosed to unauthorized individuals. When the security properties mentioned above are fulfilled, the data can be trusted to be authentic in a court of law.

1.2.2 Hardware and protocols

Modern vehicles can have over 150 ECUs and contain more than 100M lines of code [2]. Automotive hardware include sensors, actuators and ECUs. Software consists of

applications used to implement automobile's functionalities, which can be installed and running in ECUs. Typically, the in-vehicle software provides functionalities such as braking and interaction with the infotainment system. Memory chips in ECUs are ideal data sources for ADF. Besides ECUs, the infotainment system is of particular interest for investigators to examine. The infotainment system delivers information and entertainment via the dashboard touchscreen [9], which contains forensically useful information such as synchronized data from paired devices. Another device that contains useful data is the Event Data Recorder (EDR) [9]. In previous research about data storage investigation in vehicles and significance analysis, the EDR data was emphasized [10]. EDR, sometimes referred to as an automotive black box, is triggered by an event and used to record information related to crashes and accidents in a tamper-proof manner [8].

In order to access the data stored in the hardware for debugging and forensics purposes, interfaces are used. In the automotive case, On-Board Diagnostics-II (OBD-II), Joint Test Action Group (JTAG), USB, WiFi, Bluetooth are examples of existing interfaces. OBD-II, has been introduced to check the real-time parameter of all the electronic control units [11], and to record any abnormal behavior or malfunction in the system components [12].

In vehicles, there are mainly four kinds of communication buses: Controller Area Network (CAN), Local Interconnect Network (LIN), FlexRay and Media Oriented System Transport (MOST). Their functionalities are described as follows:

1) CAN
   The most common bus for data exchange and diagnostics.

2) LIN
   Used for low-speed and bandwidth applications, e.g., doors and sliding windows up and down.

3) FlexRay
   Used for safety critical and high-speed messages, e.g., vehicle stability control and embedded sensors.

4) MOST
   Used for high-speed and bandwidth multimedia related applications, e.g., music/video streaming and vehicle cameras.

Lacroix et al. [13] visualize the vehicle architecture and mention that CAN is most important since it is the backbone network in a vehicle. It forwards all traffic that needs to be relayed between the sub-networks, and is an essential source for ADF since it can contain relevant error messages. However, it is a low level protocol and does not support security features such as encryption. Thus, most applications use their own security mechanisms [14].

1.2.3 Regulations and standards

In general, ADF goes hand in hand with automotive cybersecurity, since securing the process of evidence data handling is of significant importance. There are several regulations and standards related to ADF both in Europe and around the world, which we will briefly introduce below.

ISO/IEC 27037:2012, provides guidelines for specific activities in the handling of digital evidence, which are identification, collection, acquisition and preservation of potential digital evidence that can be of evidential value. Digital evidence should be obtained with an acceptable method with maintained integrity. This standard also provides general collection guidelines for physical evidence which could be helpful for managing the digital evidence [15].

The ISO/SAE 21434:2021 "Road Vehicles - Cybersecurity Engineering", jointly developed by SAE and ISO, is a guideline for secure automotive software and hardware development, and a standard for cybersecurity in the automotive industry. It enables organizations to define cybersecurity policies and processes, manage cybersecurity risks, and foster a cybersecurity culture [16].

UN R.155, covers the uniform requirements of automotive cybersecurity and cybersecurity management systems. This regulation is closely mapped to the requirements laid out in ISO 21434.

General Data Protection Regulation (GDPR), applicable as of May 25th, 2018 in all member states to harmonize data privacy laws across Europe [17]. It imposes obligations onto organizations or individuals when they collect personal data in EU.

1.3 Challenges

ADF is still an immature area. Although there are some technical solutions and surveys in ADF, many challenges still exist and need to be addressed, e.g., in the following categories [2]:

- Evidence data management, referring to such as data collection, data extraction, data storage. Lack of a dedicated device to store forensic data is one of the challenges.
- Communication, referring to data transferred to the cloud, edge, fog or related to Vehicular Ad-Hoc Networks (VANETs), where bandwidth can be a problem because of the huge data volume.
- Algorithms, including machine learning. Although machine learning algorithms produce good results in data processing, they take considerable computation power.
- Software and hardware used in forensics, such as forensic tools and vehicle architecture design, sensors, EDR. However, many tools or software have their own data formats, and it is a great trouble to use multiple tools from different companies.

- Cryptography, including blockchain. Although encrypting data improves data security level, it is still a trade-off between security and time.
- General solutions, such as forensic process and guidelines, but the guidelines are not standardized.

The challenges we face in ADF are broad, and in our research the following ones are highlighted:

- When producing vehicles, many manufacturers prioritize usability and cost rather than fulfilling security properties, which increases the risk of cyber attacks.
- The increasing complexity of vehicle architecture exposes vehicles to more attacks.
- Increase in data volumes. Huge data volume could be generated both inside and outside of the vehicle such as data originating from the infotainment system and via communication with the cloud. This makes data storage, extraction and management difficult.
- Various data formats and interfaces. Lack of standardized data formats and interfaces in vehicles make the evidence data collection and analysis difficult because of the low performance.
- Security properties. The integrity and reliability of ADF data need to be ensured, i.e., the data must be authentic and tamper-proof.
- Privacy concerns. Data from cameras, recorders and GPS may involve privacy issues.

1.4 Purpose

Our ultimate goal is to provide guidelines and mechanisms for ADF. However, we put our emphasis on identifying the requirements for ADF at first, and then list focus categories corresponding to these requirements as follows:

- Forensic mechanisms. Investigate what types of mechanisms exist in current vehicles regarding forensics investigations, including a potential standardized forensic guideline or process, and consider existing gaps for a forensic model.
- Similar areas. Investigate similar areas that might apply to automotive, such as IoT, avionics, trains and smart cities.
- Security properties. For example, by what means we can ensure the authenticity and admissibility of the collected and stored data.
- Data extraction. Discuss and evaluate the challenges of transforming the collected evidence into data that a human investigator can interpret. Additionally, consider what measures may potentially influence the data (e.g., removing the power supply may result in losing volatile data).
- Data verification. Once data is extracted, investigate how to prove the data is authentic so it can be admissible in the court of law.

Therefore, we consider the following countermeasures regarding the aforementioned challenges. For instance,

- Investigate what has been done in ADF and other areas, to determine a general guideline.
- Given existing solutions, identify what is currently lacking concerning ADF.
- Propose an approach to ensure the CIANP properties.

1.5 Thesis outline

The rest of this thesis is organized as follows. We discuss related works in Chapter 2. In Chapter 3, we present the methods used when conducting the research. Following that in Chapter 4, we demonstrate the detailed steps, i.e., the results of our forensic model. In Chapter 5, we discuss various issues such as innovations, limitations and ethical concerns. In the last Chapter 6 we present conclusions and finally ends with future work.

2

Related work

Although ADF is still an immature area, some technical solutions exist. In this section, related work presented and classified into four categories, whereafter the gaps are analyzed.

1) **Guideline**

As mentioned in [3], [4], [18], one challenge is that there is no general forensics framework or guideline. Regarding this problem, Buquerin et al. propose a generalized approach for ADF [5]. Their solution comprises four steps: Forensic Readiness, Data Acquisition, Data Analysis and Documentation. In the first step, they identify available data sources, tools and data extraction techniques. Then they choose a specific tool and interface to implement the data extraction process. At the end of the second step, the data acquisition, the extracted data need to be duplicated and the original data should be stored in a tamper-proof way. All later actions are performed on the duplicated data set. In the third step, the data analysis, the most relevant data are filtered out and the investigators can establish the evidence chain and crime timeline based on the data. Finally, in the forth and last step, they document all the previous results and create a final report. An example is given at the end of the article, which proves their approach to be feasible.

Altschaffel et al. [19] propose another ADF model that has six steps: Strategic Preparation (SP), Operational Preparation (OP), Data Gathering (DG), Data Investigation (DI), Data Analysis (DA), Documentation (DO). The SP step refers to the forensic preparations done before accidents happening, while OP are forensic preparations after accidents happening. The rest steps are similar to those mentioned above. Moreover, they divide the forensic process into two categories: live forensic and post-mortem forensics. Live forensic focuses on extracting volatile data (e.g., data in main memory), while post-mortem forensics is carried out when the system is in power-off mode, which allows investigators to retrieve data on less volatile storage like hard disk.

In [20], Sharma et al. show that ADF can be performed in two ways: reactive and proactive. The reactive approach is analogous to post-mortem forensics. The proactive approach consists of five phases: Proactive Collection, Proactive Preservation, Proactive Event Detection, Proactive Analysis and Report. In the first phase, data are collected using live forensics based on volatility and priority.

Then data is preserved automatically. If there are any suspicious events detected in the third phase, they are analyzed and reported. Proactive forensics enable devices to record data prior to the accidents so that investigators are able to quickly determine the cause post-incident.

2) **Architecture**

Davi et al. [21] describe an architecture for autonomous cars using blockchain technology. Traditional blockchain ensures data is accessible by authorized third parties, but does not guarantee integrity. This approach implements an in-vehicle shared ledger architecture to ensure data integrity, where each ECU works as a miner and shares information with all other ECUs. When a transaction (safety and security related message) is made, the ECU signs and broadcasts it to all ECUs. The receiver first verifies the signature, and propose a new block if a certain threshold of transactions is reached. In case that the verification fails, the transaction will not be processed further. Finally all ECUs update their copies of the chain by appending the new block. Their approach is helpful for ADF but fails to meet the real-time requirement of a safety-critical system.

Lacroix et al. [13] introduce vehicle hardware and architecture in a comprehensive way. Firstly, the authors introduce different buses (CAN, LIN, FlexRay and MOST) and state that CAN is the core bus that links all buses together. Then infotainment system is discussed. Infotainment systems like Ford SYNC, BMW Assist, Lexus Enform are interfaces to the end users. They provide safety-related (lock/unlock the door) or entertainment (streaming services) functionalities, thus containing a lot of useful information from a forensic perspective. Next, the authors present challenges of ADF, including mobility, topology changing, unreliable channels and multi-hop communication issues. Finally, they give an example of Ford SYNC physical dumps and analyze the information it contains.

3) **Software and service**

Researchers have presented several applications used for ADF evidence data management and communication, for example in [22] and [23]. These two solutions are both based on the use of a 3-axis accelerometer together with other in-vehicle devices and modules, to track the vehicle location, detect accidents and provide the road condition updates. When the accelerometer's G-value (centre of gravity) on the three coordinate varies, it is denoted that the vehicle faces sudden change in the acceleration. A tracking system combines the smart phone application with microcontroller which embedded with an acceleration sensing module is developed in [22]. Based on the information of G-value changes, road condition detection and vehicle location tracking can be implemented. In [23], a wireless black box using MEMS accelerometer and GPS tracking system is developed for accident monitoring. Additionally, this application can send out emergency messages to appropriate recipients when accidents happen. However, both solutions suffer from a lack of security and privacy considerations, which is critical for ADF.

As investigated in [4], [13], [24], there are several forensic tools for digital forensics, such as Encase, Accessdata Forensic ToolKit, Xways Forensic, etc, but not many choices on ADF. From a previous survey in [4], there are very few ADF-specific tools, Bosch CDR and Berla iVe are two of them. The Bosch CDR Diagnostic Link Connector (DLC) Base Kit is an entry level kit which includes most components needed to retrieve EDR data directly from the DLC of many vehicles [25]. The Berla iVe Ecosystem is a collection of tools that supports investigators throughout the entire vehicle forensics process with a mobile application for identifying vehicles, a hardware kit for acquiring systems, and forensic software for analyzing data [26].

Besides the applications and tools, there are other choices, e.g., professional commercial companies which provide ADF services, like Digitpol [27] and Envista Forensics [28]. Digitpol's services for ADF include: investigating infotainment, GPS and command systems, and identifying, capturing and analysing critical evidence stored in embedded OEM systems. Envista Forensics experts in data recovery, extraction and investigation from the infotainment system.

4) **Data extraction**

A common challenge for ADF is that there is no standardized data format and interfaces, which makes data extraction and analysis difficult. Sladovic et al. state three different ways to extract data from vehicle's internal system: connecting to the OBD-II port, umbilical-to-ECU and umbilical-to-EPROM [29]. The word "umbilical" means using a cable to directly connect to a device. Connecting to the OBD-II port is a straightforward method, but it is worth noting that the data retrieved here are DPID (Data Packet Identification Number) rather than actual data. Besides DPID, logs and fault codes are also accessible via OBD-II. For data security reasons, there is a special mode and security mechanism to gain extended privileges that enables retrieving relevant data. For the second method, the investigator connects directly to the ECU with a cable. However, if not operated correctly, the ECU would alter or even wipe the data for protection purposes. Thus it is not forensically sound and data integrity is not guaranteed. Umbilical-to-EPROM requires physical disassembly of the printed circuit board (PCB). It enables investigators to retrieve raw binary data in hexadecimal format but is time-consuming.

Having discussed the related work as above, we conclude the following gaps. First of all, although several ADF frameworks or guidelines have been presented, they look similar to some extent but are still not unified. The lack of standardized guidelines in automotive industry is a problem. Secondly, no dedicated device storing forensic data is an issue. Furthermore, issues of data integrity have not been well addressed. Interfaces like WiFi and Bluetooth expose the internal system to external users as well as attackers [30]. Therefore, an approach for ensuring data integrity is needed. Finally, for data management, the obstacle of no standardized data format poses challenges for data extraction and interpretation, and can even leads to volatile data

loss.

In the following Chapter 3, we first present how we perform the literature review and the evaluation criteria. To address the before-mentioned gaps, we then introduce two relevant techniques: Blockchain and Cyber-investigation Analysis Standard Expression (CASE). For blockchain technology, since each block is linked by the hash value of its previous block, nodes can detect alteration of blocks, thus ensuring data integrity (I). Another feature of blockchain is that each node in the network keeps a copy of the main chain, and as long as more than half of the nodes are available and stay honest, the availability (A) and non-repudiation (N) properties can be guaranteed. However, ensuring "C" and "P" properties may require other approaches like encryption and data classification. CASE is an annotation language and has the ability to unify different data formats in a key-value pair format.

3

Methods

Our approach is divided into four steps as follows. First, we performed a literature review that includes 36 papers and 2 databases. Second, we analyzed digital forensic approaches in similar areas. Third, we described and evaluated the techniques used in our solution. Forth, and finally, the evaluation criteria is discussed.

3.1 Literature review

We followed the same approach as Strandberg et al. [2], where we first reviewed papers from different databases, and then perform Snowballing on these papers to gain additional papers. We first reviewed 33 papers from Google Scholar and IEEE Xplore. We used the following search strings, *automotive digital forensics*, *IoT digital forensics*, *avionics digital forensics*, *railway digital forensics* and *smart cities digital forensics*. By performing Backward and Forward Snowballing, we then obtained 3 additional papers closely related to our research.

Most of these papers are technical solutions on ADF or similar areas, and some of them are surveys. They are later evaluated using Technology Readiness Levels (TRLs) at the end of this chapter. TRLs are a method for estimating the maturity of technologies during the acquisition phase of a programme. It is based on a scale from 1 to 9, where 9 is the most mature technology, defined as follows [31]:

- TRL1: Basic research
- TRL2: Technology concept is formulated
- TRL3: Experimental proof of concept
- TRL4: Technology confirmed in lab
- TRL5: Technology validated in relevant environment
- TRL6: The technology demonstrated in relevant environments
- TRL7: System prototype demonstrated in operational environment
- TRL8: System complete and confirmed
- TRL9: The system proven in an operational environment

3.2 Similar areas

Digital forensics is not limited to automotive, there are other similar areas such as IoT, avionics, railway and smart cities, that have also taken digital forensics into consideration. In this section, we introduce similar areas and compare them to the automotive domain, with the aim to find their possible applicability for ADF.

3.2.1 IoT

DF that is a research hotspot in IoT, also faces multiple challenges. IoT devices usually are connected and communicate with other entities. The devices are mostly heterogeneous, i.e., have diverse data sources, and results in various file systems and interfaces. Additionally, the devices can communicate using different communication protocols, such as HTTP, Bluetooth and NFC. Furthermore, the topology of IoT systems may be dynamic because the entities may move. All these characteristics make DF in IoT challenging.

The IoT stores and transmits massive amounts of data between different devices. IoT DF can gather evidence from a variety of sources, such as sensors, communication devices, drones, smart home devices, cloud storage and vehicles. Thus, multiple DF approaches are required. The challenges of IoT DF include the increasing number of forensics entities, identify its relevance, blurry or non-existed network boundaries. As investigated in [32], although there are already many technical solutions that almost involve all aspects of IoT DF, it is still a field that needs hard work. Blockchain technology is considered suitable for IoT DF due to its immutable and distributed characteristics. As surveyed in [33], [34], [35] and [36], where a number of blockchain-based solutions have been presented.

IoT relates to the automotive area not only because of the increasing use of IoT devices in vehicles [37], but also similar features between them such as being mobile and distributed. IoT-based DF analysis may lay a foundation for the forensic soundness and reliability of digital forensic processes in automotive systems [38].

3.2.2 Avionics

In avionics, there is a device called flight recorder, often referred to as the name of "black box". There are two types of flight recorder, the flight data recorder (FDR) and the cockpit voice recorder (CVR). The outer casing of the flight recorder is designed to withstand transient damage from harsh environments, is made of special material and is painted bright orange. It is installed in the safest position of the avionics. The flight recorder can record various parameters during the flight, such as flight time, speed, altitude, temperature, and even the dialogue between the pilot and the crew. In the event of an aircraft accident, finding the flight recorder and reading out the recorded data can help investigators performing digital forensic to determine the cause of the accident.

Flight recorder technology in avionics is a mature field with a history dating back to 1950s. In recent years, there are some emerging techniques in avionics digital forensic also interested us, such as replacing the FDR and CVR with a system that provides aircraft data monitoring to support tracking of aircraft and data archiving technology from a space based platform in [39], and digital forensic on new generation aircraft [40]. With the development of technology, the flight recorder is no longer limited to be used in avionics, but can be widely used in variety of fields, such as missiles, rockets, trains, as well as within automotive digital forensics. An estimated price for a flight recorder is around $60,000, which is reasonable for an aircraft but expensive for a vehicle. So from a technological point of view, the flight recorder is suitable for automotive, but from an economic point of view it is not so ideal.

3.2.3 Railway

Railway is considered as a closed safety system and runs in a non-networked environment. Because of these characteristics of railway systems, there don't seem to be many serious attacks to analyse from a DF perspective. As a result, DF analysis of the railway system has not received much attention. Like other fields, railway is on the road of digitization and become increasingly connected to the Internet. Thus, there is also a risk for cyber attacks on the railway systems, potentially with disastrous outcomes.

The above indicates that DF in railway systems is imperative. Still, not much work has been done. J. Cosic et al. have published two papers outlining the challenges and the investigation process in railway DF [41], [42], respectively. No other specific technical solutions have been presented. In addition, railway system is complicated and distinct, consists of multiple components with particular functions. Furthermore, railway is a centralized system with central control in comparison vehicles are distributed and communicate via V2X. As a result of existing work within the railway system, it seems to have very limited value for ADF.

3.2.4 Smart cities

Smart cities is emerging and refers to a technologically modern urban area that uses different types of electronic methods and sensors to collect specific data [43]. The U.S. National Institute of Standards and Technology (NIST) proposed a model for smart cities that comprises six components: government, economy, mobility, environment, living, and people [44] [45]. Baig et al. further divide smart cities into four categories: Smart Grids, Building Automation Systems (BAS), Unmanned Aerial Vehicles (UAVs) and Smart Vehicles, corresponding to environments, living, mobility and mobility. Smart grids gather and analyze the data about energy consumption pattern and provide a more flexible power supply. BAS controls the devices inside buildings and provide services like heating, ventilation and air conditioning. UAVs or drones have wide applications such as package delivery and coastline patrol. Future Smart vehicles will have more entertainment and diagnostic functionalities.

Although smart cities have greatly improved life quality, there are vulnerabilities that can be exploited by attackers. For example, smart grids make use of thousands of smart meters to collect power usage data and upload them to the cloud for storage and analysis purpose, which opens the possibility of attacking. To mitigate the consequences of such events, DF has become an important topic in smart cities, and it has a lot in common with ADF. First of all, smart cities and vehicles have similar topology in the sense that everything is distributed and interconnected. Therefore, when performing forensics, investigators may acquire data from multiple devices instead of one particular device. Additionally, nodes status is changing, i.e., nodes are constantly joining or leaving the network, making it hard to determine the system's state. Lastly, smart vehicles are within the scope of smart cities, thus DF for smart cities and ADF have similar characteristics.

3.3 Techniques

In this section, we are going to introduce the specific techniques used in our solution. Blockchain is commonly referred to within the field of ADF and is highlighted as a promising approach due to its characteristics. CASE is a community-developed specification language that aims to advance the exchange of cyber-investigation information between tools and organizations [46]. We consider blockchain and CASE as the fundamental techniques in our solution.

3.3.1 Blockchain

When people are shopping online, traditional payment system typically involves three parties: the buyer, the seller and a trusted third party (e.g., a bank). Although the third party solves trust issues, it introduces extra cost for both buyers and sellers. The emergence of blockchain technology is to ensure secure payments without a trusted third party. A blockchain is a distributed ledger with growing lists of records (blocks) that are securely linked together via cryptographic hashes [47], where each block contains several transactions information, and each node in the distributed system keeps a copy of the main chain. Due to providing traceability in such an approach, it is also ideal for storing forensic information.

According to the bitcoin white paper [48], the blockchain system has three core components: a timestamp server, a proof-of-work system and an incentive system. It implements an implicit timestamp server by taking a hash of the previous block. The timestamp server proves that the data must have existed at the time in order to get into the hash [48]. When there are new transactions needed to be recorded, all nodes start working at the same time to find a new block. Thus, a proof-of-work system is necessary to determine which node finishes the work first. One of the approaches is to scan for a hash value that fulfills a particular pattern. The first node who has found a new block will receive a certain amount of currency as a reward. Such incentive system encourages nodes to keep working on finding next block.

The blockchain technology can be applied to ADF. As mentioned above, establishing

an event timeline is a crucial step in forensic investigations. The timestamp server plays an important role since it can record information chronologically, which ensures non-repudiation property. The proof-of-work system is helpful in proving that a node has done the work and thus ensuring authenticity. Every vehicle is a node in the blockchain network in our solution. For the incentive system, it motivates the nodes to keep storing forensic information since the sensors on vehicles produce data all the time. Another feature of blockchain is that it ensures data integrity. The whole system is not considered compromised as long as at least 50% of the nodes are honest because of its one-CPU-one-vote nature [48]. The process of computing hash is comparable to voting. In traditional one-IP-address-one-vote mode, an attacker can fake many IP addresses to have multiple votes. For one-CPU-one-vote mode, however, it is very unlikely that a limited number of individuals take control of over 50% of the nodes. Moreover, blockchain is based on distributed systems. All the running vehicles form a distributed system, where each vehicle runs as a separate node. Thus, blockchain technology is well suited for ADF.

3.3.2 Cyber-investigation Analysis Standard Expression

Evidence data for ADF can be collected from various data sources, such as components in the vehicle, entities communicating over V2X, and the cloud. The heterogeneity of devices can lead to different data formats. Similar devices, for example, ECUs commonly runs different operating systems and uses different data formats [2]. Furthermore, data formats also vary across numerous vehicle brands and models. The lack of a standardized data format is a critical challenge for ADF. For example, there are two types of message on the CAN network. The standard one supports a length of 11 bits for the CAN identifier and the extended one supports a length of 29 bits for the CAN identifier, as shown in Fig 3.1 and Fig 3.2 [49]. LIN message frame consists of a header and a response as shown in Fig 3.3 [50].

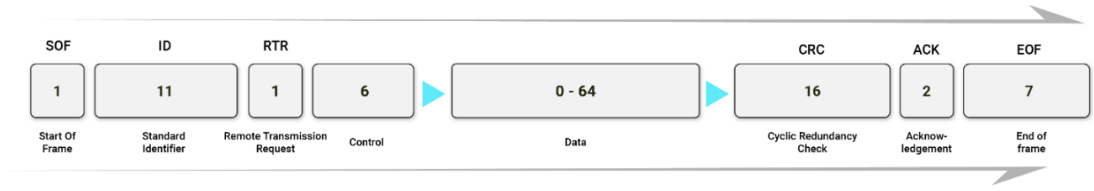

Figure 3.1: Standard CAN message frame

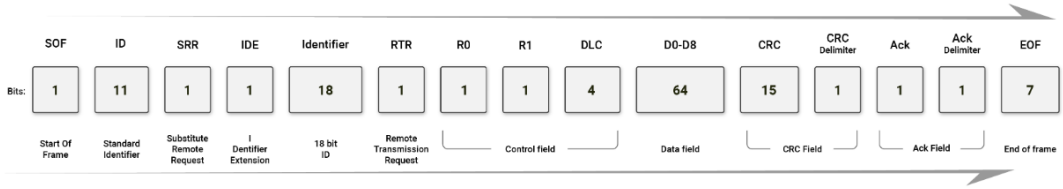

Figure 3.2: Extended CAN message frame

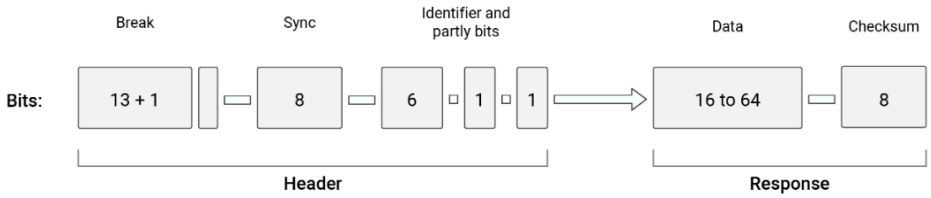

Figure 3.3: LIN message frame

There are solutions that proposes a common format. One typical example is the Navigation Data Standards (NDS). NDS is a standardized format for automotive navigation databases that aims to develop a standardized binary database format for the navigation data exchange between different systems [51]. NDS solves part of the problems, but a more general solution is still needed. From a forensic perspective, no matter where the data is extracted from or what protocol is used, a common and standardized data format is essential for the implementation of the ADF process.

As a community-developed specification language, CASE utilizes JavaScript Object Notation for Linked Data (JSON-LD) to serialize forensic information. A complete CASE representation consists of multiple *Objects*, where each *Object* is a collection of key-value pairs. It provides users several types of objects such as Identities, Relationship, Action, Investigation, Roles, Traces, Location, Annotations and Tools. Objects are uniquely identified by a 128-bit label called Universally Unique Identifier (UUID). List 1 illustrates how CASE represents information. It contains a propertyBundle which defines it as a computer. The first *"@type"* implies that it is a *"Trace"* type of object. In the *"propertyBundle"*, each *"@type"* and its associated information reveals the *"Device"* in more detail. This annotation gives a complete picture of the *Device*.

During the forensic process, a large volume of data is usually gathered, analyzed and preserved from different devices and stakeholders. As previously highlighted, having a general approach to represent information is necessary. CASE is a common format intended for expressing and exchanging cyber-investigation information [46], and digital forensics is one of the specific interest domain for CASE. Therefore, we consider that CASE is suitable for ADF investigations due to its high flexibility, usability and semantics.

```
{
  "@context": {
    "@vocab": "http://case.example.org/core#",
    "olo": "http://purl.org/ontology/olo/core#",
    "acme": "http://custompb.acme.org/core#"
  },
  "@graph": [
    {
      "@id": "forensic_lab_computer1",
      "@type": "Trace",
      "location": "forensic_lab1",
      "propertyBundle": [
        {
          "@type": "Device",
          "manufacturer": "Dell",
          "model": "Inspiron 5000",
          "serialNumber": "D1234567"
        },
        {
          "@type": "OperatingSystem",
          "name": "Windows 7 Ultimate Edition",
          "manufacturer": "Microsoft",
          "version": "6.1.7601 Service Pack 1 Build 7601"
        },
        {
          "@type": "ComputerSpecifications",
          "bios": "E1762IMS.10M",
          "cpu": "Intel Pentium i7",
          "ram": "4GB"
        },
        {
          "@type": "NetworkLocation",
          "domain": "dfl.local",
          "ipAddress": "192.168.1.145"
        },
        {
          "@type": "acme:InventoryComputer",
          "name": "DFL-03",
          "inventoryNumber": "10503"
        }
      ]
    }
  ]
}
```

Listing 1: A Device represented using CASE

3.4 Evaluation

Our intention is to analyse the usability of the CASE format aligned with blockchain technology. In order to do this, in the following subsections, we first use the TRLs to evaluate them and then carry out a gap analysis.

3.4.1 TRLs of technical solutions

We go through the technologies presented in ADF and similar areas mentioned above and evaluate them using TRLs. The results are divided into five main categories based on the areas of Automotive, IoT, Avionics, Railway and Smart Cities. For each technical solution, we identify its readiness as shown in Table 3.1. In the Applicability Level (AL) column, ● denotes applicable, ◑ partially applicable, ○ not applicable.

3.4.2 Gaps analysis

Although these techniques have solved many forensic problems, gaps still exist. For blockchain, one of the most important issues is timing and spacing overhead. The process of finding a new block involves computing hashes of a particular pattern, thus consuming considerable computation power and time. Similarly, it may require excessive storage space, because each node in the blockchain network is responsible for keeping a copy of the main chain. Therefore, timing and spacing overhead has become a main challenge for blockchain used in ADF. Moreover, an incentive/punishment mechanism is needed to ensure security [8], which is the motivation for vehicles in the networks to join and contribute to the blockchain. The blockchain technology described in [48] has an incentive system based on giving digital currency to participants. However, it does not work in ADF since there is no currency involved at all. In our case, it is an implicit incentive mechanism in the sense of "one for all, all for one". Each node records data for the benefit of all nodes. If the node itself is involved in an incident, it can also be quickly served. Thus, having an incentive mechanism can motivate vehicles to engage in forensic investigations.

Using CASE to represent information benefits forensic investigations. Firstly, different data being expressed in a unified format greatly reduces forensic complexity. For example, multi-jurisdiction is a common problem encountered in cross-border forensics. A single file can be even broken down into multiple blocks storing in different locations with different regulations [32]. In such case, a unified data format will improve cooperative investigation efficiency. Secondly, CASE supports various types of *Objects*. An investigation, a device, a file or even a location can be expressed by CASE. Such feature enables a detailed representation of information. Furthermore, each object is identified by a unique UUID, making it easier to refer other objects. However, no method exist that automatically converts other data formats into CASE.

Table 3.1: TRLs of technique solutions

Area	Ref.	Technology	TRLs									AL
			1	2	3	4	5	6	7	8	9	
Automotive	[2]	Survey	●									●
	[10]	Hardware	●									●
	[8]	Blockchain		●								●
	[9]	Introduction	●									●
	[52]	Framework					●					●
	[53]	Data Extraction					●					●
	[13]	Data Extraction			●							●
	[30]	Framework	●									●
	[21]	Framework		●								●
	[20]	Review		●								●
	[19]	Survey		●								●
	[5]	Framework					●					●
	[29]	Data Extraction		●								●
	[18]	Data Analysis			●							●
	[3]	Introduction	●									●
	[4]	Data Analysis			●							●
	[54]	Framework		●								●
	[7]	Data Extraction				●						●
	[23]	Software					●					●
	[22]	Software			●							●
	[24]	Framework			●							●
IoT	[32]	Survey	●									◑
	[38]	Survey	●									◑
	[37]	Survey	●									●
	[36]	Blockchain			●							◑
	[34]	Blockchain		●								◑
	[33]	Blockchain		●								◑
Avionics	[39]	System		●								◑
	[40]	System		●								◑
Railway	[41]	Framework		●								○
	[42]	Introduction	●									○
Smart Cities	[44]	Introduction	●									○
	[45]	Framework			●							◑

4

Results

In this chapter we describe the results in detail. The whole framework is introduced first with a table and a figure that illustrates the underlying mechanism, followed by each specific steps. At the end, a results document is designed to record all the relevant data and help the reader better understand how each procedure works.

4.1 Forensic lifecycle

Forensic lifecycle refers to the procedures required for the completion of a forensic investigation. We propose six steps, namely 1) *Data Collection (DC)*, 2) *Data Preprocessing (DP)*, 3) *Data Preservation (DV)*, 4) *Data Retrieval (DR)*, 5) *Data Analysis (DA)* and 6) *Documentation (DO)*. Below we give an overview of what should be done in each step.

Any investigation starts with the collection of information that aids in solving a case. For ADF, data can exist in vehicle, be transferred V2X, or stored in various cloud sources. In the vehicle, sensors are vital components and can generate a wide range of data. For example, the GPS module records locations of the vehicle, which is of great importance for forensic investigations. Other data sources such as ECUs and EDRs may also contain relevant data. The task of step (1) is therefore to collect data from various data sources, whereafter the data is filtered to identify the most relevant data, which is later uploaded to the cloud. Step (2) involves data preprocessing, including data classification and format unification. Data are classified into several categories according to different rules and then unified in CASE format. This allows investigators to perform ADF within a particular category. In step (3), the reduced and well-organized data set is stored in the cloud as a chunk with a unique ID, which is a concatenation of the timestamp and the hash of Vehicle Identification Number (VIN). In this scenario, the chunk refers to the pieces of data stored in the cloud, and is used to distinguish it from the block stored in the vehicle. The vehicles will then receive the corresponding IDs from the cloud, and the IDs are later used for data retrieval. As we clarified previously, each vehicle is a node in the blockchain network and can propose or erase blocks. The ID information and the hash of the previous block form a new block that will be proposed to the blockchain. The architecture of the vehicle blockchain network is shown in Fig 4.1. During step (4), if after a period of time, there are no accidents involving a specific vehicle, the blocks and chunks containing its data will be erased because of storage

limitation. Otherwise, the data should be retrieved and verified for further actions, e.g., subsequent analysis in step (5). Finally, in step (6), all steps and analysis results are documented. Moreover, each step is further divided into several sub-steps. Table 4.1 summarizes the tasks of each step and Figure 4.2 visualizes the whole process in detail.

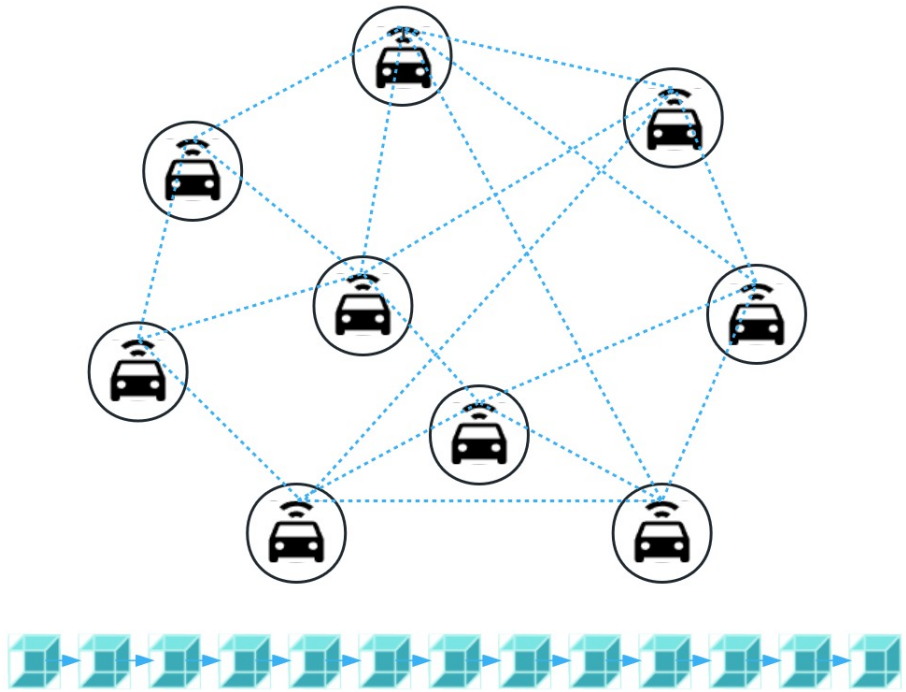

Figure 4.1: Vehicle blockchain network architecture

4.2 Data collection

The goal of this step is to collect all data pertaining to a specific vehicle and store them in the cloud. During the entire process, however, many factors can influence the success of data collection, such as stakeholders, data sources and tools. We mainly need to address two questions in *Data collection* phase: where to collect and how to collect. It comprises three sub-steps: *Preparations*, *Feasibility analysis* and *Gathering*. Firstly, identify the factors such as data sources, tools and interfaces in *Preparations*. Then determine if it is feasible to extract data in this context. Finally gathering all the data if feasible and upload to the cloud. We assume that the cloud is secure and cannot be compromised.

4.2.1 Preparations

Preparations is the first sub-step of *Data collection*, where factors that may influence data collection are identified. Such factors include stakeholders, data sources, tools and interfaces.

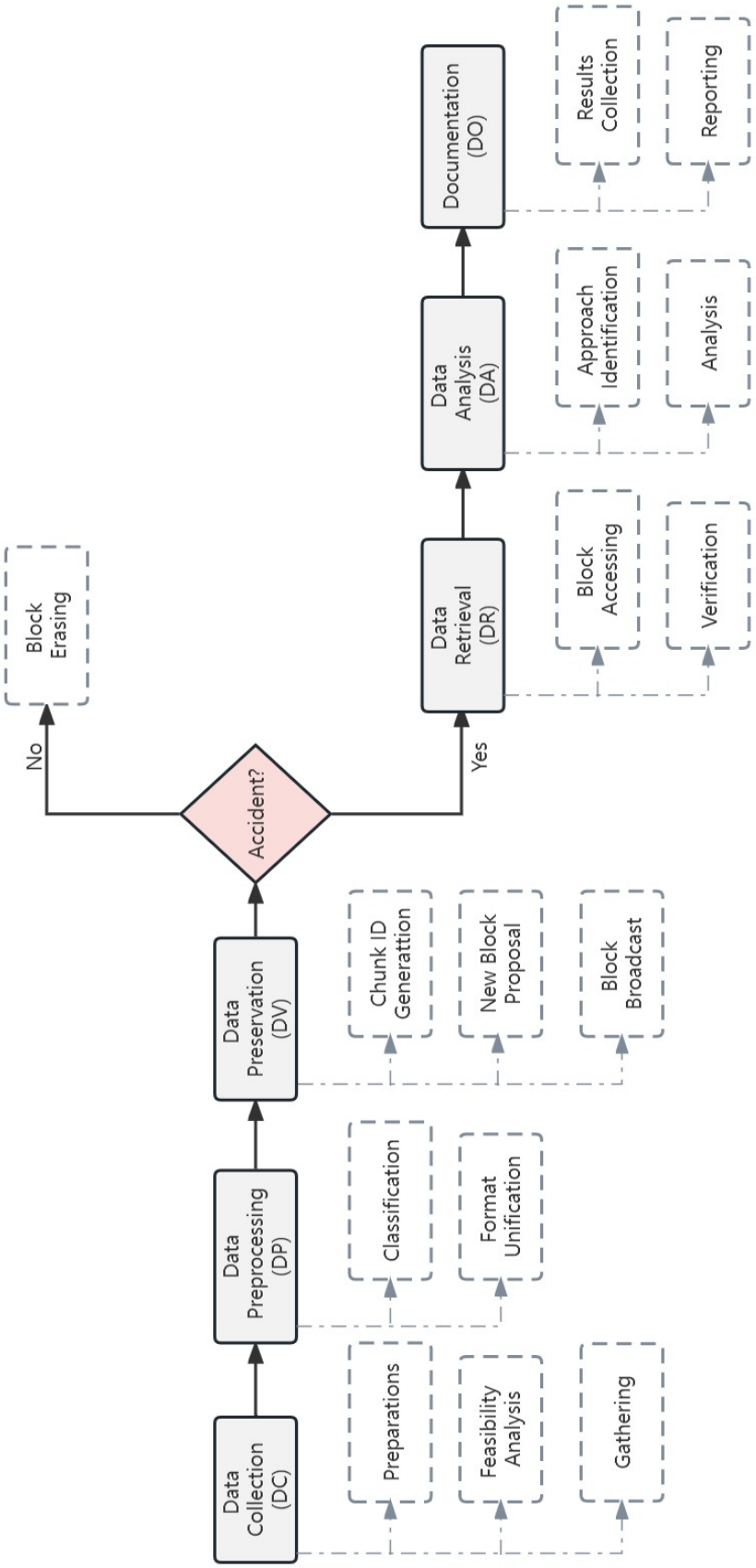

Figure 4.2: ADF framework

Table 4.1: ADF lifecycle

Phase	Description
Data Collection (DC)	Collect data from various data sources. Once the data collected, they are filtered to reduce the volume, and only the most relevant data are kept and uploaded to the cloud.
Data Preprocessing (DP)	Including data classification and format unification. Data are classified into several categories according to different strategies and then unified in CASE format.
Data Preservation (DV)	The data set is stored in the cloud as a chunk with a unique ID, which is a concatenation of the timestamp and the hash of VIN. The ID is then sent back to the corresponding vehicle and later will be used for data retrieval. The ID information and the hash of the previous block form a new block that will be proposed to the blockchain.
Data Retrieval (DR)	Two possibilities exist in DR. Blocks are erased to save space if no accidents occur after a period. Data will be read out from the cloud when accidents happen according to the chunk ID and verified for future use.
Data Analysis (DA)	The task of this step is to take the data as input, analyze them and generate an output (analysis result). It starts with identifying an approach, followed by analyzing the data using the approach. Then create a timeline for this accident based on the analysis result.
Documentation (DO)	All steps and analysis results are documented.

Stakeholders Stakeholders are the intended audiences of the collected data, and their interests in different types of data can vary. For example, law enforcement agencies prioritize data that can be used as evidence in criminal investigations, such as GPS locations, video or audio recordings. Insurance companies, on the other hand, pay more attention to EDR data, which provides critical information for determining liability in accidents, such as brake, airbag and seatbelt status, as well as vehicle speed at the time of collision. EDR data can help insurers make informed decisions when compensating accident victims. For manufacturers, regular vehicle diagnoses are of great importance, as the diagnostic reports provide valuable insights for improving the vehicle's performance and safety. On-vehicle service (e.g., navigation system) providers are primarily interested in user data to enhance user experience. Drivers themselves care more about dashboard data, such as fuel indicators, speedometers, and odometers, to monitor their vehicle's performance. By identifying stakeholders and understanding their data needs, we can make better decisions about what data to collect and analyze, ultimately leading to better outcomes for all parties involved.

Data sources The underlying mechanism of a vehicle can be summarized as follows: the ECUs continuously collect data from various sensors, analyze them,

and send instructions to actuators to control the vehicle. In case of a trigger event, such as a sudden change in speed, the EDR records the vehicle's status prior to the event. Additionally, drivers can interact with the infotainment system to access valuable information from the internet. Throughout the entire process, sensors contain environmental data, while ECUs contain controlling data. EDRs are a source of safety-critical data, while user-generated data are stored in the infotainment system. Moreover, other vehicles, Roadside Units (RSU), the cloud, and mobile devices are all valid data sources due to V2X communication. We can significantly reduce complexity by identifying the appropriate data source based on the data type we require.

Tools Data can reside in vehicle storage or internet. Prevalent forensic tools such as Forensic Toolkit (FTK), EnCase and autopsy extract the on-vehicle data by scanning the hard drive. Other tools like Wireshark can be used to monitor the data from internet. However, tools have limitations with respect to file systems. Various file systems exist in vehicle modules due to the differences in brands, OEMs and the models. For example, FAT, FAT32 and NTFS file systems in Windows embedded OS, QNX4 and QNX6 in QNX OS, HRFS and DosFS in VxWorks. QNX is considered as the most widely used file system in automotive industry for its helpful in building a safety, security and reliable automotive system, currently being used in more than 215 million vehicles, whereas most of the existing tools do not support the QNX file system. In general, forensic tools have their own applicable file systems, e.g., Table 4.2 from [4] shows the supported file systems of three forensic tools. Therefore, it is necessary to verify the suitability firstly and then choose an appropriate tool for data collection.

Table 4.2: Tools and support file systems

Support file system	**Encase 8.x**	**Access forensics**	**X-Ways 20.x**
FAT 12/16/32	●	●	●
NTFS	●	●	●
EXT2/3/4	●	●	●
HFS+	●	●	●
UFS 1/2	●		●
QNX			●

Interfaces As mentioned earlier, vehicles have multiple interfaces, such as OBD-II, JTAG, USB, WiFi, and Bluetooth. It is important to choose the appropriate interfaces based on the specific needs. OBD-II is primarily used to access diagnostic data, while JTAG is used to access test data. If we want to acquire external data, WiFi or Bluetooth may be better options. For most scenarios, a combination usage of multiple interfaces will be concerned to meet the forensics requirements.

One approach to identify factors is to focus on specific scenarios. For example, if a crash occurs on a BMW vehicle, we can identify the stakeholders, data sources, tools, and interfaces as follows. Firstly, the traffic police may investigate the cause

of the accident, and the insurance company may need to provide compensation. The driver, traffic police, and insurance company are all stakeholders. Secondly, the accident could have been caused by a variety of factors, such as driver's misoperations, environmental factors like fog or slippery roads, or vehicle system failure. In this case, the EDR is the most important data source to consider, with the ECU being an auxiliary data source. Since the BMW vehicle uses the QNX operating system, the X-Ways Forensics tool is a good candidate because it supports QNX [55]. Additionally, the OBD-II interface may be useful in extracting data. Then we can move on to the next sub-step.

4.2.2 Feasibility analysis

Feasibility analysis is an important step in the sense that we might abort the forensic process if it has too much time and economic costs. We will investigate this issue from three aspects: forensic difficulties, time cost and economic cost.

As discussed in Chapter 2, connecting to OBD-II, connecting to ECU and physical chip disassembly are three common methods to extract data from vehicles. They differ in terms of forensic difficulties and costs. Extracting data via OBD-II port only requires a cable, a customized device and a customized software from the manufacturer. The OBD-II port is standardized and each pin has a distinct meaning, which enables fast and accurate diagnoses. Extracting data directly from ECU can be more challenging since it may require professional software to dump the entire file system from the ECU. An article [56] has compared three forensic software: Forensic Toolkit (FTK), EnCase and X-ways in price and performance, where performance is expressed by the time spending on searching a specific string, as shown in Table 4.3. For individual researchers, X-ways is a better choice.

Table 4.3: Comparison of three forensic tools

Cost	Encase	FTK	X-ways
Price (per year)	$3500	$3000	$1000
Time for string searching	3m31s	3h56m3s	1m52s

Physically chip disassembly also requires hardware experts besides a professional software. It has the highest costs but is the most comprehensive method of extracting data. Table 4.4 summarizes the characteristics of each approach and their corresponding costs.

Table 4.4: Characteristics and costs of extraction methods

Evaluation	Connecting to OBD-II	Connecting to ECU	Physical chip disassembly
Forensic difficulties	Easy	Middle	Hard
Time cost	Low	Depends on the tools	High
Economic cost	Low	High	High
Summarize	Accurate	Challenging	Comprehensive

4.2.3 Gathering

In the *Preparations* sub-step we have addressed the question of where to collect the data by identifying the data sources, and in this *Gathering* sub-step, we aim to address the question of how to gather the data.

Various data sources make data gathering challenging, not only because of the large number of data sources related to large number of devices and entities, but also because of the heterogeneity of these entities. Data should be gathered and uploaded to the cloud in a forensically sound manner to ensure the integrity and confidentiality. To meet the security properties, the cloud is managed by a trusted third party, and the data are updated and stored the data in a cryptographic way.

In general, there are several levels of data extraction strategies in ADF: Network Level, Board Level and Chip Level [53]. Due to its high usability and ability to be performed automatically without damaging vehicle modules, Network Level extraction is the most commonly used method and is usually sufficient in most scenarios. As our approach is based on blockchain implemented in a distributed V2X network, we primarily focus on Network Level data extraction. This method uses specific manufacturer software to gather data from the multiple modules of the IVN. The ECUs continuously collect data from various sensors, and the infotainment system stores the user-vehicle interaction information.

This data are huge in volume and needs to be filtered first and then uploaded to the cloud. After filtering, the data that is not relevant to forensics is filtered out. This reduces the volume of the data, resulting in lower storage costs and higher performance. The gathered data stored in the cloud is preprocessed in the next sub-step. To implement data gathering, we can utilize software to listen on CAN buses and other communication channels.

4.3 Data preprocessing

Finding valuable information from the raw data is a rather difficult task because they are uncategorized and chaotic. Unifying data formats and reducing data complexity have become essential steps. We tackle this problem with two sub-steps: *Classification* and *Format unification*, which address complexity and data formats issues respectively.

4.3.1 Classification

Data classification is a process that divides a data set into different categories based on a standard. In our work, three classification standards are identified: data sources, confidential levels, and the contents, as shown in Figure 4.3.

1) Firstly, data can originate from IVN or V2X. However, this data source based method is imprecise and requires further sub-categorization. In IVN, sensors, EDR and infotainment are all possible data sources. V2X contains sub-categories like V2D (Vehicle-to-Device), V2G (Vehicle-to-Grid) and V2N (Vehicle-to-Network).

2) Confidentiality level is another standard for data classification. There are four levels in general: Restricted, Confidential, Internal and Public. This classification method helps the investigators determine the sensitivity level of data. The data with the highest confidentiality level play the most important role, and special attention should be paid to the protection of such data.

3) Since the obtained data are all forensics-related data after filtering, the data can be classified into three categories in terms of contents: safety-related data, security-related data and personal data. Safety-related data primarily involves safety incidents caused by, for example, hardware or software failure or driver's misoperations. The data falling in this category include the accelerator, the status of the safety belt and the speed. Security-related data are the information related to the security events or modules of the vehicle, e.g., the malicious code implanted in the vehicle, the remote manipulation of the vehicle, or the vehicle's Intrusion Detection System (IDS). Personal data are the user data that involves the interaction between the stakeholders and the vehicle, such as the data originated from the smartphone.

A supervised machine-learning approach is a good option to perform data classification. Supervised machine learning uses labeled datasets to train algorithms to predict results. In the context of ADF, this approach works as follows.

1) Prepare the dataset. The filtered data from the previous sub-step is the dataset in this phase.

2) Split the dataset into training and testing sets, for example, 10% training and 90% testing set using *train_test_split()* function in sklearn. The training set is annotated with the labels of *safety*, *security*, and *personal*, corresponding to the three categories based on the contents.

3) Train the model using several classification algorithms such as Random Forest, Support Vector Machines, K-Nearest Neighbours and Neural Networks. These algorithms will result in different accuracy, and the one with the highest accuracy will be selected.

4) Predict the results using the fitted model with the sklearn's *predict()* function, and the result is the classification with the three categories, i.e., *safety*, *security*, and *personal*.

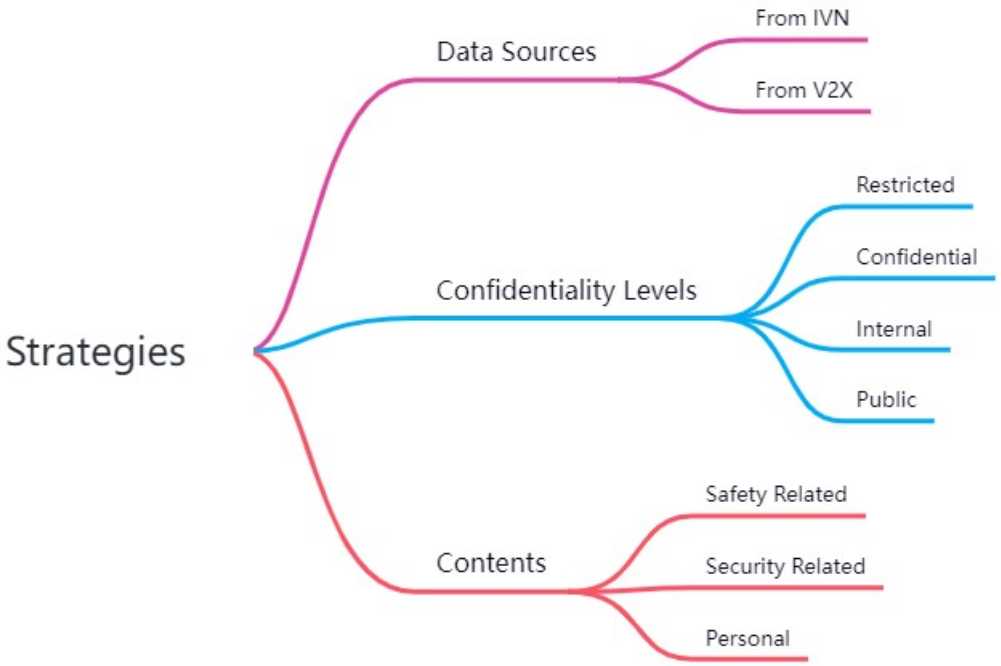

Figure 4.3: Data classify strategies

After data classification, the data are presented in a clearer, more structured way and is easier to access, giving the investigators a better understanding of the data. It is also easier to identify the most relevant data about a specific incident. For instance, if an incident is caused by an attacker remotely controlling the car, more attention should be paid to the ECU category data. Therefore, a proper data classification can greatly improve efficiency of forensics. In addition, classification can protect the confidential and sensitive data, which makes the data more secure.

The filtered and classified data will then be unified in CASE format in the next sub-step.

4.3.2 Format unification

The last sub-step aims at standardizing the data formats. Due to the variety and quantity of on-vehicle devices, many problems can arise if there lacks an exchangeable data format. Firstly, data interpretation has become an intractable issue. For instance, service providers and manufacturers have their own data formats, and some insurance companies require drivers to carry a plug-in device that records their driving behavior [13], which has introduced additional data to be dealt with. In order to interpret the diverse data formats, specialized software are needed, leading to a slower forensic process. On the other hand, modern forensics is usually not accomplished by a single person or authority, but a collaborative activity. For example, cross-border investigation refers to the investigation that involves multiple parties

(e.g., countries, jurisdictions). A unified data format is of particular importance here, as files may need to be reconstructed from multiple pieces held by different parties. In our work, CASE is applied to achieve this goal.

Object and property bundle are two fundamental components of CASE, where object depicts an entity and property bundle contains the corresponding properties. Table 4.5 lists all types of objects and their descriptions [46].

Table 4.5: CASE objects and descriptions

Object	**Description**
Investigation	An exploration of the facts involved in a cyber-relevant set of suspicious activities
PropertyBundle	A group of properties characterizing a particular aspect of an object
Identity	Characterization of the identifying properties of an individual or organization
Location	A geophysical place, site or position
Tool	Characteristics of a tool used in a cyber context
Relationship	An association or link between two objects
Annotation	A statement asserted to be true in relation to one or more other objects
Action	Something that may be done or performed within the digital domain
Trace	A distinct article or unit within the digital domain
ProvenanceRecord	A provenantial connection between a forensic action and a set of observations

A simple example is presented to show the usage of CASE. Suppose researchers obtain the following data in an investigation of a car crash:

On May 6, 2020, at 14:32 pm, Bob was driving a BMW X1 from Stockholm to Oslo. However, at 16:05 pm, the car crashed in Örebro at the geographic coordinates of 59° N, 15° E. The video captured by a Garmin dash camera revealed that the road conditions and visibility were good at the time of the crash. A recent maintenance report indicated that all functions of the car were normal. However, the EDR data showed the car was traveling at a speed of 120 km/h and in 5th gear prior to the crash. Based on this information, a preliminary conclusion has been made that the accident was caused by speeding.

In this scenario, various types of data, devices, stakeholders and judgements are involved, but they can easily be unified with CASE as follows:

```
{
  "@id": "investigation-42dec83a-ec19-11ed-a05b-0242ac120003",
  "@type": "Investigation",
  "name": "Accident X",
  "focus": "Crash",
  "description": "In Örebro, Bob's BMW car crashed",
  "object": ["Bob-uuid","car-uuid","dashcam-uuid",
  "maintenance-report-uuid","EDR-uuid","location-uuid",
  "forensic-action1-uuid","annotation1-uuid"]
```

```
},
{
  "@id": "Bob-uuid",
  "@type": "Identity",
  "propertyBundle": [
    {
      "@type": "SimpleName",
      "firstName": "Bob",
      "lastName": "Smith"
    },
    {
      "@type": "BirthInformation",
      "birthdate": "2000-01-01"
    }
  ]
},
{
  "@id": "car-uuid",
  "@type": "Trace",
  "propertyBundle": [
    {
      "@type": "Car",
      "brand": "BMW",
      "model": "X1"
    }
  ]
},
{
  "@id": "dashcam-uuid",
  "@type": "Trace",
  "description": "good road condition and visibility",
  "propertyBundle": [
    {
      "@type": "Device",
      "brand": "Garmin",
    }
  ]
},
{
  "@id": "maintenance-report-uuid",
  "@type": "Trace",
  "propertyBundle": [
    {
      "@type": "File",
      "filePath": "\Bob\home",
      "fileName": "maintenance-report.pdf"
```

```
      }
    ]
  },
  {
    "@id": "EDR-uuid",
    "@type": "Trace",
    "propertyBundle": [
      {
        "@type": "EDR",
        "data": {
          "speed": "120km/h",
          "isBraked": "true",
          "gear": "5"
        }
      }
    ]
  },
  {
    "@id": "location-uuid",
    "@type": "Location",
    "propertyBundle": [
      {
        "@type": "SimpleAddress",
        "locality": "Örebro",
        "region": "Sweden",
        "postalCode": "70210"
      },
      {
        "@type": "LatLongCoordinates",
        "latitude": "59.293257",
        "longitude": "15.199069"
      }
    ]
  },
  {
    "@id": "forensic-action1-uuid",
    "@type": "ForensicAction",
    "name": "extracted",
    "startTime": "2020-05-06T15:36:20Z",
    "endTime": "2020-05-08T09:30:48Z",
    "propertyBundle": [
      {
        "@type": "ActionReferences",
        "location": "location-uuid",
        "target": "Bob-uuid",
        "object": ["car-uuid","dashcam-uuid",
```

```
      "maintenance-report-uuid","EDR-uuid"],
      "result": "annotation-uuid"
    }
  ]
},
{
  "@id": "annotation-uuid",
  "@type": "Annotation",
  "tag": ["forensic"],
  "description": "The accident is likely caused by speeding",
  "object": ["forensic-action1-uuid","EDR-uuid"]
}
```

Listing 2: CASE representation of the scenario

There is an overall *Investigation* object that refers eight sub-objects, where each sub-object stores some information. Specially, an object representing the forensic action and another representing the preliminary forensic result are shown. Referring other objects is accomplished by the unique UUID.

Up to now, the collected data have been filtered locally resulting in a reduced data set. The data are then uploaded to the cloud for classification and formatting, after which they will be preserved in the cloud as a chunk for further steps.

4.4 Data preservation

A reduced and classified data set with a unified format is available from the previous steps. One approach for preserving the data is to have a dedicated device on vehicle. In our approach, however, they will be preserved in a permissioned blockchain, which provides more privacy as it is only accessible to members who have been granted permissions by the administrator. Nodes are permitted to perform certain actions by presenting certificates. A permissioned blockchain has the benefits of a blockchain, as well as the authority aspect of a centralized system. Nodes in the network are identified, rather than being anonymous in a permissionless blockchain. Since it is not publicly accessible, the contents of blocks are more transparent in the network among the members. In addition, permissioned blockchain is faster because of the partially decentralized structure and fewer nodes, resulting in better performance. Moreover, an important step in traditional blockchain is finding a proof-of-work, such as a hash with particular pattern. However, the real-time requirement of ADF make it not applicable. In our work, each vehicle proposes its own blocks and broadcasts. Table 4.6 summarizes the differences between permissioned and permissionless blockchain. In previous steps, data have been filtered locally and uploaded to the cloud for classification and formatting, and then stored as a chunk in the cloud. In this step, the cloud generates a unique ID for each data chunk by concatenating the timestamp and the hash of the VIN. This approach offers several benefits. Firstly, the ID is

Table 4.6: Comparison of permissioned and permissionless blockchain

Property	**Permissioned**	**Permissionless**
Privacy	Accessed by members	Open
Decentralization	Partially decentalized	Totally centralized
Anonymity	Identified	Anonymous
Consensus Mechanism	PBFT, FBA, RRC	PoW, PoS, PoC
Speed	Faster	Slower
Performance	Higher	Lower

guaranteed to be unique even if a vehicle generates multiple blocks with identical VINs, due to the uniqueness of the timestamp. Secondly, investigators can search for blocks within a specific time period by indexing the timestamps. Furthermore, the VIN is private information for vehicle owners, and hashing is an effective way to protect it. After the ID being generated, the cloud returns it back to the vehicle. The vehicle then proposes a new block containing the hash of the previous block and the received ID. Blocks are linked through hashes, and there is a one-to-one correspondence between a data chunk in the cloud and a block in the vehicle. The ID is similar to a pointer in terms of functionalities, i.e., they are both addresses and point to a location where the actual contents are stored in. It is then broadcast to all nodes over the network so that each node creates a copy of this block. Finally, if the hash is checked to be correct, the nodes accept the block by appending it to the top of the chain, reaching a consensus and continue waiting for new blocks. In case that the vehicle is disconnected, we have a buffer mechanism that can store the generated block temporarily. When it is online again, the blocks are sent out. Figure 4.4 illustrates the procedures of *data preservation*.

4.5 Data retrieval

Two possibilities exist in this step. Blocks and chunks are erased to save space and reduce the load of the system if no accidents occur after a period of time (e.g., one month). Otherwise, the data will be read out from the cloud according to the chunk ID stored in the blockchain.

4.5.1 Block erasing

Due to limited storage capacity of both the vehicle and the cloud, it is infeasible to preserve all blocks or chunks without deleting old ones. To estimate the required storage space, we conduct a simple calculation on blocks as follows.

In traditional blockchain, a new block is generated every ten minutes, and we can apply this to our work as well. To further reduce the data stored in each vehicle, the V2X network can be divided into sub-networks that consist of a certain number of nodes , for example, 100000 nodes per sub-network. In this case, vehicles

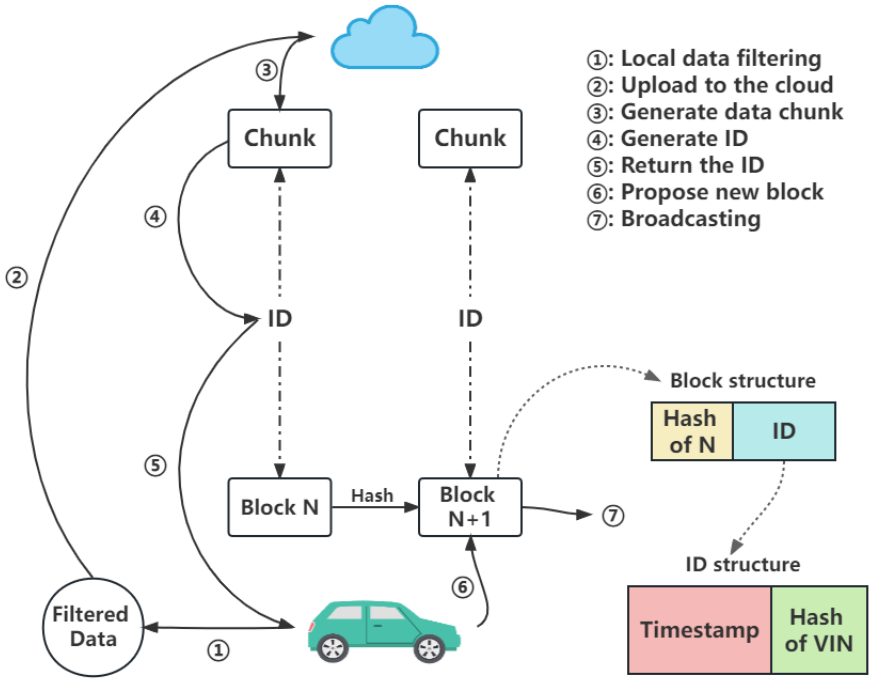

Figure 4.4: Procedures of data preservation

only need to store the blocks generated by those in the same sub-network, which requires less on-vehicle storage space. Suppose we use SHA-256 as hash algorithm and a 32-bit timestamp, the size of a single block is the sum of a 256-bit hash of the previous block, a 32-bit timestamp and another 256-bit hash of the VIN, i.e., $256 + 32 + 256 = 544$bit/block. The average daily running time of a vehicle is assumed to be ten hours. The total space required for a sub-network with 100000 nodes in one month is $544/8 * 10 * 60/10 * 30 * 100000/1024^3 \approx 11.4$GB/month, which is not an issue for most of modern vehicles as they typically have much larger storage space. Therefore, it is reasonable to erase blocks older than one month.

Erasing old blocks is implemented by removing them from the main chain. Additionally, the hash field of next block should also be reset to zero, indicating the initial block of the chain.

4.5.2 Block accessing

When accidents occur, investigators acquire the relevant data by accessing the corresponding blocks. Specifically, estimate the time of the accident, create a range of timestamps and index the blocks using these timestamps, with an additional step of checking the hashes to be correct. Since multiple vehicles may generate blocks within this range, we select the necessary ones based on the VIN hash. Then retrieve the data chunks from the cloud using IDs. In next step, the data contained in the chunks will be analyzed.

4.6 Data analysis

Forensic data analysis examines the structured data with regard to criminal incidents, aims to reveal the truth of the incidents. Specifically for ADF, the task of this step is to take the chunks of data that have been processed in the previous steps as input, analyze them, and produce an output, i.e., the analysis result. It starts with narrowing down the scope of the analysis data, followed by analyzing the specific data, and finally establishing a timeline for this incident and reconstructing the incident based on the analysis result. It is important that the results are reproducible. This is because the results should be the same in order to provide reliable evidence no matter how many times the DA process is repeated.

During the occurrence of incidents, there are expected some anomalies such as the vehicle over speeding, running red lights, crashing or being involved in hit-and-run incidents. Numerous data are stored in the cloud, but not all of them are related to the specific incident. Firstly, we narrow down the data scope based on the time when the event took place, which can be an exact time or a time range. We then retrieve the data chunks by the ID associated with the exact time or time range. Finally, all necessary chunks are presented to the investigators. Here we borrow the naming method of the terminologies in interaction provenance [52], and define: 1) *Actor*, the subjects interacting with each other in the IVN or V2X, such as the driver, the vehicle, the pedestrian, the signal light and the RSU. 2) *Interaction*, is a CASE file containing a description of a message or an action exchanged between the *Actors*. 3) *Event*: is a set of *Interactions* triggered by the *Actors* to perform an operation and, 4) *Story*: is a list of *Events* consisting of ordered *Interactions*. 5) *Incident*: a series of *Stories* ordered chronologically.

Different incidents are always closely related to specific vehicle modules, which motivates investigators to focus on analyzing data of specific components. For example, analyzing the state of acceleration, braking, and steering is of great importance when in a crash. In this scenario, the forensic information is dug up by the path in the chunks as follows, *category*, *subcategory* of specific components, finally to the exact *CASE files*. When we get in touch with the CASE files, for its unified and forensic manner designed format, the information of the time, the location, the speed, the operation of the driver, the VIN of the vehicle, the messages and the actions between them are presented in a clear way, that is, the *Interactions*. The CASE files in the same subcategory of component form an *Event* of this component interacts with other actors. Whereafter the related *Events* combined together, the *Story* is presented. When all the *Stories* have been created, the timeline is established and the reconstruction of the *Incident* is complete. The entire procedure of data analysis is shown in Figure 4.5.

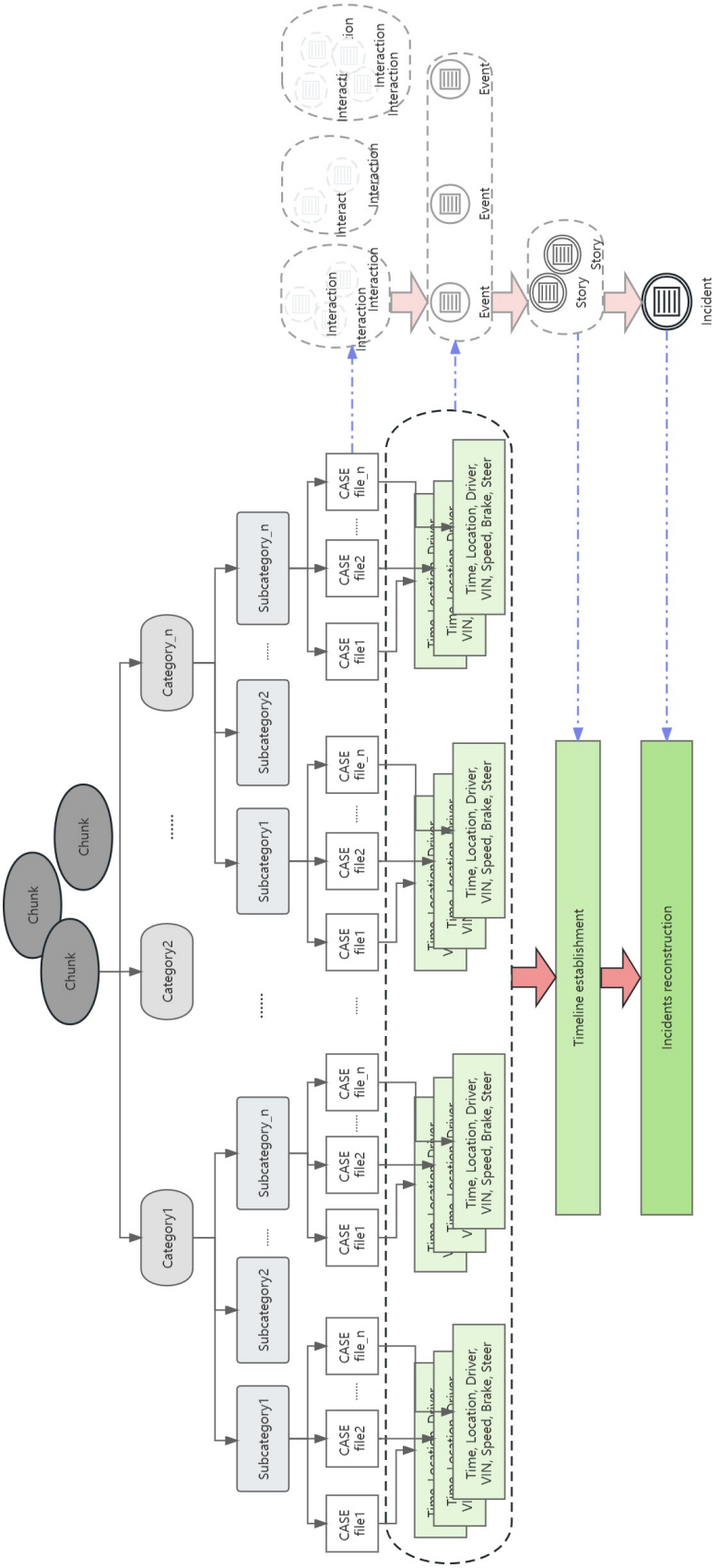

Figure 4.5: Data analysis

4.7 Documentation

This is the last step of forensic lifecycle. Previous procedures and results are documented in this step, such as timestamps, tools, data sources and data analysis results. It is worth noting that only facts can be recorded. Any judgements, deductions or hypotheses should be excluded from the final report, which is later presented on the court.

4.7.1 Results collection

Each of the previous steps takes inputs and generates outputs. Our task is to record all the outputs in a comprehensive and unified way. Figure 4.6 is an example of the results document that used to record all the relevant data of one investigation. The investigators are responsible for filling the form after completing each step.

In the results document, the *Case ID* field, which is the serial number of the documents, is created first. Then the five *W* elements are recorded, namely Who (*Investigator*) and Whom (*Driver*), When (*Date* and the *Timestamp*), Where (*Location*), What (*Type*) and How (*Step* and *Tools*). Specifically, the *Investigator* is the subject, responsible for performing ADF and filling this document, while the *Driver* and Vehicle are the objects, including the driver's personal information, the vehicle *Model* and *VIN*. *Type* indicates the type of incident, such as a crash, a hardware failure or a cyber attack. Details of the forensic process are also recorded, including forensic sub-steps, *Block IDs* along with their *Validity*, *Chunk Size*, *Tools* used in this step, and the *Data Sources*. Finally, a *Description* field records the entire life cycle of the incident in detail. This results document records the entire ADF procedures step-by-step and helps both the DA and DO steps to be repeatable and the results to be reproducible.

4.7.2 Reporting

During previous steps, the timeline has been established, the accidents have been reconstructed and a number of results documents are available. In this final step, investigators report these findings on the court, marking the end of the entire forensic lifecycle.

Results Document					
Case ID		Date (Timestamp)			
Location					
Investigator		Type			
Step	DC DP DV DR DA DO				
Block IDs		Validity		Chunk Size	
Tools		Data Sources			
Driver		Model		VIN	
Description					

Figure 4.6: Results document form

5
Discussion

In this chapter, we discuss our approach, with special attention on highlighting innovations and limitations. We then identify the ethical issues. In addition, we analyze the capability of the adversary in the attacker model.

5.1 Innovation

As mentioned before, ensuring data security properties and unifying forensic data formats are identified as challenges of ADF. We addressed these problems by utilizing CASE and blockchain. More specifically, we summarized the following innovative aspects of our work:

1) Other solutions pay more attention to usability instead of privacy. In our solution, however, privacy is an important concern that should be well protected. We implemented it on the cloud by classifying data into several categories by different classification strategies, such as safety-related data, security-related data and personal data in terms of content. When performing ADF, the investigators only need to access necessary data in corresponding category and subcategory. For example, in case of a car crash accident, safety-related data is of more interest than personal data, thus personal data related to privacy will not be disclosed. Classifying data is implemented by machine learning algorithms, which requires tremendous computing powers, we thus deploy it on the cloud. Another measure concerning privacy is that each block stores the hash of VIN, as it can uniquely identify a vehicle. Generally, knowing the hash of a string, one cannot compute the original text reversely. Therefore, due to this one-way characteristic of hash function, the VIN information can be protected.

2) Unifying forensic data formats is a problem for not only ADF, but all branches of digital forensics. Vehicle manufacturers, service providers and software developers have their own data formats and can lead to information being out of synchronization. CASE solves this problem by converting other data formats into JSON-LD, which makes data management and interpretation easier. With a unified data format, only one interpreter is needed, resulting in reduced complexity and improved interoperability.

3) The blockchain technology is originally designed as a distributed ledger, with the ability of ensuring data security properties by employing hash algorithms.

We adopted this concept to guarantee data integrity, availability and non-repudiation in ADF. Since blocks are linked through hash values, any attempt to modify one block can be detected by examining its hash, thus ensuring data integrity. Furthermore, non-repudiation is implied by the fact that a block mush have existed at the time to get into the chain. Due to duplication of the main chain in each node, data are always available as long as more than half of nodes stay honest. For confidentiality, we assume the cloud to be secure and will not be compromised. As a result, the data preserved on the cloud can also be considered confidential. However, ensuring the security of the cloud itself is out of our work's scope.

4) A vehicle can produce a huge volume of data every day, it is infeasible to store or transmit them all. We have applied three methods to reduce data size. Firstly, there is a sub-step of local data filtering that aims at extracting the forensic-relevant data. Secondly, the actual data are stored on the cloud, which can be considered having unlimited storage space, while the ID is transmitted between the cloud and vehicle. ID is typically a 288-bit number and can easily be sent via internet. In addition, unnecessary blocks are erased after some time. All these measures can reduce data volume to a great extent.

5) The ID field consists of a timestamp and a hash of VIN. The timestamp allows investigators to search for certain chunks, as well as recording the time of accident. The VIN field distinguishes the vehicle since it is unique around the world. Therefore, the concatenation of these two values is able to uniquely determine the data chunk of a desired vehicle and time.

5.2 Limitation

We used multiple strategies to reduce the data volume and the complexity of the data structure, making it a lightweight solution. However, potential limitations still exist.

1) **Storage limitation**

As assumed in subsection 4.5.1, for a sub-network of 100000 nodes, each running for 10 hours per day, the total space needed in a vehicle is approximately 11.4GB per month. On the other hand, the preprocessed data are all stored as chunks in the cloud. According to an estimation, a vehicle produces 25GB of data per hour, most of which are HD video and music streaming, or generated by web browsing [57]. Assuming 1/10000 of them are valuable for forensics, then 75TB of data are uploaded to the cloud per month in a sub-network.

This is an ideal assumption. In a real scenario, the size of the sub-network may be larger than 100000 nodes, for example in the super cities such as New York. In addition, a certain number of vehicles are used for business, such as transportation vehicles and taxis, which usually run for more than 10 hours a day.

And the proportion of valuable data would be more than 1/10000. Furthermore, we unified the data in CASE format, which can add additional information, for example, the predefined fields in the format that form a CASE framework, such as the *"@id"*, *"@type"*, *"description"*. The numerous number of such fields significantly increase the data volume.

As a result, the actual storage space required for the vehicle or cloud might be slightly larger than assumed, but in an acceptable range. Vehicle storage space usually depends on the model and the specific configuration. In the state-of-the-art modern vehicles, there are larger storage spaces, and then it is not a critical issue, but the storage limitation is still worth considering.

2) **Computation and time consumption**

Machine learning algorithms usually imply massive computational requirements. For example, in Convolutional Neural Networks (CNNs), which are useful and effective for classifying data, the training speed and prediction time of the models depend on the time complexity. The time complexity is the overall number of computations that the algorithms perform, and it depends on the computation times at each layer and the layer numbers of the CNNs models. In our framework, machine learning based approaches are widely used during data pre-processing sub-steps, i.e., data filtering and classification. Besides the algorithms, data format standardization can increase time complexity as well. Currently, there lacks an automated approach to convert data into CASE format and the conversion is tentatively done by human, which consumes a huge amount of time.

In summary, our approach, with its innovations and advantages, is suitable for most common scenarios. Nevertheless, in extreme cases there are still concerns about the storage, computation and time consumption, which may be limitations of our approach.

5.3 Ethics

Due to the immaturity of research in the field of ADF, there are ethical issues relating to both individuals and organizations.

1) **Privacy of the stakeholders**

We mentioned previously that there are privacy concerns for ADF. To be more specific, when collecting accident-related data, we may unintentionally touch upon stakeholders' private information, whether sensitive or not. For example, if the camera has taken a photo of an innocent pedestrian, then his

privacy is violated. Similarly, if the sound recorder has recorded irrelevant information about the driver's private life, this may cause great inconvenience or trouble to the driver. With GPS, we also have to consider the problem that whether it is legal to share someone's location. In particular, the state-of-the-art vehicles are usually equipped with a multi-function infotainment system, which increases the risk of revealing the privacy of those involved, not only the driver, but also the passengers and the pedestrian.

GDPR is a comprehensive data privacy law that all industries and organizations in the European region must comply with. For ADF, GDPR provides a legal guidance on the collection, processing, storage and transfer of personal data, meaning that all personal data must be handled in a secure manner to protect the privacy of the stakeholders.

2) **Ethics of AI algorithms**

As machine-learning algorithms are used in our approach, the ethical issues of AI cannot be ignored. If AI algorithms are misused in criminal identification, it will inevitably compromise people's rights. For example, attackers can imitate others biological characteristics like voice using AI algorithms. There can also be deviations by using different algorithms and data sets. An algorithm is in essence a process of using past data to predict the future, where the outcomes are determined by algorithms and the input data together. Therefore, these two factors become the main sources of deviation. In addition, the availability and accuracy of the input data can also affect the accuracy of the prediction. If there are malicious nodes existing in the networks, the data provided by them can affect the results of ADF.

There are also human rights concerns arising from the ethics of algorithms. Algorithms are opinions expressed mathematically or in computer code. They are subjectively designed and chosen by the designers and the developers based on their own judgement [58]. In the ADF scenario, algorithms can help identify the criminal. There is a risk that just because someone has been a criminal once, they will always be a criminal in the algorithms. In this case, an innocent person can become a victim due to the mistakes of the algorithms.

5.4 Attacker model

Here we will hypothesise attacks from adversaries and evaluate the impact on the system, the cloud and the vehicles respectively. In the attacker model, we state the potential purpose of the attackers, analyze their capabilities, and argue about how our solution is able to mitigate the threats.

Vehicles are in an insecure environment and there are various threat actors sabotaging them constantly, such as a person trying to alter digital evidence to avoid prosecution. Attackers can get access to forensic data from the cloud or from the vehicle. The cloud is connected to the vehicles via the internet, which opens the possibility of attacking. Therefore, a common attack vector is to compromise cloud servers to retrieve, manipulate or delete digital evidence. As the cloud stores forensic data with highest confidentiality level, it should not be compromised under any circumstances, which requires strict security measures. We assume that the cloud has applied state-of-the-art security measures to minimize potential threats. In addition, our assumption also includes a backup mechanism for cloud storage to keep a copy of the forensic data.

For vehicles, the attack vectors are mostly passive attacks aimed at eavesdropping. Hardware and software could be controlled or compromised depending on the access to the IVN and V2X communication, either wired or wireless. So there are several attack vectors, such as hacking through the OBD-II port to eavesdrop on traffic between ECUs and communication buses, or remote hacking of the infotainment systems to compromise personal information, or even physically destroy vehicle components. Thus, we cannot protect the actual data itself. However, it can be considered infeasible for the adversary to carry out undetected attacks to tamper the IDs stored in the chain with the nature of use of the blockchain.

Our solution can mitigate the threats on vehicles since the blockchain network is resilient to data tampering. It is very unlikely for any individual attacker to control more than half of global vehicles. Therefore, as long as at least 51% nodes stay honest we can guarantee the digital evidences are reliable. Nevertheless, the security of cloud servers is not in our scope and left to be the future work.

6
Conclusion

In this section we identify what we have done, summarize and evaluate our work. We also identify what efforts can be made in future work based on our approach.

6.1 Conclusion

Currently, there are many solutions for digital forensics, but few of them regard ADF. Forensics on vehicles is also an important topic for researching since there are over 5M car accidents every year [59]. Fast and accurately determining the causes of accidents has become an essential requirement. Some scholars proposed macro-level frameworks for ADF, while others presented solutions targeting specific problems, but they all have limitations. Lacking of standardized forensic data formats and difficulties in ensuring data security properties are two main limitations. To address them, we proposed an innovative solution utilizing CASE and blockchain.

In this thesis, we started by introducing the background of ADF and related concepts, identifying challenges and describing purposes, giving readers an overview of what is lacking on existing solutions. Follow that we introduced related works from four aspects, which showed what efforts have been done. Then we discussed details of literature review and similar areas, and borrowed some ideas from others' work, e.g., utilizing blockchain to protect data security properties. In chapter 4, we described our research results, i.e., the detailed steps of the framework and how it ensures data security properties. Finally we analyzed the innovations and limitations, and conclude the thesis.

Compared to other solutions, we pay special attention to data formats unification and security properties, where the former is implemented by converting any data format into a JSON-LD format and the latter is primarily implemented by blockchain. Although making forensic investigation more convenient, limitations still exist. Due to the increasing number of vehicles around the world, data volume can be a main challenge for ADF, since the storage space and transmission bandwidth are limited. Further reducing data size is also an important topic for future works.

6.2 Future work

The innovations and advantages of our approach have been identified so far, there are still efforts to be made in the future. First of all, we have not specified the data

filtering strategy in detail, but it is an important topic to solve in the future. Data filtering aims to reduce the huge volume of data by excluding and rearranging raw data based on certain criteria. Defining the filtering criteria is a critical step at the beginning of the filtering, as it determines whether data is retained or removed. If the criteria are not accurate, it may result in valuable data being filtered out.

In addition, data format being CASE-ed can improve the efficiency of data pre-processing. CASE is a unified data expression that helps ADF interpret data from a unified single format, significantly reducing the complexity of interpreting data from various data formats in previous forensic approaches. However, the process of unifying multiple different data formats into CASE is not straightforward. Our vision for the future is to use programming languages such as C or python to implement automated format standardization, like an Application Programming Interface (API).

Furthermore, future work will also focus on identifying the potentially hidden malicious nodes in the network. When handling cyber security issues, it is inevitable to consider the presence of adversaries. In ADF, if there are malicious nodes, whether they are scattered or organized, it will bring unpredictable consequences. In our permissioned blockchain based framework, malicious nodes cannot take control of the entire network, but if they upload malicious code or fake data to the cloud, it may more or less affect the results of ADF. In future work, the membership of nodes in the chain should be authenticated by Vehicular Public Key Infrastructure (VPKI).

Bibliography

[1] *FORENSICS | definition*, https://dictionary.cambridge.org/us/dictionary/english/forensics.

[2] K. Strandberg, N. Nowdehi, and T. Olovsson, "A systematic literature review on automotive digital forensics: Challenges, technical solutions and data collection," *IEEE Transactions on Intelligent Vehicles*, 2022.

[3] R. Rak and D. Kopencova, "Actual issues of modern digital vehicle forensics," *Internet of Thinks and Cloud Computing*, vol. 1, pp. 12–16, 2020.

[4] N.-A. Le-Khac, D. Jacobs, J. Nijhoff, K. Bertens, and K.-K. R. Choo, "Smart vehicle forensics: Challenges and case study," *Future Generation Computer Systems*, vol. 109, pp. 500–510, 2020.

[5] K. K. G. Buquerin, C. Corbett, and H.-J. Hof, "A generalized approach to automotive forensics," *Forensic Science International: Digital Investigation*, vol. 36, p. 301 111, 2021.

[6] *Automotive digital forensics*, https://taltech.ee/en/news/article-automotive-digital-forensics.

[7] A. Attenberger, "Data sources for information extraction in automotive forensics," in *Computer Aided Systems Theory–EUROCAST 2019: 17th International Conference, Las Palmas de Gran Canaria, Spain, February 17–22, 2019, Revised Selected Papers, Part II 17*, Springer, 2020, pp. 137–144.

[8] M. Cebe, E. Erdin, K. Akkaya, H. Aksu, and S. Uluagac, "Block4forensic: An integrated lightweight blockchain framework for forensics applications of connected vehicles," *IEEE Communications Magazine*, vol. 56, no. 10, pp. 50–57, 2018. DOI: 10.1109/MCOM.2018.1800137.

[9] E. A. Bates, "Digital vehicle forensicitecs," *online].[cit, 2019-11-17]. Available at https://abforensics.com/wp-content/uploads/2019/02/INTERPOL-4N6-PULSE-IssueIV-BATES.pdf*, 2019.

[10] N. Vinzenz and T. Eggendorfer, "Forensic investigations in vehicle data stores," in *Proceedings of the Third Central European Cybersecurity Conference*, ser. CECC 2019, Munich, Germany: Association for Computing Machinery, 2019, ISBN: 9781450372961. DOI: 10.1145/3360664.3360665. [Online]. Available: https://doi.org/10.1145/3360664.3360665.

[11] K. Khorsravinia, M. K. Hassan, R. Z. A. Rahman, and S. A. R. Al-Haddad, "Integrated obd-ii and mobile application for electric vehicle (ev) monitoring system," in *2017 IEEE 2nd International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, 2017, pp. 202–206. DOI: 10.1109/I2CACIS.2017.8239058.

[12] M. Ammar, H. Janjua, A. Thangarajan, B. Crispo, and D. Hughes, "Securing the on-board diagnostics port (obd-ii) in vehicles," Jul. 2020.
[13] J. Lacroix, K. El-Khatib, and R. Akalu, "Vehicular digital forensics: What does my vehicle know about me?," Nov. 2016, pp. 59–66. DOI: 10.1145/2989275.2989282.
[14] *CAN bus*, https://en.wikipedia.org/wiki/CAN_bus.
[15] *ISO/IEC 27037:2012*, https://www.iso.org/standard/44381.html.
[16] *ISO/SAE 21434:2021*, https://www.iso.org/standard/70918.html.
[17] *GDPR*, https://gdpr-info.eu/.
[18] D. Jacobs, K.-K. R. Choo, M.-T. Kechadi, and N.-A. Le-Khac, "Volkswagen car entertainment system forensics," in *2017 IEEE Trustcom/BigDataSE/ICESS*, 2017, pp. 699–705. DOI: 10.1109/Trustcom/BigDataSE/ICESS.2017.302.
[19] R. Altschaffel, K. Lamshöft, S. Kiltz, and J. Dittmann, "A survey on open automotive forensics," Apr. 2017.
[20] P. Sharma and J. Gillanders, "Cybersecurity and forensics in connected autonomous vehicles: A review of the state-of-the-art," *IEEE Access*, vol. 10, pp. 108 979–108 996, 2022. DOI: 10.1109/ACCESS.2022.3213843.
[21] L. Davi, D. Hatebur, M. Heisel, and R. Wirtz, "Combining safety and security in autonomous cars using blockchain technologies," in *Computer Safety, Reliability, and Security: SAFECOMP 2019 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Turku, Finland, September 10, 2019, Proceedings 38*, Springer, 2019, pp. 223–234.
[22] A. D. Sathe and V. D. Deshmukh, "Advance vehicle-road interaction and vehicle monitoring system using smart phone applications," in *2016 Online international conference on green engineering and technologies (IC-GET)*, IEEE, 2016, pp. 1–6.
[23] N. Watthanawisuth, T. Lomas, and A. Tuantranont, "Wireless black box using mems accelerometer and gps tracking for accidental monitoring of vehicles," in *Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics*, IEEE, 2012, pp. 847–850.
[24] K. Dološ, C. Meyer, A. Attenberger, and J. Steinberger, "Driver identification using in-vehicle digital data in the forensic context of a hit and run accident," *Forensic Science International: Digital Investigation*, vol. 35, p. 301 090, 2020.
[25] *Bosch CDR*, https://cdr.boschdiagnostics.com/cdr/.
[26] *The iVe Ecosystem*, https://berla.co/ecosystem/.
[27] *Digitpol*, https://digitpol.com/our-company/.
[28] *Envista Forensics*, https://www.envistaforensics.com/.
[29] D. Sladović, D. Topolčić, K. Hausknecht, and G. Sirovatka, "Investigating modern cars," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2019, pp. 1159–1164. DOI: 10.23919/MIPRO.2019.8756732.
[30] H. Mansor, K. Markantonakis, R. N. Akram, K. Mayes, and I. Gurulian, "Log your car: The non-invasive vehicle forensics," in *2016 IEEE Trustcom/BigDataSE/ISPA*, 2016, pp. 974–982. DOI: 10.1109/TrustCom.2016.0164.

[31] *Technology Readiness Level*, https://en.wikipedia.org/wiki/Technology_readiness_level.

[32] M. Stoyanova, Y. Nikoloudakis, S. Panagiotakis, E. Pallis, and E. K. Markakis, "A survey on the internet of things (iot) forensics: Challenges, approaches, and open issues," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 1191–1221, 2020. DOI: 10.1109/COMST.2019.2962586.

[33] D.-P. Le, H. Meng, L. Su, S. L. Yeo, and V. Thing, "Biff: A blockchain-based iot forensics framework with identity privacy," in *TENCON 2018-2018 IEEE region 10 conference*, IEEE, 2018, pp. 2372–2377.

[34] S. Brotsis, N. Kolokotronis, K. Limniotis, *et al.*, "Blockchain solutions for forensic evidence preservation in iot environments," in *2019 IEEE Conference on Network Softwarization (NetSoft)*, IEEE, 2019, pp. 110–114.

[35] Z. Tian, M. Li, M. Qiu, Y. Sun, and S. Su, "Block-def: A secure digital evidence framework using blockchain," *Information Sciences*, vol. 491, pp. 151–165, 2019.

[36] S. Singh, I.-H. Ra, W. Meng, M. Kaur, and G. H. Cho, "Sh-blockcc: A secure and efficient internet of things smart home architecture based on cloud computing and blockchain technology," *International Journal of Distributed Sensor Networks*, vol. 15, no. 4, p. 1 550 147 719 844 159, 2019.

[37] M. A. Rahim, M. A. Rahman, M. Rahman, A. T. Asyhari, M. Z. A. Bhuiyan, and D. Ramasamy, "Evolution of iot-enabled connectivity and applications in automotive industry: A review," *Vehicular Communications*, vol. 27, p. 100 285, 2021, ISSN: 2214-2096. DOI: https://doi.org/10.1016/j.vehcom.2020.100285. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214209620300565.

[38] Á. MacDermott, T. Baker, P. Buck, F. Iqbal, and Q. Shi, "The internet of things: Challenges and considerations for cybercrime investigations and digital forensics," *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 12, no. 1, pp. 1–13, 2020.

[39] G. T. Coll, J. F. Pellegrino, and J. Pilchuk, "Black box: Improving aircraft safety by bringing the black box from the bottom of the sea to outer space," in *AIAA SPACE and Astronautics Forum and Exposition*. DOI: 10.2514/6.2017-5130. eprint: https://arc.aiaa.org/doi/pdf/10.2514/6.2017-5130. [Online]. Available: https://arc.aiaa.org/doi/abs/10.2514/6.2017-5130.

[40] D. Mink, A. Yasinsac, K.-K. R. Choo, and W. B. Glisson, "Next generation aircraft architecture and digital forensic.," in *AMCIS*, 2016.

[41] J. Cosic, C. Schlehuber, and D. Morog, "Digital forensic investigation process in railway environment," in *2021 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, 2021, pp. 1–6. DOI: 10.1109/NTMS49979.2021.9432658.

[42] J. Cosic, C. Schlehuber, and D. Morog, "New challenges in forensic analysis in railway domain," in *2019 IEEE 15th International Scientific Conference on Informatics*, 2019, pp. 000 061–000 064. DOI: 10.1109/Informatics47936.2019.9119288.

[43] Wikipedia contributors, *Smart city — Wikipedia, the free encyclopedia*, [Online; accessed 4-April-2023], 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Smart_city&oldid=1147124936.

[44] R. Khatoun and S. Zeadally, "Smart cities: Concepts, architectures, research opportunities," *Commun. ACM*, vol. 59, no. 8, pp. 46–57, Jul. 2016, ISSN: 0001-0782. DOI: 10.1145/2858789. [Online]. Available: https://doi.org/10.1145/2858789.

[45] X. Feng, E. S. Dawam, and S. Amin, "A new digital forensics model of smart city automated vehicles," in *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, IEEE, 2017, pp. 274–279.

[46] E. Casey, S. Barnum, R. Griffith, J. Snyder, H. van Beek, and A. Nelson, "Advancing coordinated cyber-investigations and tool interoperability using a community developed specification language," *Digital Investigation*, vol. 22, pp. 14–45, 2017, ISSN: 1742-2876. DOI: https://doi.org/10.1016/j.diin.2017.08.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1742287617301007.

[47] Wikipedia contributors, *Blockchain — Wikipedia, the free encyclopedia*, [Online; accessed 11-April-2023], 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Blockchain&oldid=1149194310.

[48] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Cryptography Mailing list at https://metzdowd.com*, Mar. 2009.

[49] *CAN in Automation (CiA)*, https://www.can-cia.org/.

[50] *LIN*, https://www.autopi.io/blog/lin-bus-protocol-explained/.

[51] *Navigation Data Standard (NDS)*, https://nds-association.org/.

[52] M. Hossain, R. Hasan, and S. Zawoad, "Trust-iov: A trustworthy forensic investigation framework for the internet of vehicles (iov)," in *2017 IEEE International Congress on Internet of Things (ICIOT)*, 2017, pp. 25–32. DOI: 10.1109/IEEE.ICIOT.2017.13.

[53] J. Daily, M. DiSogra, and D. Van, "Chip and board level digital forensics of cummins heavy vehicle event data recorders," *SAE International Journal of Advances and Current Practices in Mobility*, vol. 2, no. 2020-01-1326, pp. 2374–2388, 2020.

[54] X. Guo, T. Aoki, and H.-H. Lin, "Model checking of in-vehicle networking systems with can and flexray," *Journal of Systems and Software*, vol. 161, p. 110 461, 2020, ISSN: 0164-1212. DOI: https://doi.org/10.1016/j.jss.2019.110461. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0164121219302353.

[55] *X-Ways Forensics: Integrated Computer Forensics Software*, https://www.x-ways.net/forensics/index-m.html.

[56] *Encase vs FTK vs X-Ways Review*, https://cloudyforensics.medium.com/encase-vs-ftk-vs-x-ways-review-2b7b075333ef.

[57] *Connected-Car-Data-Generation*, https://www.statista.com/chart/8018/connected-car-data-generation/.

[58] K. Martin, "Ethical implications and accountability of algorithms," *Journal of Business Ethics*, vol. 160, Dec. 2019. DOI: 10.1007/s10551-018-3921-3.

[59] *Car Accident Statistics for 2023*, https://www.forbes.com/advisor/legal/car-accident-statistics/.