EXPLORING DATA EXHAUST IN IOT DEVICES WITH A FOCUS ON VOICE ASSISTANTS

by

Mahdieh Mellaty

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
November 2023

I dedicate this thesis to my cherished husband, Hamed. His unwavering support, constant encouragement, and boundless love have been my pillars of strength throughout this journey. Hamed's belief in me and the sacrifices he has made have enabled me to pursue my academic aspirations, for which I am eternally grateful.

Additionally, I dedicate this thesis to my parents, Farzaneh and Masoud, who are currently far from me in our beautiful homeland of Iran. Despite the physical distance, their enduring love, unwavering support, and numerous sacrifices have played a pivotal role in shaping the person I have become.

This thesis serves as a testament to the love, belief, and support I have received from my husband, parents, and little sister, which have illuminated my path and made this academic achievement possible. I extend my heartfelt gratitude to them for always being there for me.

Table of Contents

List of Tables

List of Figures

Abstract

In the realm of the Internet of Things (IoT), data exhaust encompasses the unintended data generated during device interactions over the Internet. This unintentional data collection can be analyzed by businesses and third parties to gain insights into user behaviour. Consequently, there exists a keen interest among various parties in accessing the details of this data.

This thesis has two main objectives. Firstly, we surveyed approximately 48 research papers, with thirty of them focusing on IoT architecture, layers, components, and ecosystem attributes. Another thirteen papers concentrated on the pressing issue of data privacy within IoT devices and ecosystems. Surprisingly, only five papers broached the concept of data exhaust, and none explored its nuances across different IoT device types. We aim to bridge this knowledge gap by providing a comprehensive analysis of data exhaust issues across various types of IoT devices.

The second objective of the thesis is to design and develop a predictive modelling scheme to introduce a solution designed to safeguard users' privacy while utilizing Voice Assistants (VAs). providing more details regarding VAs. We discuss the structure of a VA ecosystem. Following that, we delve into the journey of data within this ecosystem. Our findings highlight the critical challenge of user awareness regarding data collection in VAs, emphasizing the potential privacy risks this opacity entails.

To present our solution, we use a real dataset namely Amazon Alexa Traffic Traces, provided by Barcel et al. that tracked network traffic involving a VA and included all communications between the user, VA, and the VA server. After obtaining this data, we carefully analyzed it and used different machine learning methods to predict data exhaust. The proposed approach brings users more clarity when interacting with these types of smart devices.

In conclusion, this study outlines a potential solution to address this significant concern, ultimately ensuring that users can make informed choices when engaging with VAs and their data.

List of Abbreviations Used

ACK	Acknowledgement
ANN	Artificial Neural Network
CDR	Call Detail Records
CSV	Comma Separated Value
FTC	Federal Trade Commission
GBRAM	Goal-Based Requirements Analysis Method
GPS	Global Positioning System
HIoT	Healthcare Internet of Things
IIoT	Industrial Internet of Things
IoT	Internet of Things
K-NN	K-Nearest Neighbor
LAN	Local Area Network
MLP	Multi-Layer Perceptron
NFC	Near-field communication
PCAP	Packet Capture
PIoT	Personal Internet of Things
RFID	Radio Frequency Identification
SADLM	Scenario Agnostic Data Lifecycle Model
SSE	Sum of Squared Errors

SVM Support Vector Machine

t-SNE T-distributed Stochastic Neighbor Embedding

VA Voice Assistant

WCSS Within Cluster Sum of Squares

Wi-Fi Wireless Fidelity

WSN wireless sensor network

ZigBee Zonal Intercommunication Global-standard

Acknowledgements

I am deeply thankful to my supervisor, Dr. Srinivas Sampalli, for his consistent belief in my abilities and for entrusting me with this significant journey. From our very first conversation, he has consistently demonstrated persistent support, kindness, and genuine humanity. He will always hold a special place in my heart as one of the most important individuals in my life.

I also want to express my sincere appreciation to Calian, especially to Kevin de Snayer, National Cyber Security Strategist, and Terry Dougall, Vice President, ESG and Industrial Development, for their invaluable collaboration. Their support has been instrumental in my progress. Additionally, I extend my thanks to the MITACS project for their generous support. Their financial support was indispensable in enabling me to pursue my education here in Canada. I am also grateful for the guidance provided by Dr.Jaume Manero and Dr.Nur Zincir-Heywood. The collective trust and support of all these individuals have played a pivotal role in my achievements, and I am genuinely thankful for the opportunities they have given me.

Chapter 1

Introduction

This thesis concentrates on IoT devices, with a specific focus on VAs. Initially, it offers a comprehensive overview of the IoT and subsequently narrows its scope to examine VAs as a distinctive category within the realm of IoT devices.

1.1 Internet of Things

IoT emerged as an important concept that is transforming the world today. It is now possible for anything to be connected to anyone at any time and from anywhere, as long as it has the requisite technology. Using sensors, actuators, Radio Frequency Identification (RFID) tags, and readers, the system is able to enable both physical and virtual interactions with the surrounding environment. IoT typically refers to a network of connected objects that have unique identifiers and are equipped with sensors (e.g., cameras, motion sensors) and actuators, enabling them to transmit generated data over a network like the Internet [56]. The current number of connected IoT devices is estimated to be approximately 13.1 billion, with a projected increase to over 125 billion by 2030 [15]. Globally, there are more than 200 million smart personal assistants installed, and current trends indicate that the number will exceed 500 million by 2030[68]. This growth leads to a data explosion in the current data age, with estimates suggesting that total data storage will exceed 200 zettabytes by 2025 [39][67].

Smarter, faster, and with a higher quality of life, IoT devices are being used in a wide range of fields. From factories to homes, from hospitals to roads, the footprint of the Internet of Things can be seen throughout the world. Health, security, and transportation are a few of the aspects of smart city citizens' lives that are impacted by the Internet of Things. Furthermore, it can play a significant role at the national level in regard to policy decisions (such as energy conservation, pollution reduction,

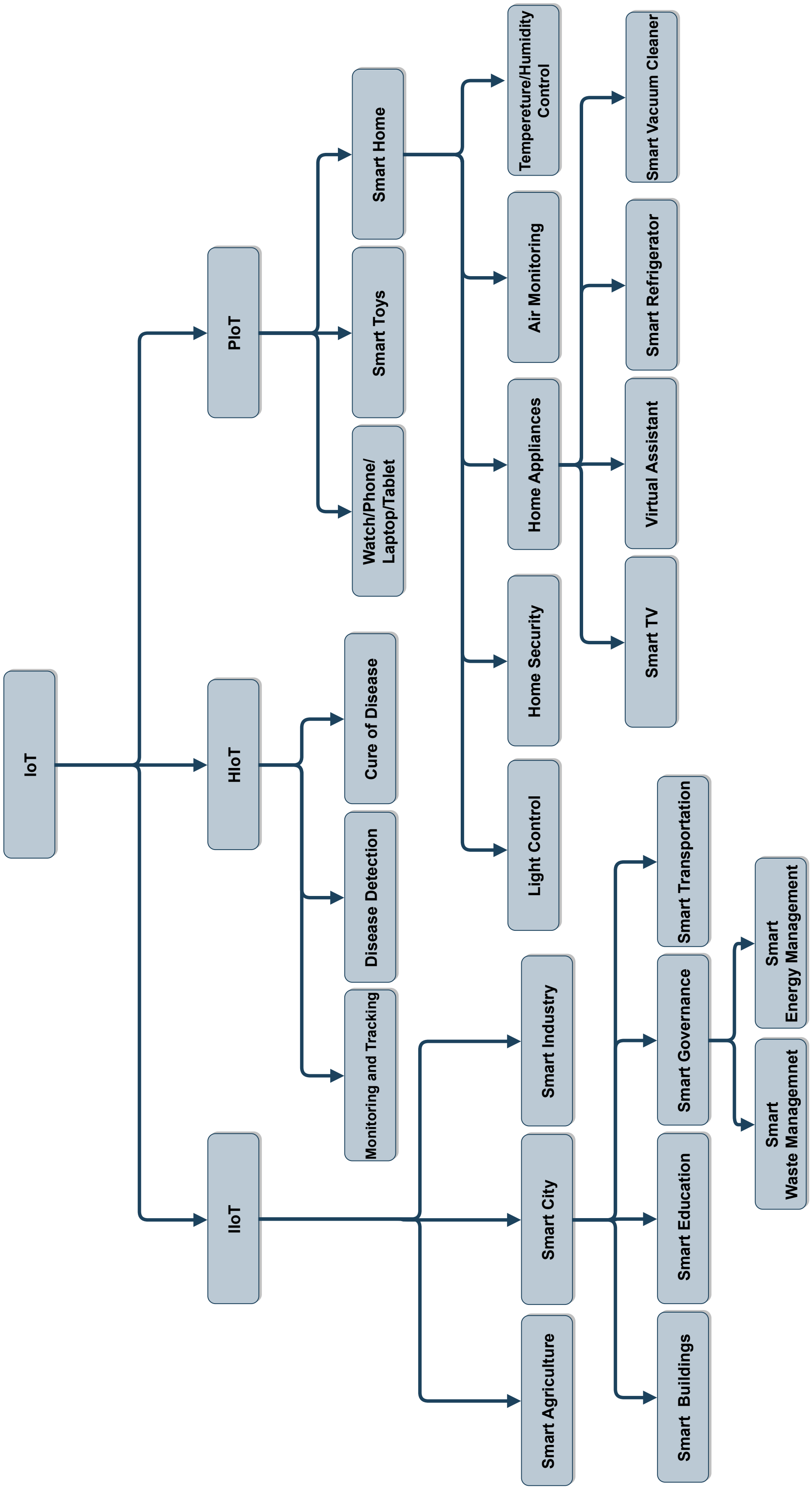


Figure 1.1: IoT Application Categories

etc.), remote monitoring, and infrastructure development. Our systems can operate more efficiently, economically, and securely by leveraging the Internet of Things based on a variety of factors, such as energy-saving policies, economic considerations, and reliability levels. This leads us to categorize the areas in which IoT is being used, as shown in Figure 1.1[14]. According to the figure, there are three general types of IoT applications. Industrial Internet of Things (IIoT), Healthcare Internet of Things (HIoT), and Personal Internet of Things (PIoT). A brief description of each of these types is provided below.

IIoT: This category of application pertains to industrial activities such as agriculture, smart cities, and factories. Remote monitoring solutions, for instance, can simplify and speed up agricultural production. Smart cities are also one of the key areas in which the Internet of Things is advancing rapidly in order to facilitate the implementation of tools that improve city life. Many smart solutions are already being used in countries around the world, from smart parking to smart waste management. In many factories, a smart monitoring system is being introduced to replace traditional production methods[45].
HIoT: Using IoT in healthcare can make patient care better, reduce costs, and improve efficiency. IoT can be used in the healthcare system to control medication, and medical equipment, manage information, supervise patients, and telemedicine[37].
PIoT: "A group of connected devices focused mainly in homes and the immediate proximity of an individual"[58]. This category includes smart devices we use every day. Watches, homes, phones, laptops, tablets, and even smart toys are all examples of how personal IoT is taking off.

The rise of cloud computing, artificial intelligence, and the IoT has led to the increasing popularity of voice assistants in households[65]. Many top technology companies worldwide, including Apple, Amazon, and Google, have created their own VAs such as Siri, Alexa, and Google Assistant. These VAs are not only limited to mobile devices but are also available on various platforms such as smart cars, speakers, and televisions[24].

In addition to deliberately generated core data, IoT devices generate unintentional data, which is referred to as data exhaust generated during internet interactions by humans or smart devices holds significant value for businesses and third parties[48]. Unintentionally generated data can provide valuable insights, enabling a comprehensive understanding of users and customers. By analyzing collected data using data mining techniques [18], extract valuable information. However, the data generated, stored, and transmitted by sensors and actuators may contain sensitive personal information, raising privacy concerns[1].

1.2 Voice Assistnats

VAs, a very popular type of PIoT devices, are always listening devices, activated by a specific wake word such as *Alexa*, *Hey Siri*, or *Okay Google*. On smartphones, this feature can usually be turned off, but on smart speakers, it's essential.[28] Once activated, the device records the sound, which is sent to the cloud for processing, understanding the user's intention, and executing them in real-time[34].

This intention can trigger actions in the smart assistant's cloud or be sent to third-party services. The response is then relayed back to the user's device, sometimes involving other cloud services to control IoT devices[28].

By leveraging connections with third-party services and devices, the VA can carry out a range of functions simply by responding to a user's voice prompt. These functions may include responding to queries, playing music, setting alarms or timers, making phone calls or sending messages, completing purchases, offering updates on the weather, and managing other smart devices[49]. Due to the vast computing power and resources available on cloud platforms, voice assistants can rely on cloud-based servers to handle the mentioned complex tasks such as natural language processing, speech recognition, and machine learning algorithms that require intensive computational power[65].

However, privacy remains a significant concern for consumers, in different aspects. An incident in 2018 revealed the unintentional leak of sensitive military base locations via a fitness tracker social network. Both business and private spheres are affected by these challenges, including security and privacy concerns. Despite the potential benefits of smart assistants for business, IT security professionals are hesitant to adopt them, potentially resulting in missed optimization opportunities. While smart assistants have quickly gained popularity among private users, individuals who are concerned about privacy may have difficulty avoiding them. There is also a concern in society regarding the possibility of large corporations aggregating and exploiting vast quantities of personal information collected by these devices.[28].

In light of this perspective, there are two types of threats to privacy:

(1) Sensitive information exposure by external attackers.

(2) Privacy disclosures to voice service providers [70].

Many studies have addressed the first issue, and various types of attacks have been previously discussed. Li et al.[31], for example, classified different forms of privacy and security attacks in their study. Regarding the latter, this study addresses the relatively under-explored topic of data exhaust in VAs. In this context, data exhaust refers to the traces left behind as a result of user interactions with VAs. This often-overlooked aspect of privacy threats in IoT devices poses significant challenges, prompting an in-depth investigation in this research.

This thesis provides a comprehensive insight into the IoT ecosystem by identifying the different types of data exhaust, with a particular emphasis on personal IoT devices including VAs. It considers privacy-preserving laws and regulations and identifies a potential solution to detect data exhaust. Detecting data exhaust and distinguishing it from core data is crucial in addressing the challenge of transparency within a VA ecosystem. Through the analysis of various works, it has become evident that transparency is a major problem in VA systems. Properly identifying and differentiating data exhaust from core data can significantly contribute to resolving this issue.

1.3 Thesis Outline

The remainder of this thesis is structured as follows: Chapter 2 will provide a comprehensive survey of data exhaust in IoT devices, covering various aspects such as IoT terminology, the IoT ecosystem, architecture, data exhaust in IoT devices and voice assistants, ad targeting and data profiling, as well as Privacy-Preserving Protocols and Laws. In Section 3, the focus will be on presenting the proposed solution to the issues discussed in the second chapter. This section will involve examining the dataset, detailing the methodology, and outlining the different steps of analysis.

Chapter 4 will encompass the conclusion, discussions on future work, and an exploration of challenges faced in the research.

Chapter 2

Systematic survey of Data Exhaust in IoT Devices

The objective of this chapter is to provide a comprehensive overview of data exhaust in IoT devices. In order to gain an understanding of issues, challenges, privacy and security matters with regard to the data collection in these devices, it is important to understand their architecture models and the general data flow.

2.1 Literature Review

This section presents various related works and research to provide a deeper understanding of the topic at hand. The papers are listed according to their publication date.

Iqbal et al [21] designed an auditing framework that utilized online advertising to assess the data collection, usage, and sharing practices of smart speaker platforms. Based on the evaluation results, Amazon and third-party providers collect data on smart speaker interactions and use it to infer user interests to serve targeted ads.

Jiang et al. [22] address privacy concerns associated with the new generation of cyberspace data collection, which include tracking browsing activities, disclosing user input data, making data available via mobile devices, maintaining data security during transmission, protecting participation sensing privacy, and protecting identity in opportunistic networks.

Zainuddin et al.[74] addresses a variety of privacy and security issues pertaining to IoT devices, including unintentional data collection. In this study, various types of IoT applications are enumerated and the privacy concerns associated with each application are discussed.

O'Leary et al. [48] propose a framework for locating and transforming exhaust data. Specifically, they investigate four case studies, namely, Internet search data, accounting entries, social media disclosures, and the use of Edgar logs. While other studies have examined data exhaust as a threat to users' privacy, this study explores the subject as a potential opportunity for businesses to gain valuable insights into their users' preferences.

Ren et al. [56] conduct an analysis of information exposure from approximately 80 devices. They answer a number of questions during their experiment, including "Does the device expose information unexpectedly?" One of the most intriguing points about this research is that they identify unexpected behaviour from audio and video recording devices. In addition, they identify several cases in which exposure varies depending on the device location.

Pierce et al. [52] discuss some of the vulnerabilities of smart home security cameras, highlighting how they monitor and track the most personal and intimate interior spaces. Three key concepts have been highlighted in this study namely digital leakage, hole-and-corner applications, and foot-in-the-door devices. By using these concepts, the paper shows how user experience, interactive technology, and concerns relating to privacy, security, accountability, trust, and fairness are interconnected.

Maher et al [33] also examined ethical concerns as well as solutions related to each ethical concern such as passive data collection, secondary data use, and storage of passive data.

O'Leary et al [47] analyze and suggest an analysis framework for data exhaust. Moreover, through the analysis of a case study, they illustrate the concepts related to data exhaust, including its potential and limitations, as well as how it can be used effectively. According to this study, data exhaust may provide a comprehensive overview of how individuals or groups of individuals processed transactions, providing details regarding the information they accessed and the resources they did not utilize. Inferring an individual's needs, desires, or intentions with this information is possible, making data exhaust an effective tool for gaining insight.

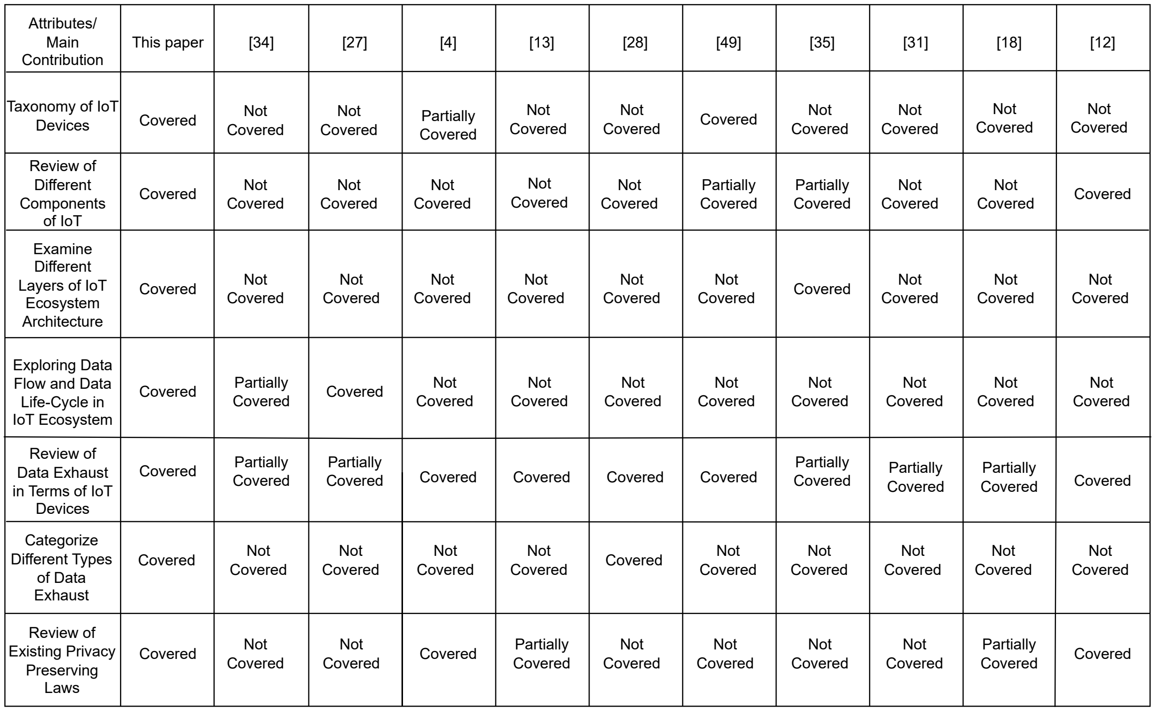
Rutledge et al [57] conducted an exploratory case study of the privacy policy to determine who is collecting what data in the context of IoT devices, focusing their attention on Smart TVs as an example of IoT devices. Using the Goal-Based Requirements Analysis Method (GBRAM) goal mining and refinement, the authors identified 293 privacy-related goal statements for a Samsung SmartTV. These goals were classified according to Anton-Earp's privacy taxonomies and compared with another study of online finance and healthcare websites. They found that almost

90% of the data SmartTV viewers collect is unobservable, which poses a significant privacy vulnerability.

Cunningham et al. [10] examine the impact of legacy privacy laws on the collection and use of data from IoT devices. This study points out that the focus of privacy laws should be shifted from data collection to data usage instead of data collection. For privacy laws to be more effective, they must recognize and address the particular dangers and hazards associated with the use of sensitive information in certain situations.

Bolton et al. introduced PrivExtractor, an awareness dashboard tool[7]. They conducted a comparative analysis of privacy practices from four major vendors, evaluating their compliance with data protection laws. The research findings reveal that these companies usually adhere to legal standards. However, the current regulations are insufficient to ensure transparent disclosure of data practices.

Each of the articles mentioned above covers some of the most pertinent points relating to data exhaust in IoT devices. In this paper, we seek to cover all these points in a comprehensive manner, as there are no articles that address both IoT ecosystem properties and data exhaust in terms of IoT ecosystem extensively in a single publication. For the next steps to be taken in this area, it seems necessary to fill this gap. In Table2.1 you can find a comparison between this work and other related studies.



Attributes/ Main Contribution	This paper	[34]	[27]	[4]	[13]	[28]	[49]	[35]	[31]	[18]	[12]
Taxonomy of IoT Devices	Covered	Not Covered	Not Covered	Partially Covered	Not Covered	Not Covered	Covered	Not Covered	Not Covered	Not Covered	Not Covered
Review of Different Components of IoT	Covered	Not Covered	Not Covered	Not Covered	Not Covered	Not Covered	Partially Covered	Partially Covered	Not Covered	Not Covered	Covered
Examine Different Layers of IoT Ecosystem Architecture	Covered	Not Covered	Not Covered	Not Covered	Not Covered	Not Covered	Not Covered	Covered	Not Covered	Not Covered	Not Covered
Exploring Data Flow and Data Life-Cycle in IoT Ecosystem	Covered	Partially Covered	Covered	Not Covered	Not Covered	Not Covered	Not Covered	Not Covered	Not Covered	Not Covered	Not Covered
Review of Data Exhaust in Terms of IoT Devices	Covered	Partially Covered	Partially Covered	Covered	Covered	Covered	Covered	Partially Covered	Partially Covered	Partially Covered	Covered
Categorize Different Types of Data Exhaust	Covered	Not Covered	Not Covered	Not Covered	Not Covered	Covered	Not Covered	Not Covered	Not Covered	Not Covered	Not Covered
Review of Existing Privacy Preserving Laws	Covered	Not Covered	Not Covered	Covered	Partially Covered	Not Covered	Not Covered	Not Covered	Not Covered	Partially Covered	Covered

Table 2.1: Comparison of this survey with other survey papers

2.2 Background

In this section, we will begin by offering different explanations of IoT to ensure clarity. Moving forward, we will dissect the intricate web of IoT Ecosystem Components, elucidating the diverse elements that collaborate. Understanding these components is pivotal in comprehending the intricate workings of IoT technologies. Subsequently, we will describe the intricacies of IoT Ecosystem Architecture, peeling back the layers to reveal the underlying frameworks that support the functionality of IoT systems. Lastly, we will navigate through the IoT Data Life-cycle, elucidating how data is generated, transmitted, processed, and ultimately utilized within the IoT ecosystem.

2.2.1 Terminology

IoT has been defined from a variety of perspectives in numerous publications published over the past two decades. In this section, we explore some of these definitions to gain a better understanding of the term.

Definition 1: Sensors and actuators are interconnected in an Internet of Things. This includes everything that can be uniquely addressed and visible to the entire globe, including products and physical objects. Using standard communication protocols, machines gather, transmit, and analyze data to make the world smarter, more efficient, and more effective[18].

Definition 2: IoT is a term used to describe extending the Internet into the physical world by creating spatially distributed devices that can sense, track, and act on things[36].

Definition 3: In the IoT, things are connected to each other using unique identifiers, like IP addresses. Rather than requiring human-to-human interaction or human-to-computer interaction, their data can be transferred over a network to provide high-level e-services by gathering and processing information[14].

Definition 4: Using information-sensing devices, the IoT enables devices to identify, operate, and manage themselves via the internet intelligently[62].

Definition 5: An integrated network of interconnected (physical and virtual) things that provides advanced services with existing and evolving technology that's interoperable[51].

Definition 6: The IoT is a network of connected devices that integrates the cyber and physical worlds[63].

Definition 7: The IoT is a concept in which things and people can be connected to anything and anyone at any time, anywhere, using any path or network[54].

2.2.2 IoT Ecosystem Components

Smart devices/sensors: As the main units of an IoT system, smart devices are capable of sensing, monitoring, controlling, and actuating[26]. In fact, one of the reasons that IoT has become drastically popular in recent years is the embedding of sensors and actuators into everyday devices to capture data from the IoT environment (Such as smart watches) and exchange the gathered data with the other components. A smart device can be equipped with a variety of sensors, depending on its functionality.

Table 2.2 summarizes the different types of sensors based on their functional characteristics.

Connectivity: An embedded sensor is called a sensor "node" Although it is relatively straightforward to deploy a single sensor, it is more challenging to ensure

Table 2.2: Different Sensor Types in IoT Devices

Sensor Type	**Functionality**	**Sensor Name**	**Area of Use**
Object Detection	Detecting the presence or distance of nearby objects by emitting electromagnetic radiation. Detecting motion is done by taking continuous screenshots.	Occupancy Sensors, Proximity Sensors, Motion Sensor	-Industry -Smart Home -Healthcare -Agriculture -Smart City
Voice Detection	Detecting human voices. It converts vibrations into audio signals (proportional voltages and currents).	Speech Recognition	-Smart Home
Velocity Meter	Calculation of the rate of change of constant position values. Velocity sensors can be linear or angular.	Gyroscope sensors	-Industry -Smart City
Temperature Meter	Measuring heat energy to detect an individual's body or surroundings.	Temperature Sensors	-Industry -Agriculture -Healthcare
Pressure Meter	Force measurement and conversion into signals.	Pressure Sensors	-Industry -Healthcare
Chemical Meter	Measurement of chemical composition in the environment.	Chemical Sensors, Water Quality Sensors	-Industry -Smart City
Humidity Meter	Measuring and signaling humidity in the environment.	Humidity Sensor	-Smart City - Agriculture
Infrared Meter	Detection of some characteristics of a certain object. Heat emission can also be measured by these sensors.	Infrared Sensor	-Smart Home -Smart City
Optic Meter	Detection of electromagnetic energies.	Optic Sensor	-Industry -Healthcare

connectivity between multiple sensors[54]. IoT devices require specific wireless connectivity technologies based on their type[58]. RFID, Near-field communication (NFC), Wireless Fidelity (Wi-Fi), Zonal Intercommunication Global-standard (ZigBee), Bluetooth, Z-Wave, Thread, and Wireless Sensor Network (WSN) are some of these standards[59]. We can divide these types of technologies according to different features. The key features that may affect the type of connectivity standard in a smart device are Data Rate, Latency, Coverage, Power, Reliability, and Mobility[12].

Edge: The edge serves as a bridge between devices/sensors and the cloud server. The edge is the point of communication for all devices and sensors. For these components, there are two primary functional considerations:

A) Providing local addresses to sensor nodes in wireless personal area networks for short-range communication.

B) Translating between local addresses in wireless personal area networks and IP addresses on the Internet [51].

Cloud: The cloud is the central component of the IoT ecosystem, and it is responsible for accumulating and processing sensor data[64]. Besides offering storage space, cloud servers also provide the infrastructure required for real-time processing and operations. In order to provide the user with the desired response, the data collected in the cloud may be transmitted to specific services[9].

Services: To respond to the user's requests, an IoT device may need to communicate with several parties and service providers. In the course of these interactions, post-processed data may be shared with different parties[60].

2.2.3 IoT Ecosystem Architecture

A three-layer architecture consists of the following components:

Physical layer: This layer consists of sensors, devices, NFC devices, RFID tags, etc. Among the components mentioned above, smart devices and gateways can be considered to be part of this layer.

Network layer: At the top layer, the Network Layer is responsible for the transfer of data collected from the previous layer, initiating the connection between sensors and IoT applications. It includes the connectivity component previously described in this section. Wireless connectivity technologies like Wi-Fi, Signal towers, NFC, Bluetooth, NFC, Zigbee and the like are incorporated into this layer[25].

Application layer: A major responsibility of this layer is to manage users' requests and provide responses from third parties. Requests from users are handled by servers in this layer. In this layer, it is determined whether the IoT ecosystem should be labelled as smart home, smart city, smart healthcare, or another type of IoT ecosystem[25].

To address some of the shortcomings of a three-layer architecture, a five-layer architecture was proposed. Two new layers have been added to this model in addition to the three previously mentioned:

Processing layer: Known as the middleware layer, this layer collects data from the network layer and stores and processes it. It is the cloud processing, the servers, and the information storage at this layer that handle these operations.

Business layer: Besides determining the business mode and data management, this layer is also responsible for managing all aspects of the ecosystem and ensuring the privacy of users[76].

Components of the IoT can be assigned to any of these layers. Smart devices/sensors live under the physical layer; connectivity protocols are located in the network layer. The edge component resides under the processing layer, and the cloud/data center and Services reside under the application layer and business layer [43] [2].

2.2.4 IoT Data Life-cycle

The components listed in section 2.2.2 are intended to capture, communicate, analyze, and act. An IoT ecosystem begins with observation of the environment to *gather data* on a physical phenomenon. Communication technologies are then used to enable the device to be *linked* to other devices or servers. For the purpose of extracting information, all of the gathered data should be *processed and analyzed.* Eventually, this enormous process leads to appropriate *action.*

Considering all phases of data management, the Scenario Agnostic Data Lifecycle Model (SADLM) proposes three main blocks namely, data acquisition, data processing, and data preservation. Further, a detailed breakdown of each block is provided in more detail in each phase[66]. Data Acquisition is the process of collecting data from a variety of sources, evaluating its quality, and tagging it with additional

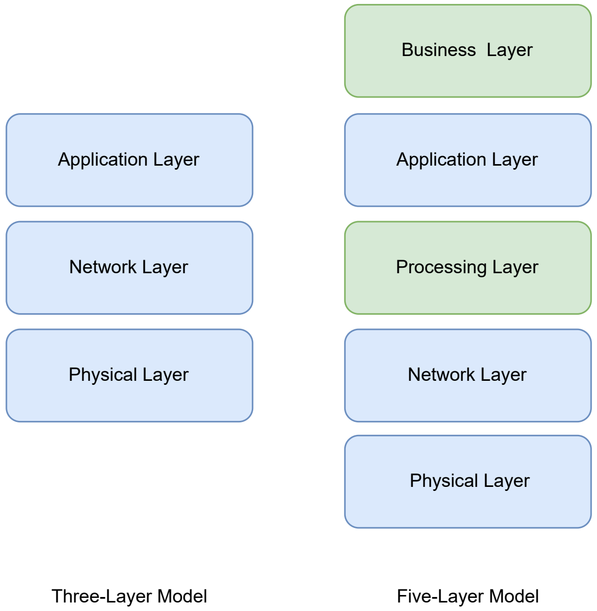


Figure 2.1: IoT Architecture Models

information. Data is generated consecutively or triggered by an external event by IoT sensors. The data generated by sensor networks is not the only source of data. Other sources also provide data streams. Therefore, the raw data generated must be aggregated, warehoused, and streamed at a specified network rate to remote locations for further analysis[27]. Besides sensor data, stored data and activity data are other key sources of data that are aggregated with sensor data. Stored data includes device identifiers and personally identifiable information provided by the user during device activation, activity logs, device state, etc. Activity data, on the other hand, refers to data that describes how a user interacts with a device (e.g., via a mobile device or a button on an IoT device), as well as which functionality has been utilized (e.g., toggling a light)[56].

Once collected, data can be preserved, through the Data Preservation block, or processed, through the Data Processing block. In the Data Preservation block, all data storage and preservation-related tasks are handled. In this step, the data is prepared for further processing or publication. Using sophisticated data analysis techniques, the Data Processing block generates additional value from big data [66].

2.3 Data Exhaust in IoT Devices

It is important to note that IoT systems are fundamentally reliant on data. In order to collect data, smart devices are equipped with sensors. As data is collected, it is transmitted between the components through each of their respective connectivity technologies. At the edge, data is stored and processed locally. All gathered data is aggregated in the cloud and in databases, enabling analysis to be performed. Lastly, services respond to users based on sensor data. In this regard, data is the primary input and output for each component.

We are confronted by real-time, complex and massive streaming data -Big Data- in this ecosystem[17]. *Big Data* is the term applied to massive sets of largely unstructured data that we are able to collect, process, and analyze[55].

In an IoT ecosystem, the collected data can be divided into two categories:
Core Data that are deliberately generated.
Data Exhaust that are unconsciously generated.
In the first group, the outcome is directly related to the operation of the service, whereas in the second group, the outcome is due to the interaction between the user or device over the Internet with other devices[11].

As we move into the all-connected era, there will be a tremendous increase in the amount of data generated, as a consequence, there will be an increase in unwanted data generated as well.

Businesses and third parties analyze these unintended generated data consisting of virtual trails left behind by users to learn more about their behaviours. Telling a meaningful story about users' preferences, data exhaust is a valuable source of information for two purposes. Firstly, targeted advertising and secondly, market research.

In the opinion of businesses and companies, targeted advertisements can be personalized in order to get a better return on investment[38]. With a deep understanding of what is important to you, they may be able to provide you with exactly what you are seeking, for example on social media! Furthermore, the generated data regarding how the application is used may help them to improve their products in later versions and enhance the user experience.

2.3.1 Different Types of IoT Data Exhaust

Devices that are part of the IoT are designed to perform specific functions. There may be additional data generated during the process of executing those specific tasks. The location data generated by your smartphone (IoT device) can be considered unwanted generated data.

As a source of Big Data, data exhaust can be classified according to the type of IoT device. Unstructured and semi-structured data types, including textual, signal/vocal, transactional, pictorial, and positional data[47].

Textual data: Browser-generated data, such as cookies, log files, temporary browsing history, and files.

Signal/Vocal data: Those types of data that are generated by interacting with a virtual assistant.

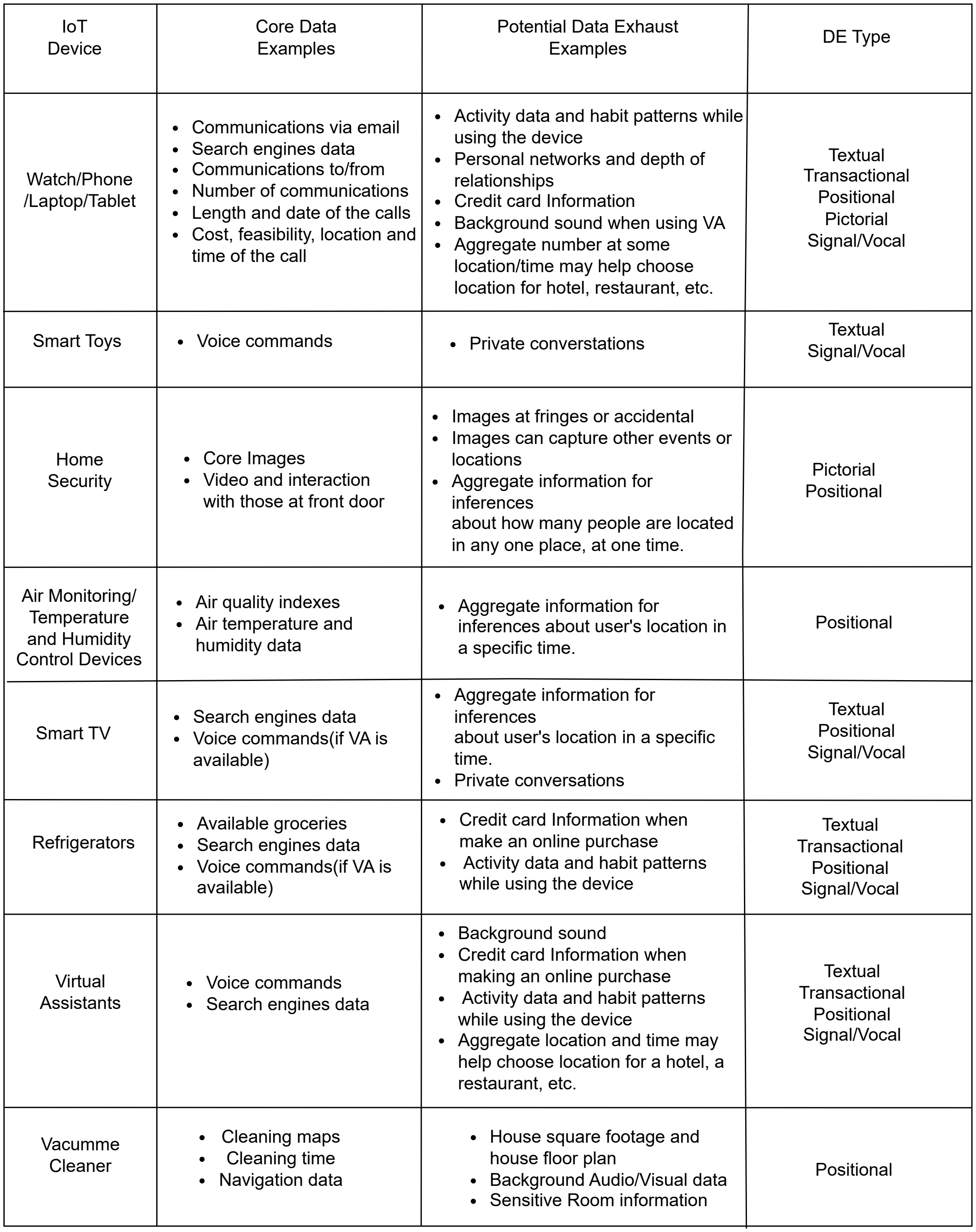
Transactional data: In the course of interacting with a payment application, unwanted data is generated, such as sales orders, invoices, credit card payments, and shipping documents.

Pictorial data: All additional data that may be captured when collecting data from IoT devices equipped with cameras. For example, a security camera may unintentionally collect these data.

Positional data: Any generated data related to the location of the user/device. Several billion mobile phone users around the world have their locations tracked and recorded by mobile phone companies. The users do not voluntarily and continuously log and submit their positional information[10].

2.3.2 Exploring Data Exhaust in Personal IoT Devices

Personal IoT devices refer to a collection of connected devices primarily designed for use in personal settings and within close proximity to an individual[58]. Personal IoT devices, such as smartwatches, smartphones, laptops, tablets, smart homes, and smart toys, have become increasingly popular as they enhance daily life experiences. However, the rise of these devices has led to a need to explore the various types of data exhaust.



IoT Device	Core Data Examples	Potential Data Exhaust Examples	DE Type
Watch/Phone /Laptop/Tablet	• Communications via email • Search engines data • Communications to/from • Number of communications • Length and date of the calls • Cost, feasibility, location and time of the call	• Activity data and habit patterns while using the device • Personal networks and depth of relationships • Credit card Information • Background sound when using VA • Aggregate number at some location/time may help choose location for hotel, restaurant, etc.	Textual Transactional Positional Pictorial Signal/Vocal
Smart Toys	• Voice commands	• Private converstations	Textual Signal/Vocal
Home Security	• Core Images • Video and interaction with those at front door	• Images at fringes or accidental • Images can capture other events or locations • Aggregate information for inferences about how many people are located in any one place, at one time.	Pictorial Positional
Air Monitoring/ Temperature and Humidity Control Devices	• Air quality indexes • Air temperature and humidity data	• Aggregate information for inferences about user's location in a specific time.	Positional
Smart TV	• Search engines data • Voice commands(if VA is available)	• Aggregate information for inferences about user's location in a specific time. • Private conversations	Textual Positional Signal/Vocal
Refrigerators	• Available groceries • Search engines data • Voice commands(if VA is available)	• Credit card Information when make an online purchase • Activity data and habit patterns while using the device	Textual Transactional Positional Signal/Vocal
Virtual Assistants	• Voice commands • Search engines data	• Background sound • Credit card Information when making an online purchase • Activity data and habit patterns while using the device • Aggregate location and time may help choose location for a hotel, a restaurant, etc.	Textual Transactional Positional Signal/Vocal
Vacumme Cleaner	• Cleaning maps • Cleaning time • Navigation data	• House square footage and house floor plan • Background Audio/Visual data • Sensitive Room information	Positional

Table 2.3: Potential DE in different types of PIoT devices

Watch/Phone/Laptop/Tablet: These devices utilize a variety of sensor types, including voice detection, optical meters, and velocity meters, which can provide textual, transactional, positional, and signal/vocal data. Human mobility patterns in urban areas for example, can be analyzed through the analysis of Scenario Agnostic Data Lifecycle Model (CDR) collected from mobile phones. There is usually information about the user's unique ID, a time stamp, and the location of the cell phone tower in these records[72].

Smart Toys: There may be several types of sensors that may be used in this device, including an accelerometer, temperature, voice detection, humidity, and pressure. As a result, possible data output types include textual and signal/vocal data. As an example, Hello Barbie is a smart toy that is capable of collecting, storing, and processing information about children. It was one of the earliest attempts to develop a smart toy. Private conversations were found to be shared by the toy with multiple parties, thereby undermining the authority of parents and potentially impacting a child's trust[23].

Smart Homes: In this category, there are different types of sensors installed in different types of devices, such as: Home security (pictorial data, positional, textual), Air monitoring (textual data), temperature and humidity control devices (textual data), and home appliances. The categories of home appliances can be divided into Smart TVs (textual, positional, signal/vocal data - if VAs are included), VA (positional, signal/vocal, transactional, and textual data), refrigerators (transactional and positional data), and vacuum cleaners (positional data).

For this subcategory, we can refer to different smart devices. Smart security cameras for example, equipped with artificial intelligence analyze human behavior and environmental conditions. Users can also receive smart alerts via their mobile device when certain types of activity are detected. Additionally, these cameras are capable of detecting ambient light and device temperature. Google's NestCam and Amazon's Cloud Cam have not been limited to recording footage of intrusions or carelessness by their owners, as marketing and advertising materials indicate. Instead, these devices are advertised as being capable of capturing personal moments such as pets, children, and strange events. It even has a platform called *Best of Nest* which encourages users to submit their most compelling and entertaining Nest videos. In fact, the website suggests that such videos may even result in the emergence of a newly created, branded subcategory of social video called *Nestie* [52][71].

Another example would be iRomba generating maps of the user's house during the vacuuming process[52]. Roomba can use the collected data to optimize its cleaning patterns by analyzing the layout of the home and the placement of furniture. IRobot assures its users that any data collected will not be shared with third parties without their knowledge, and users can decide whether or not to send their data to the cloud[69].

Based on a study conducted by Nshimba et al.[46] smart devices owners are subject to countless privacy risks. A number of privacy issues are addressed by this study based on the use cases of SAMSUNG, LG, and Sony smart televisions. By taking into account the architecture of IoT devices, the types of data that is collected, vulnerabilities, threats, and policy statements for each television model, we can assign each privacy issue to a specific architecture layer. Listed below are some of these privacy issues:

- Data on view habits of users collected.
- Data collected which can be classified as sensitive.
- A microphone which is present.
- Transmission of data vulnerable to interception.
- Cloud storage of data collected from user.
- Data is shared with external companies.

Another example of a smart device that has received significant usage in the past few years is the personal voice assistant. According to another study conducted by Iqbal et al., [21] a portion of the data collected by Amazon Echo and third parties, such as advertising services, will be used by Amazon for advertising and tracking purposes. The fact that Amazon hosts more than 200k third-party skills can pose a privacy threat to users. Approximately 41 advertisers share their cookies with Amazon, which may contain personal information. Other than these advertisers, 247 other third-party entities, including advertising services, also receive cookies from these advertisers. Nevertheless, according to their study, there appears to be a lack of transparency regarding Amazon's policies and claims concerning its operation practices and third-party skills. It does not appear to be consistent with Amazon's public statements that they infer advertising interests from their users' voice interactions. Over 70 percent of third-party skills do not mention Alexa or Amazon in their privacy policies, and only 2.2 percent provide clear information about their data collection practices.

The examples provided above are only a few examples of cases where unwanted data was generated that might have been collected by the sensor device. Table2.3 summarizes the core data and potential data exhaust for each IoT device discussed above.

2.3.3 Data Exhaust in Voice Assistants

Depending on the type of sensors embedded in the smart device, different types of data exhaust are expected to be generated. In this section, we concentrate on voice assistants, specifically delving into the concept of data exhaust within this context.

The rise of cloud computing, artificial intelligence, and the Internet of Things has led to the increasing popularity of voice assistants in households[65]. Many top technology companies worldwide, including Apple, Amazon, and Google, have created their own VAs such as Siri, Alexa, and Google Assistant. These VAs extend their functionalities through third-party-developed *Skills* allowing users to automate various tasks using voice commands. These skills range from ordering a drink to automating morning routines. VAs are not limited to mobile devices; they have expanded their reach to various platforms such as smart cars, speakers, and televisions[24]. Amazon dominates the market with its Echo products, while Google is expanding its presence with home speakers and integration into various devices. Apple entered the market with HomePod, and Microsoft focuses on integrating Cortana into Windows devices and partnering with other brands[20].

As always listening devices, are activated by a specific wake word such as *Alexa*, *Hey Siri*, or *Okay Google*. Once activated, they interpret user requests and execute them in real-time[34]. By leveraging connections with third-party services and devices, the VA can carry out a range of functions simply by responding to a user's voice prompt. These functions may include responding to queries, playing music, setting alarms or timers, making phone calls or sending messages, completing purchases, offering updates on the weather, and managing other smart devices[49]. Due to the vast computing power and resources available on cloud platforms, voice assistants can rely on cloud-based servers to handle the mentioned complex tasks such as natural language processing, speech recognition, and machine learning algorithms that require intensive computational power[65].

Offering numerous benefits, advantages, and convenience, the global market of smart speakers is expected to gain more global popularity with a market growth from $2.8 billion in 2021 to $11.2 billion by 2026[24]. Globally, there are more than 200 million smart personal assistants installed, and current trends indicate that the number will exceed 500 million by 2030[68].

However, privacy remains a significant concern for consumers, in different aspects. There are two types of privacy threats:

(1) Sensitive information exposure by external attackers.

(2) Privacy disclosures to voice service providers [70].

Many papers have addressed the first issue, and various types of attacks have been previously discussed. Li et al.[31], for example, classified different forms of privacy and security attacks in their study. Regarding the latter one, this paper addresses the relatively under-explored topic of data exhaust.

Data exhaust pertains to the kind of data that is unconsciously generated by the user, as opposed to the intentionally produced core data. To gain a deeper understanding of this matter, it is necessary to examine the flow of data within an IoT ecosystem.

The journey of data within a VA ecosystem can be examined by referencing the architecture model. Following the Five-Layer model, data is collected from the perception layer and transmitted through the network layer, before being processed in the cloud which forms the third layer. In certain cases, the application and business layers may require external services or other parties, which may involve calling upon additional resources[43][66]. Figure 2.2 illustrates this journey in detail.

Businesses and third parties analyze unintentionally generated data, composed of virtual trails left behind by users, to gain insights into user behaviour. Data exhaust transforms user preferences into a meaningful narrative that can be used for two purposes: targeted advertising and market research. As categorized in the previous

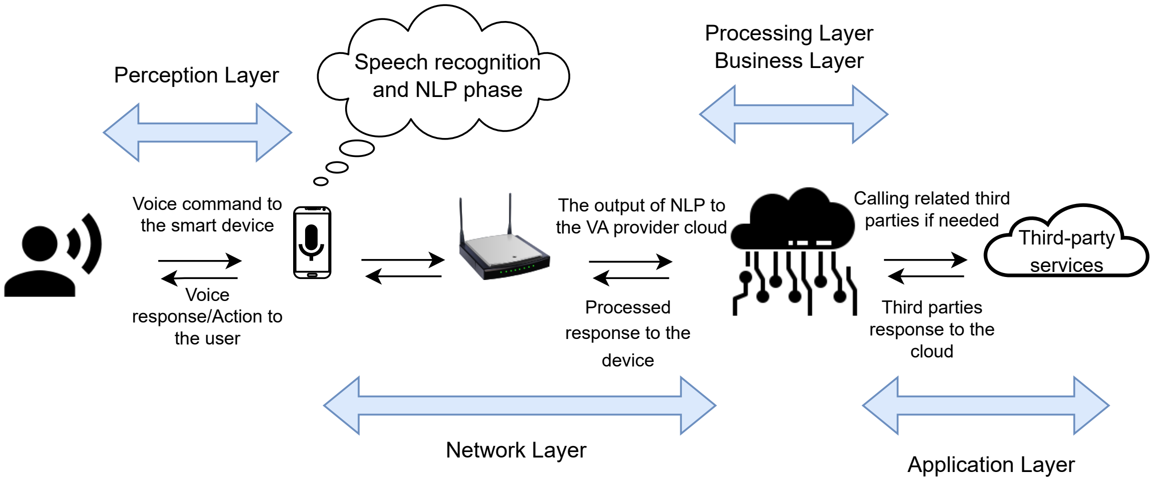

Perception Layer
Speech recognition and NLP phase
Processing Layer
Business Layer
Voice command to the smart device
Voice response/Action to the user
The output of NLP to the VA provider cloud
Processed response to the device
Calling related third parties if needed
Third parties response to the cloud
Third-party services
Network Layer
Application Layer

Figure 2.2: Data flow in Voice Assistant ecosystem

study, based on the sensors integrated into the smart device, VA in this case, the anticipated core data includes voice commands and search engine data. Meanwhile, potential data exhaust may encompass background sound, credit card information, device activity data, habit patterns, and the aggregation of location and time to facilitate locating a specific place.

There needs to be a clear understanding of whether all data generated, transmitted, and shared by the user are what the user intended to generate, transmit, and share. Data exhaust in particular may contain some personal information that the user does not wish to share with third parties. More specifically, a user may not be aware of who is collecting what data while interacting with a virtual assistant, and this can constitute a serious threat to their privacy.

Although microphones are the primary sensors embedded in voice assistants, some smart speakers, such as Amazon Echo Plus, also include temperature sensors, humidity sensors, and ambient light sensors. Furthermore, some newer smart speakers, such as the Amazon Echo Show, feature additional sensors, such as video cameras and motion sensors[35]. Consequently, these smart devices generate a greater volume and type of data. The interaction between a voice assistant and other applications and devices generates various types of data. For example, a Global Positioning System (GPS) may be used by a voice assistant to locate the user when it is used as a navigational aid. Due to this, the data exhaust from a voice assistant may consist of a variety of different types of data, such as audio/vocal signals, text, transactional, positional, etc.

Lack of user awareness regarding the data collected by VAs, both personal and household-related, leads to a limited understanding of data processing and the implications it has on their privacy. As a result, users face challenges in making informed decisions about their privacy. Trust in the vendors rather than informed consent becomes the primary factor influencing the adoption and usage of VAs[53].

In recent years, however, there has been increased public concern about VA security and privacy. Amazon and Google have admitted in news articles that VA recordings are listened to by employees, and VAs have been reported as recording or activating actions without the user's knowledge or consent. As a result of these incidents, there have been concerns raised among the general public regarding privacy[8].

The violations of privacy in VAs can be categorized into four broad categories:

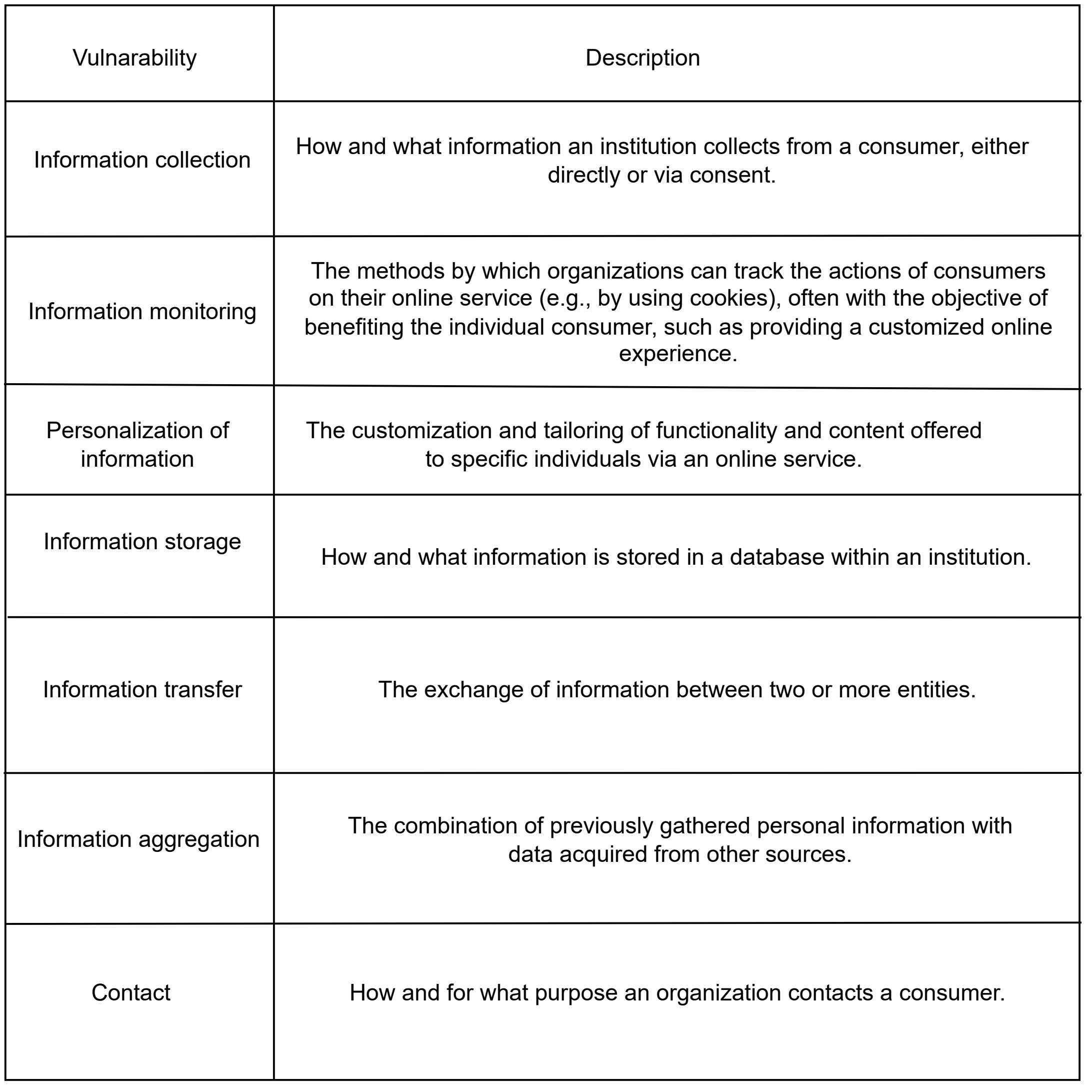
Contextual Privacy: Through the use of virtual assistants, users are exposed to social privacy breaches, such as being spied on and shown their current location without their consent. In cases in which user data is shared with third parties without consent, access-control privacy is violated. Furthermore, ambient privacy and visual privacy may be compromised if VA devices read aloud notifications or messages.

Bystander Privacy: VAs are capable of collecting data not only from legitimate users but also from their surroundings, potentially violating the privacy of third parties. In social privacy situations, voice conversations, for instance, can be recorded without consent. The recording and analysis of private conversations by VA manufacturers constitute a serious violation of speech privacy.

Data Sharing Privacy: Cloud services are often used by VAs, and collected data can be shared with third-party providers. Data sharing privacy can be violated when data is shared without user awareness or consent. It includes privacy related to access control, location privacy, criminal intent privacy when data is shared with malicious third parties, user rights privacy regarding data ownership and control, and user-consent privacy when users agree to terms and conditions without understanding how the data will be collected and utilized.

Environmental Privacy: As the IoT grows, environmental privacy concerns will arise. Surveillance and monitoring programs conducted by the government for national security purposes may violate user rights as well as government privacy. Moreover, VA service providers may track user patterns for business purposes, raising privacy concerns[50].

In light of these violations, seven vulnerability categories can be identified in Table3.2 [57].



Vulnarability	Description
Information collection	How and what information an institution collects from a consumer, either directly or via consent.
Information monitoring	The methods by which organizations can track the actions of consumers on their online service (e.g., by using cookies), often with the objective of benefiting the individual consumer, such as providing a customized online experience.
Personalization of information	The customization and tailoring of functionality and content offered to specific individuals via an online service.
Information storage	How and what information is stored in a database within an institution.
Information transfer	The exchange of information between two or more entities.
Information aggregation	The combination of previously gathered personal information with data acquired from other sources.
Contact	How and for what purpose an organization contacts a consumer.

Table 2.4: Vulnerability of privacy in VAs

2.3.4 Data Profiling and Ad Targeting

IoT device users are at risk of being identified, profiled, and tracked. Tracking and profiling users can also be done using their personal information, such as their name and address. As mentioned before, smart devices may access the user's personal information in many ways. Consequently, it is possible that users' profiles are being developed based on not only their personal information but also long-term activities and behaviours. Then it is critical that users be aware that the collection of their personal data without their consent can result in targeted marketing and the loss of privacy. Overall, the prevalence of identification, profiling, and tracking poses a significant threat to the privacy of users[25]. It is possible to reduce the accuracy of data mining by restricting access to private or personal data, but there is an inherent conflict between privacy and profiling that highlights the risks associated with identification and tracking. Such risks can increase the possibility of profiling and lead to private data leakage through black market data hunting[61]. In conclusion, data collection should be conducted with the consent of the user, and privacy policies should be clearly put in place to ensure that the data is protected.

2.4 Privacy-Preserving Protocols and Laws

In the IoT era, privacy and security issues have become more complex due to the ability to collect personal data from users. In view of the fact that data is collected both actively and passively, a set of protocols and regulations is imperative. Smartphones, smart watches, fitness trackers, and mobile phones are now equipped with more resources between the virtual and physical worlds. Such devices can be used for recording, storing, and processing data pertaining to health, daily routines, and other activities[37]. Mobile phones are capable of recording and transmitting images, sounds, voices, and videos with or without the consent of the user. A growing number of data collection methods have created new privacy concerns, and countermeasures are necessary to ensure the privacy of users[22].

Rutledge et al [56] analyzed 81 devices information exposure. They conducted an experiment to see whether there are unexpected exposures of private and/or sensitive information (e.g., video surreptitiously transmitted by a recording device) or not.

According to a study conducted by Rutledge et al [57] some of the IoT devices may not meet the fair information practices principles recommended by the U.S. Federal Trade Commission (FTC) due to the fact that they do not notify consumers nor collect their consent before collecting data. Physical limitations may also affect their ability to comply with rules and regulations. In that case, it may be unclear for the IoT device users that **who** collects **what data**. Using a Samsung Smart TV as an archetypal example of an IoT device and an exploratory case study of the privacy policy, the study explored how it applies to this device. Their research focused on retrieving Samsung's privacy policies applicable to Smart TVs for analysis through the use of goal-oriented techniques which they applied. According to the paper, from the 77 pieces of information collection and monitoring goals included in the Samsung Smart TV Privacy policies document, 8 (10.4%) could be observed by the user, while 69 (89.6%) could not be observed. Accordingly, most of the data collection and monitoring is not visible to the average viewer. For another instance, Pal et al [50] examined different aspects of privacy in terms of voice assistants. According to this work, in a more sophisticated scenario, once a VA has been activated, the VA not only collects data about the person who activated it but also collects information from background voice conversations with non-users.

Furthermore, due to the rise of machine learning applications, big data analysis is being developed for analyzing the exhaustion of people's daily browsing habits. This is a result of the rise of machine-learning applications. This is a result of the increase in machine-learning applications. In the advertising industry, third-party domains are often connected with publishers' websites, and cookies put unique identifiers on users so that browsing data exhaust can be tracked and used to reconstruct individual browsing histories. Ads are displayed based on the analysis of users' features, including behavioural targeting, frequency capping, re-targeting, and conversion tracking. At the same time, publishers also tailor information to users' conditions and predict requirements based on evaluated preferences that users have never chosen. It is undeniable that the collection of data and its subsequent analysis have benefited both parties. Users benefit from the automatic customization enabled by a wide variety of websites and network services, while publishers earn an increase in revenue of approximately 52 percent when third-party cookies are used. As a result of the trend of collecting exhaust data, the concept of identity tracking has become a major privacy concern[22]. In terms of smart homes, for instance, the White House recently released a report on smart meters and smart homes that outlined both their advantages and disadvantages. It is not only the approximate electricity consumption of residents that smart meters provide information about. It is reported that devices powered by electricity have unique signatures. With the help of this unique signature, some meters are able to distinguish between microwave ovens and refrigerators, or even between a lightbulb in the bathroom and a lightbulb in the dining room. Smart devices can detect when the user is at home, cooking, watching television, or on vacation. An analysis of this information can provide information regarding the wealth, cleanliness, health, and sleeping habits of a resident. Analyzing a person's electrical signal can pinpoint their exact TV show or movie with 96 percent accuracy, according to a study[10].

In 1995, the European Union issued a directive on privacy that seeks to protect personal information by requiring notice and consent from entities collecting data, as well as allowing users to access and correct their data. However, it does not address the IoT or passive data collection and assumes that all personal information is voluntarily provided by users. As far as passive data collection is concerned, the notice and consent requirements are difficult to apply, leaving questions about how they will be applied to new technologies such as cameras and smart meters[10]. Furthermore, concerns exist regarding the ownership and security of these unwanted generated data. Data containing highly sensitive information can be claimed by users, researchers, companies, and academic institutions. In the case of HIoT devices, for example, the storage and security of data is a significant concern due to the ease with which encryption methods can be broken, as well as the potential impact on the confidentiality, safety, and efficacy of clinical care[33].

It is concluded that the privacy of IoT device users is susceptible to various threats, and the existing regulations and policies are insufficient in ensuring the protection of personal information generated during the usage of these devices.

In response to these concerns, the industry has begun to take steps to address them. Google Home, Amazon Alexa, and Siri now support speaker recognition for distinguishing between the speakers in a household. As part of an effort to enhance privacy control, Amazon introduced a command that allows users to delete their recordings. In recent years, projects such as Alias and Mycroft have been developed to give users a greater degree of control over their virtual assistants and to prioritize privacy. By utilizing constant white noise, Alias disables the smart speaker, which can then be reactivated when needed. In order to avoid cloud-based analysis of recordings, Mycroft was designed with privacy in mind[8].

Chapter 3

Predictive Modeling of Data Exhaust in Voice Assistant Network Traffic Data

3.1 Dataset

For the purpose of addressing privacy concerns relating to data exhaust, we are analyzing data related to voice assistant network traffic. In this thesis, network traffic data for Amazon Alexa Echo Dot will be examined. In order to generate the dataset, we used a paper entitled Amazon Alexa traffic traces in 2022, which includes raw Packet Capture (PCAP) files[4]. It contains 150,000 English raw PCAP files that contain all of the network traffic communications between the Amazon Echo Dot and Alexa servers. Their experiment involved setting up a Raspberry Pi as a WiFi hotspot. The Amazon Echo Dot device connects to the WiFi generated by the Raspberry Pi, and all communication between the device and the Alexa Voice Service servers is channelled through the Raspberry Pi's Local Area Network (LAN) port.

Considering that the specific network traffic data described in the referenced paper is only available as PCAP files, our approach encompasses a dual objective. The primary objective was to convert these files into a Comma Separated Value (CSV) format. A comprehensive information extraction process was also undertaken in order to capture a significant amount of data that could be analyzed further. The dataset for our research was generated using a tool named CICFlowMeter[29]. Using the CICFlowMeter, all the PCAP files were processed, and 87 features were extracted into a CSV file.

The study utilizes a conventional definition of a network flow based on a sequence of packets sharing common values for Source IP, Destination IP, Source Port, Destination Port, and Protocol (TCP or UDP). Flows are considered bidirectional, and the research introduces an application called ISCXFlowMeter, written in Java, for flow generation and feature calculation. Unlike some existing tools, ISCXFlowMeter offers greater flexibility in feature selection, addition, and flow timeout control. The tool generates bidirectional flows, distinguishing between forward and reverse directions, especially for time-related statistical features. TCP flows terminate upon connection teardown, while UDP flows terminate based on a flow timeout, a value set arbitrarily by the individual scheme.

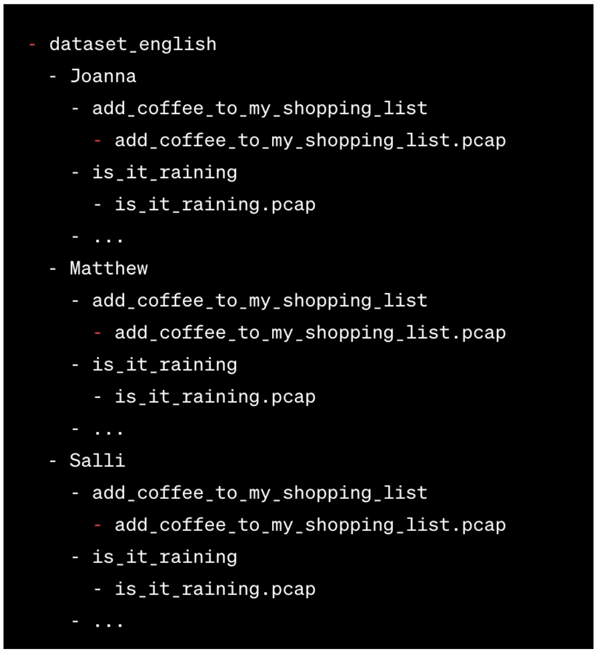
The study explores various flow timeout values, specifically durations of 15, 30, 60, and 120 seconds, studying their impact on classifier accuracy using the same dataset. Experimental results indicate that the maximum accuracy is achieved with a flow timeout of 15 seconds for all classifiers. The classifier response time includes flow time, feature extraction time, and machine learning algorithm time. Notably, the study focuses on time-related features, considering two approaches: measuring time between packets or the duration of flow activity, and fixing time and measuring other variables like bytes or packets per second.

The extracted features are categorized into six groups: "fiat", "biat", and "flowiat" for forward, backward, and bidirectional flows respectively; "idle" and active for states of inactivity and activity; and "psec" for size and number of packets per second. The research emphasizes time-related features to analyze their impact on classifier accuracy[29].

Upon integrating all of the generated CSV files, the dataset was finally ready for analysis and consisted of 380194 rows. Over the period of 334 days and 23 hours, this dataset contains such features as Protocol, Flow Duration, Flow Bytes/s, and so on. There is no label feature in the dataset, making data exhaust detection challenging since there is no ground truth. On the other hand, it is unclear what data is being transmitted over the network due to the encryption of the payload. Therefore, we must analyze the features in order to be able to train a model that can detect data exhaust. Barcel et al. have provided a PCAP file with two distinct features that can be extracted. As well as being organized according to the commander's designations, each PCAP file has also been renamed to correspond to the command issued to the Amazon Alexa Dot.

The provided PCAP files are organized into three distinct folders, with each folder bearing the name of a specific commander. Within each commander's folder, there are 100 subfolders, each renamed to a specific command. Consequently, every PCAP file can be associated with both a commander and a specific command.

Some of the voice commands include: ”Tomorrow will rain?”, ”Wake me up at 9:00 AM”, ”When is Black Friday?”, ”Which is the closest supermarket?”



```
- dataset_english
  - Joanna
    - add_coffee_to_my_shopping_list
      - add_coffee_to_my_shopping_list.pcap
    - is_it_raining
      - is_it_raining.pcap
    - ...
  - Matthew
    - add_coffee_to_my_shopping_list
      - add_coffee_to_my_shopping_list.pcap
    - is_it_raining
      - is_it_raining.pcap
    - ...
  - Salli
    - add_coffee_to_my_shopping_list
      - add_coffee_to_my_shopping_list.pcap
    - is_it_raining
      - is_it_raining.pcap
    - ...
```

Figure 3.1: IoT Architecture Models

Consequently, we systematically generated individualized PCAP files, associating each with the commander’s identity and the specific command directed to the Amazon Alexa Dot device. For every CSV file generated, we incorporate both the commander’s name and the issued command as distinct attributes within the dataset. By leveraging each datapoint voice command text and the commander name, were generated to offer a broader perspective.

In the final dataset, there are a total of 85 columns and 380,194 rows. Most of the features are numerical, except for a few ones such as Flow ID, Src IP, Dst IP, Timestamp, and Label. All the generated features with the related description are gathered in table 3.1. These features were extracted using CICFlowMeter, with the feature label filled as ”No Label” for all rows.

The *Label* column generated by CICFlowMeter serves as a vital element, indicating the classification or categorization assigned to network flows. This categorization discerns whether a particular network flow is benign or potentially malicious. During the analysis process, CICFlowMeter employs algorithms and rules to assess network flows based on their behaviour, patterns, and anomalies. As a result, flows are assigned specific labels, such as *normal*, *suspicious*, or even designated with names corresponding to specific threats[29].

The interpretation of these labels can vary, contingent upon the configuration settings and the dataset used for training the detection system. In our specific dataset, the *Label* column has been filled with *No label*, indicating that no distinct anomalies were identified. Consequently, this feature does not provide meaningful insights for analysis. Hence, it is imperative to exclude this column from further consideration, as it does not contribute to our understanding of the dataset.

Feature Name	**Description**
Flow duration	Duration of the flow in Microsecond
total Fwd Packet	Total packets in the forward direction
total Bwd packets	Total packets in the backward direction
total Length of Fwd Packet	Total size of packet in the forward direction
total Length of Bwd Packet	Total size of packet in backward direction
Fwd Packet Length Min	Minimum size of packet in forward direction
Fwd Packet Length Max	Maximum size of packet in forward direction
Fwd Packet Length Mean	Mean size of packet in forward direction
Fwd Packet Length Std	Standard deviation size of packet in forward direction
Bwd Packet Length Min	Minimum size of packet in backward direction
Bwd Packet Length Max	Maximum size of packet in backward direction

Bwd Packet Length Mean	Mean size of packet in backward direction
Bwd Packet Length Std	Standard deviation size of packet in backward direction
Flow Bytes/s	Number of flow bytes per second
Flow Packets/s	Number of flow packets per second
Flow IAT Mean	Mean time between two packets sent in the flow
Flow IAT Std	Standard deviation time between two packets sent in the flow
Flow IAT Max	Maximum time between two packets sent in the flow
Flow IAT Min	Minimum time between two packets sent in the flow
Fwd IAT Min	Minimum time between two packets sent in the forward direction
Fwd IAT Max	Maximum time between two packets sent in the forward direction
Fwd IAT Mean	Mean time between two packets sent in the forward direction
Fwd IAT Std	Standard deviation time between two packets sent in the forward direction
Fwd IAT Total	Total time between two packets sent in the forward direction
Bwd IAT Min	Minimum time between two packets sent in the backward direction
Bwd IAT Max	Maximum time between two packets sent in the backward direction
Bwd IAT Mean	Mean time between two packets sent in the backward direction
Bwd IAT Std	Standard deviation time between two packets sent in the backward direction
Bwd IAT Total	Total time between two packets sent in the backward direction

Fwd PSH flags	Number of times the PSH flag was set in packets traveling in the forward direction
Bwd PSH Flags	Number of times the PSH flag was set in packets traveling in the backward direction
Fwd URG Flags	Number of times the URG flag was set in packets traveling in the forward direction
Bwd URG Flags	Number of times the URG flag was set in packets traveling in the backward direction
Fwd Header Length	Total bytes used for headers in the forward direction
Bwd Header Length	Total bytes used for headers in the backward direction
FWD Packets/s	Number of forward packets per second
Bwd Packets/s	Number of backward packets per second
Packet Length Min	Minimum length of a packet
Packet Length Max	Maximum length of a packet
Packet Length Mean	Mean length of a packet
Packet Length Std	Standard deviation length of a packet
Packet Length Variance	Variance length of a packet
FIN Flag Count	Number of packets with FIN
SYN Flag Count	Number of packets with SYN
RST Flag Count	Number of packets with RST
PSH Flag Count	Number of packets with PUSH
ACK Flag Count	Number of packets with ACK
URG Flag Count	Number of packets with URG
CWR Flag Count	Number of packets with CWR
ECE Flag Count	Number of packets with ECE
down/Up Ratio	Download and upload ratio
Average Packet Size	Average size of packet
Fwd Segment Size Avg	Average size observed in the forward direction

Bwd Segment Size Avg	Average size observed in the backward direction
Fwd Bytes/Bulk Avg	Average number of bytes bulk rate in the forward direction
Fwd Packet/Bulk Avg	Average number of packets bulk rate in the forward direction
Fwd Bulk Rate Avg	Average number of bulk rate in the forward direction
Bwd Bytes/Bulk Avg	Average number of bytes bulk rate in the backward direction
Bwd Packet/Bulk Avg	Average number of packets bulk rate in the backward direction
Bwd Bulk Rate Avg	Average number of bulk rate in the backward direction
Subflow Fwd Packets	Average number of packets in a sub-flow in the forward direction
Subflow Fwd Bytes	Average number of bytes in a sub-flow in the forward direction
Subflow Bwd Packets	Average number of packets in a sub-flow in the backward direction
Subflow Bwd Bytes	Average number of bytes in a sub-flow in the backward direction
Fwd Init Win bytes	Total number of bytes sent in initial window in the forward direction
Bwd Init Win bytes	Total number of bytes sent in initial window in the backward direction
Fwd Act Data Pkts	Count of packets with at least 1 byte of TCP data payload in the forward direction
Fwd Seg Size Min	Minimum segment size observed in the forward direction
Active Min	Minimum time a flow was active before becoming idle
Active Mean	Mean time a flow was active before becoming idle
Active Max	Maximum time a flow was active before becoming idle

Active Std	Standard deviation time a flow was active before becoming idle
Idle Min	Minimum time a flow was idle before becoming active
Idle Mean	Mean time a flow was idle before becoming active
Idle Max	Maximum time a flow was idle before becoming active
Idle Std	Standard deviation time a flow was idle before becoming active

Table 3.1: List of Features and Descriptions

3.2 Analysis

A comprehensive methodology was employed, consisting of four main phases. These phases, namely data preprocessing, data points classification, false prediction clustering, and model evaluation and visualization, form the foundation of our study. In the following section, we will expand on each of these phases, outlining the specific steps we took in our thesis. This detailed breakdown will help clarify the methods we used and showcase the depth and accuracy of our research approach.

3.2.1 Data Preprocessing

The preprocessing phase plays a foundational role in the entire analysis, facilitating the effective utilization of the dataset. Specifically, within a network, routine activities such as packet retransmission and duplicated Acknowledgment (ACK) can lead to alterations in the distribution of packet-level features[13]. Despite collecting traffic data for the network, it may contain null and irrelevant values, particularly due to the conversion of files into CSV format using the CICFlowMeter tool, which introduces potential noise that must be addressed. Thus, essential preprocessing steps have been undertaken to address these issues. During this phase, duplicate or irrelevant data must be removed from the collected data by cleaning, filtering, and aggregating it. The dataset underwent several data cleaning procedures, wherein null, unnecessary, and duplicate values were removed, along with features exhibiting zero variance.

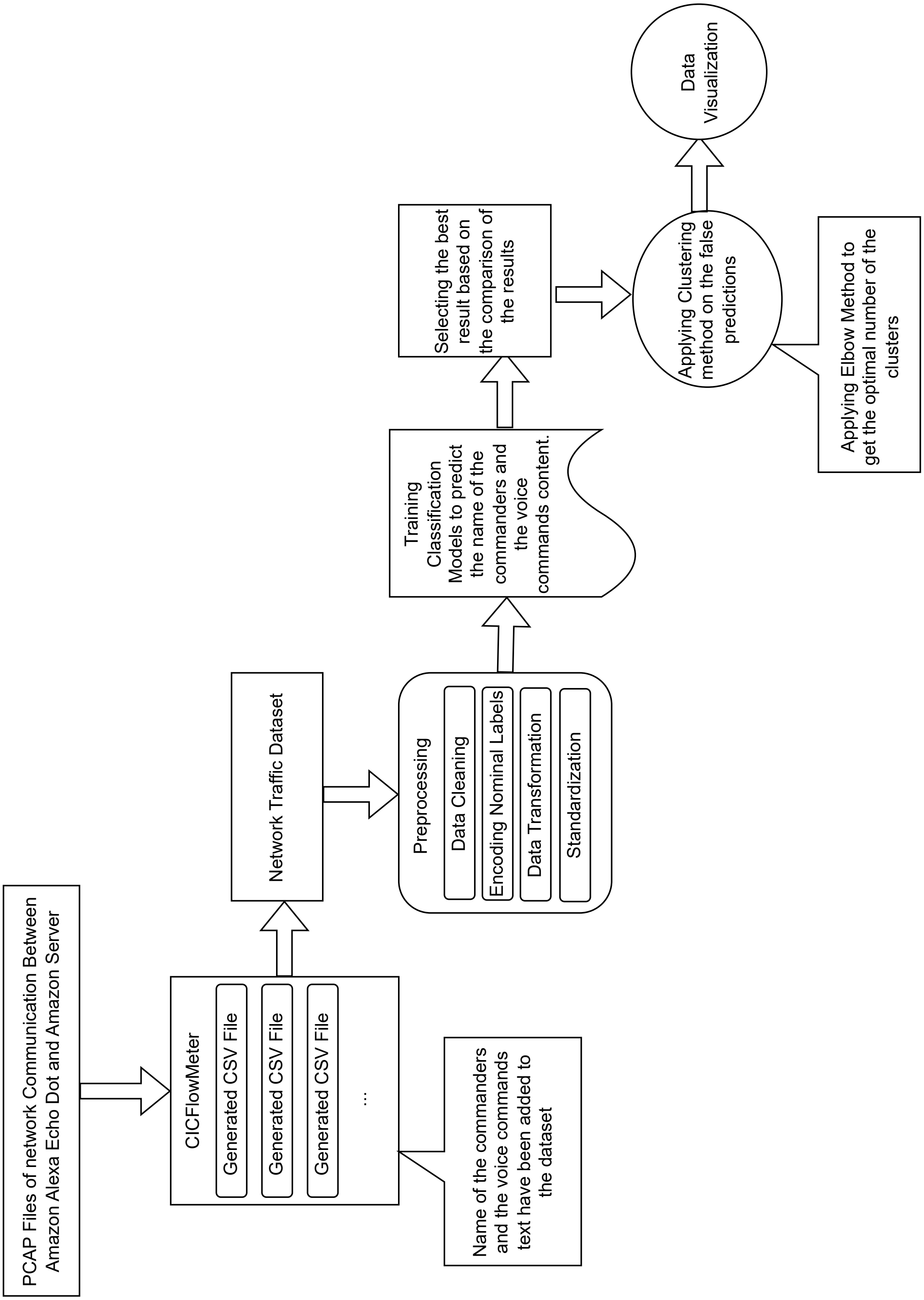

PCAP Files of network Communication Between Amazon Alexa Echo Dot and Amazon Server
CICFlowMeter
Generated CSV File
Generated CSV File
Generated CSV File
...
Name of the commanders and the voice commands text have been added to the dataset
Network Traffic Dataset
Preprocessing
Data Cleaning
Encoding Nominal Labels
Data Transformation
Standardization
Training Classification Models to predict the name of the commanders and the voice commands content.
Selecting the best result based on the comparison of the results
Applying Clustering method on the false predictions
Applying Elbow Method to get the optimal number of the clusters
Data Visualization

Figure 3.2: Analysis Workflow

A data validation step was performed to ensure IP addresses followed the expected format. Subsequently, rows with both destination and source IPs containing valid IP addresses were retained.

Various data transformation steps were applied, including converting "timestamps" into the "datetime" datatype to enable meaningful analysis. Additionally, certain features were treated as objects, despite representing numerical values, necessitating conversion to their corresponding float values.

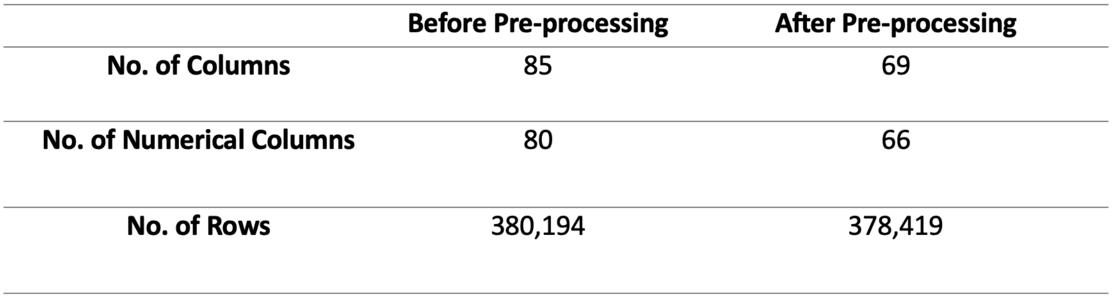
As a critical step, standardization was employed using the StandardScaler to bring certain features to a consistent scale.

Removing zero variance columns can also help to reduce the dimensionality of the data, which can improve the performance of machine learning models by reducing the risk of overfitting and speeding up the training process. As a result, we removed all the features with zero variance which includes 15 features. All the removed features are "Bwd PSH Flags", "Fwd URG Flags", "Bwd URG Flags", "URG Flag Cnt", "CWE Flag Count", "ECE Flag Cnt", "Fwd Byts/b Avg", "Fwd Pkts/b Avg", "Fwd Blk Rate Avg", "Bwd Byts/b Avg", "Subflow Bwd Pkts", "Active Mean", "Active Std", "Active Max", "Active Min", and "Label".

Following the applied preprocessing techniques, our dataset comprises 69 numerical attributes, with the exception of the *protocol*, *commander*, and *voice command* variables. These encompass three distinct protocols, three individual commanders, and 105 unique voice commands. With the utilization of label encoding, the dataset is now primed for subsequent analysis. Upon successfully applying these preprocessing steps, the dataset has been rendered clean and ready for subsequent analytical processes.

3.2.2 Datapoints Classification

We created labels for the dataset by introducing two new features called *Commander* and *Voice Command*. Our strategy involves training classification methods on the dataset based on the mentioned labels while making certain assumptions. Specifically, we assume that any data points that are incorrectly identified with respect to the two labels indicate a potential data exhaust. Data points that exhibit incorrect classification for both commanders and voice commands could suggest two potential



	Before Pre-processing	After Pre-processing
No. of Columns	85	69
No. of Numerical Columns	80	66
No. of Rows	380,194	378,419

Table 3.2: Dataset Features Before and After Preprocessing

scenarios: either the voice network traffic data wasn't generated due to the user's intent, or the transmitted data doesn't correspond accurately to the original voice command.

Regarding the selection of classification methodologies, an assessment of various classifiers was undertaken. However, certain classifiers were excluded from consideration due to their incompatibility with the dataset's characteristics and the nature of the classification aim. Specifically, logistic regression and naive Bayes were deemed unsuitable as the task at hand doesn't align with binary classification paradigms. Additionally, the application of Support Vector Machine (SVM) proved sub-optimal due to challenges posed by the dataset's high dimensionality, resulting in extended computation times.

Subsequently, the decision was made to employ the K-Nearest Neighbor(K-NN), Random Forest and Multi-Layer Perceptron (MLP) classifiers.

For all three methods, the dataset was stratified into two distinct labels: *voice command* and *commander*. Accordingly, two separate models were developed to predict the corresponding labels.

K-NN is a simple and intuitive machine learning algorithm used for both classification and regression tasks. Leveraging the K-NN algorithm, which is a non-parametric and instance-based machine learning technique, this thesis employed a systematic approach[16]. By considering the K nearest neighbours of each network packet within the multidimensional feature space, the K-NN algorithm made classifications based on the majority voting principle among these neighbours, taking into account their similarities and differences. The K-NN algorithm doesn't require consideration of the relationship between historical data and easily operates in complex environments. This simplicity makes it suitable for application in traffic information management systems[73].

Minkowski distance, a widely used metric in traditional K-NN, expresses actual distances between points. In this context, the exponent is set to 2, representing Euclidean distance. The shorter the distance, the greater the similarity[32].

The choice of an appropriate K value was pivotal in ensuring the accuracy of the classifications. Cross-validation accuracy is a statistical measure used in machine learning to assess the performance and generalizability of a predictive model. In the context of this thesis, it refers to the accuracy score obtained through cross-validation techniques. Cross-validation involves partitioning the dataset into subsets, training the model on some of these subsets, and evaluating its performance on the remaining subsets. This process is repeated multiple times, and the average accuracy score across all iterations provides a reliable estimate of the model's predictive performance on unseen data.

In the context of this thesis, a cross-validation function was applied to determine the optimal value for K in the K-NN algorithm. For each k value in the specified range, a KNN classifier is instantiated with Manhattan distance as the metric for proximity measurement and optimal parallelization using multiple processors. Subsequently, a cross-validation procedure is executed with five folds on a subset of the data. The accuracy scores obtained from each fold are computed, and the mean accuracy score is calculated as it is plotted in Figure 3.3 and Figure 3.4. The result obtained from the cross-validation process guided the selection of the most suitable K value for classifying network traffic data, predicting both *commander* and *voice command.* By employing this approach, the study ensured that the K-NN algorithm was fine-tuned and optimized to make accurate predictions on new, unseen data, thereby enhancing the reliability and effectiveness of the classification process.

Random Forest is an ensemble learning method widely used for both classification and regression tasks. A combination of decision trees, constructed using the bagging method, employs both random data selection and feature selection. As an ensemble

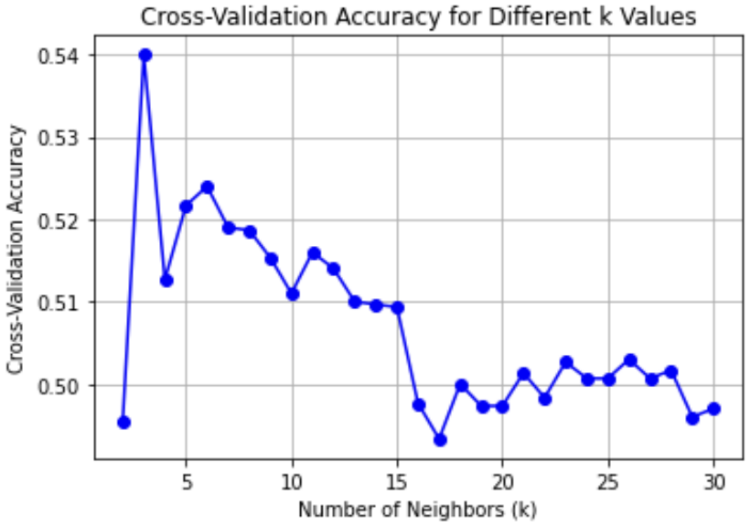


Figure 3.3: Cross Validation - Commander Classification

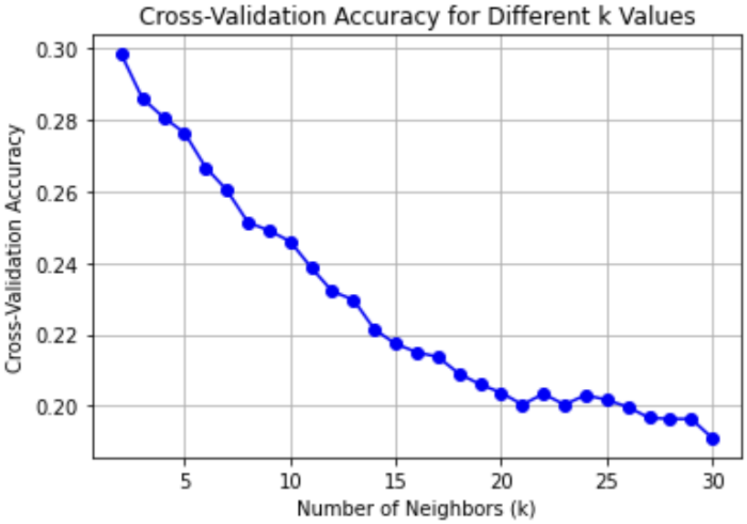


Figure 3.4: Cross Validation - Voice Command Classification

technique, Random Forest utilizes bootstrapping, averaging, and bagging to train multiple decision trees. This integration of decision trees enhances overall accuracy and stability in predictions. Random Forest builds a "forest" of decision trees, each trained with different subsets of randomly selected data and features, ensuring greater generalization for diverse scenarios. The sub-datasets, constructed through random sampling with replacement, are used to build sub-decision trees. Each sub-decision tree produces a result, and the final classification is determined through majority voting among these sub-trees.[5]

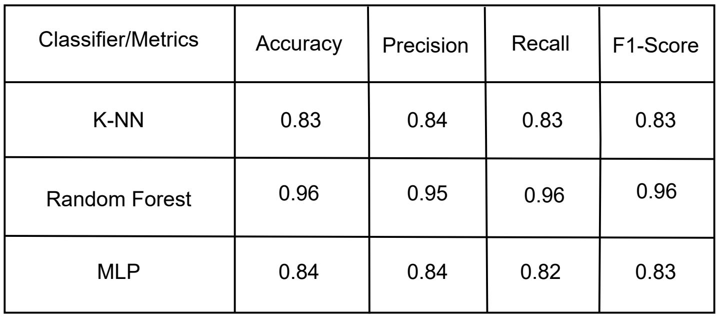
Through the use of distinct subsets of available features, multiple independent decision trees are simultaneously constructed on different segments of the training samples. Bootstrapping ensures the uniqueness of each decision tree in the Random Forest, reducing variance. Once all decision trees are built, they collectively make predictions on new, unseen data. For classification tasks, each tree "votes" for a class, and for regression tasks, they provide individual predictions. The final prediction is determined by aggregating the votes (classification) or averaging the predictions (regression). The Random Forest classifier amalgamates the decisions of multiple trees for the final judgment, resulting in strong generalization. This classifier consistently aims to outperform various other classification techniques in terms of precision, overcoming challenges associated with imbalanced datasets and overfitting. This method allows for highly accurate and stable predictions, crucial in the context of voice commander detection within network traffic data analysis [75][3]. By implementing the randomization of both data and features, our aim is to assess whether this method outperforms other classification techniques.

MLP is a type of feed-forward Artificial Neural Network (ANN) that falls under the broader category of deep learning with three or more layers. MLP is specifically designed to map input data to appropriate outputs through layers of simple neurons, also called perceptrons. These networks feature hidden layers situated between inputs and outputs, aiding the model in achieving deeper learning. Each perceptron computes a single output by combining multiple real-valued inputs using weighted sums and applying a nonlinear activation function. These networks find widespread application in supervised learning scenarios, where both training and testing datasets are essential components for training and evaluating the model, respectively[44]. MLP employs a backpropagation algorithm, adjusting connection weights based on discrepancies between expected and actual outputs[42]. MLP can learn both linear and non-linear functions, approximating any continuous function and solving problems that are not linearly separable. It also learns tasks from training data, minimizing the loss function to achieve optimality and reducing loss to an acceptable level. Deep MLPs, with multiple hidden layers, can represent certain functions more efficiently than shallow ones, showcasing the capacity to compute complex functions like the parity function with a linear-sized network[42].

In the context of classifying network traffic data into categories such as *commander* and *voice command*, this thesis utilized MLP as the final classifier in this stage of analysis, emphasizing its significance in accurately distinguishing and categorizing various network interactions.

The outcomes of our comparative analysis pertaining to the performance of these three classification models across the distinct labels are presented in the ensuing Figure3.5 and Figure3.6. According to these tables an accuracy of 0.96 for the Random Forest model when predicting the *Commander* target feature, and an accuracy of 0.85 for the same model when predicting the *Voice Command* target feature.

As anticipated, the Random Forest outperformed the other model, aligning with its reputation as the preeminent classification model for network traffic data.

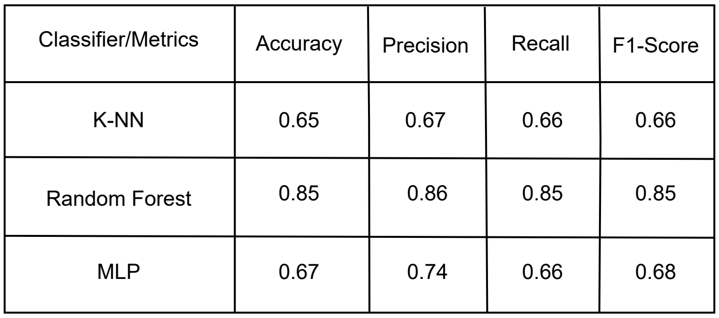


Classifier/Metrics	Accuracy	Precision	Recall	F1-Score
K-NN	0.83	0.84	0.83	0.83
Random Forest	0.96	0.95	0.96	0.96
MLP	0.84	0.84	0.82	0.83

Figure 3.5: Commander Classification Comparison Report

3.2.3 False Prediction Clustering

Following the application of the classification model to the initial dataset, all incorrectly predicted rows were integrated into a single data frame, comprising a total of



Classifier/Metrics	Accuracy	Precision	Recall	F1-Score
K-NN	0.65	0.67	0.66	0.66
Random Forest	0.85	0.86	0.85	0.85
MLP	0.67	0.74	0.66	0.68

Figure 3.6: Voice Command Classification Comparison Report

12,118 rows among all 380,194 features. For this new dataset containing the model's erroneous predictions, we employed a clustering model to categorize distinct types of data points.

K-Means is a widely used centroid-based clustering method known for its simplicity and efficiency. It is especially suitable for handling numerous variables. The process involves choosing the desired number of clusters (k) and calculating centroids for each group by iteratively assigning data points to clusters. This iteration continues until no further changes occur in the clusters, or until a predefined stopping criterion is satisfied[41].

Utilizing the k-means clustering technique necessitates the identification of the ideal number of clusters. In pursuit of this objective, we employed the Elbow Method on the recently generated dataset.

The Elbow method is commonly used to determine the best number of clusters in a dataset by examining the sum of the Sum of Squared Errors (SSE) among various groups. A noticeable change in SSE values, leading to a distinctive bend in the graph, indicates the ideal cluster count. The Elbow method calculates the squared distances between each cluster element and its centroid. The optimal number of clusters is identified through a significant shift in the Within Cluster Sum of Squares (WCSS) value for different setups, resembling an angle[19].

As illustrated in Figure3.7, the discerned optimal cluster count is 3. In accordance with this determination, we proceeded to execute the k-means clustering algorithm with 3 clusters.

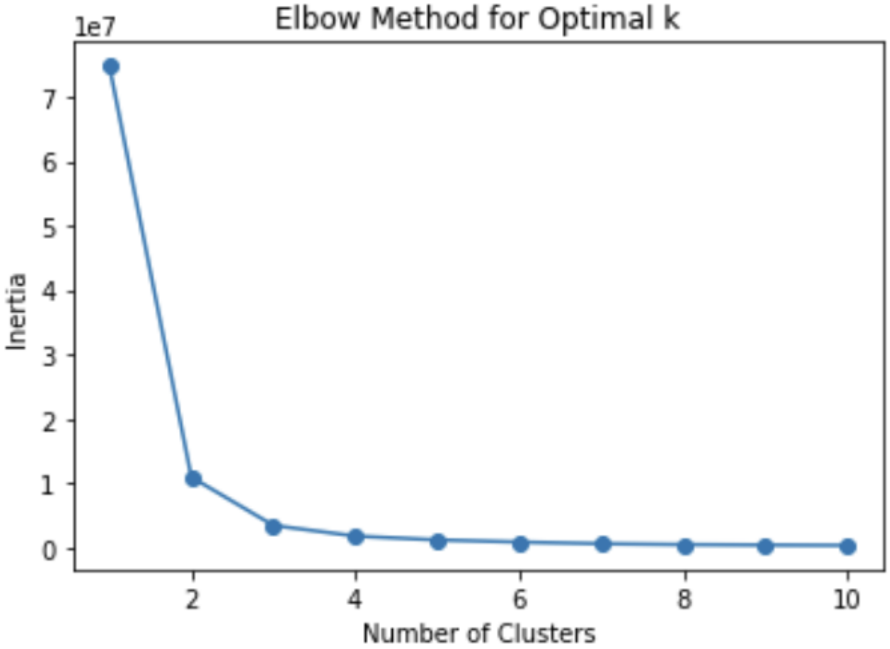


Figure 3.7: Elbow Method for Optimal k

3.2.4 Model Evaluation and Visualization

In the realm of data analysis, visualization serves as a powerful tool, enabling researchers to delve deeper into complex datasets and discern underlying structures and relationships. In this thesis, where the dataset is notably intricate, comprising a multitude of features, the challenge lies in unravelling meaningful insights from this wealth of information. T-distributed Stochastic Neighbor Embedding (t-SNE), a sophisticated method for reducing dimensionality, proved invaluable in addressing this specific challenge. Unlike some traditional methods, t-SNE excels in preserving both local relationships (the proximity of data points in the original high-dimensional space) and global structures (the overall layout and clustering patterns) of the data[6].

With 69 features to contend with, applying t-SNE became pivotal. By transforming the dataset into a lower-dimensional representation, we effectively unlocked the ability to visualize the data in a more intuitive and interpretable manner. This transformation paved the way for the creation of a three-dimensional plot, an invaluable visual representation that condensed the complexity of the dataset into a comprehensible form. The resulting visualization, as showcased in Figure 3.8, not only provided a visually striking representation of the clustered data but also served as a lens through which intricate patterns, previously obscured by the dataset's complexity, became apparent.

This application of t-SNE didn't just simplify the data; it enhanced our understanding of the inherent structures within the dataset. By revealing clusters and groupings that might not have been immediately evident in the high-dimensional space, t-SNE empowered us to make more informed interpretations and draw insightful conclusions about the underlying relationships among the data points. This visualization not only bolstered the depth of our analysis but also significantly contributed to the overall richness and comprehensibility of our research findings.

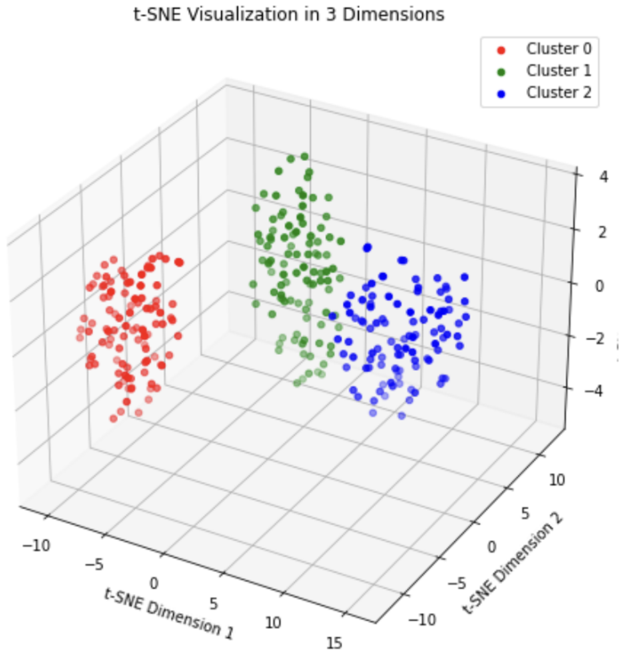


Figure 3.8: Clusters of false voice commands and commanders predictions

3.2.5 Results

As you may have noticed there are three different issues regarding data exhaust in an IoT ecosystem. First, unwanted data generating, second, passive data collection, and finally data ownership and data use. We need to specifically be clear about each issue and discuss solutions for each of them separately. Depending on the type of sensors they may use in their smart device, businesses and applications may generate more data than you expect. For instance, once a VA has been activated, the VA not only collects data about the person who activated and is supposed to command it but also collects information from background voice conversations like two other people's conversation[50]. The embedded sensors generate data in response to environmental events, and it is almost impossible to determine whether the data should be generated or not. In other words, the sensor cannot determine if the user aimed to generate specific data consciously or unconsciously. In this scenario, we cannot prevent sensors from generating data exhaust. Therefore, we may have to seek a solution in two other steps. The physical layer is responsible for generating data in an IoT ecosystem. Once the device generates data, it decides what data will be collected and transmitted via the network layer. As a result, it may be possible to distinguish between generated data and data that was not supposed to be generated. Depending on the type of data package, for example, a specific size may be expected, it may be considered data exhaust if any additional data is provided. For this scenario, we must design and implement our IoT ecosystem in order to collect the necessary data and prevent it from collecting unnecessary raw data. This solution targets the first layer of a five-layer IoT ecosystem architecture.

Regarding the second issue, passively collected data may be employed for a wide variety of purposes in both healthcare and PIoT devices like smartwatches, such as early symptom recognition, postoperative monitoring, clinical research, basic science, and public health. Passive data offers a continuous and quantifiable insight into a user's health and lifestyle. The remarkable thing about these devices is that they can be used to collect and analyze passive information, such as GPS and accelerometer data, to provide real-time insight into human behaviour without requiring participants to take part [33]. In this situation, the user is unaware of what is being transmitted in the first place, raising concerns about privacy issues. As we are referring to the transmission of data, this issue is the responsibility of the physical layer. It is the nature of these types of devices to passively collect data; we cannot prevent them from doing so, but the user should be informed of what is being transmitted while the device is in use.

However, the third issue concerns the businesses and third parties, which are currently suffering from a lack of clear and strict regulations. Data ownership and data usage in the context of IoT devices need to be clarified. As a result, both the

Application Layer and the Business Layer are responsible for addressing this issue.

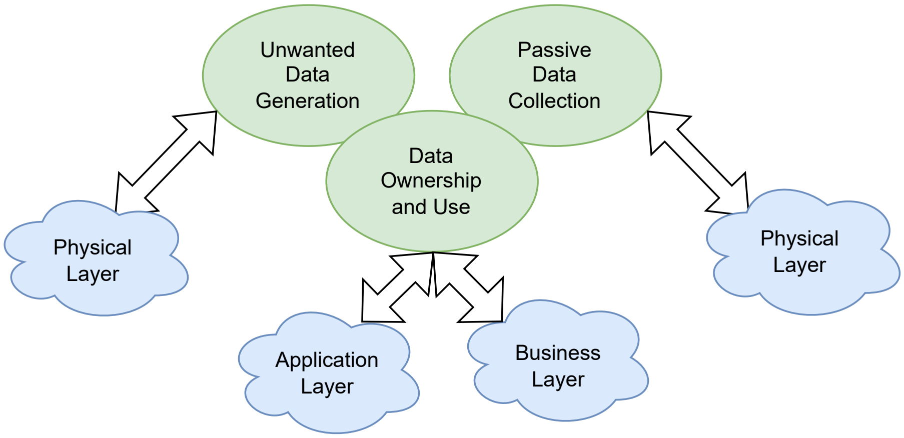


Figure 3.9: Data Exhaust Issues In An IoT Ecosystem

In a study conducted by Iqbal et al. [21], a solution was proposed to address this concern. In order to measure the amount of data collected, its usage, and its sharing by smart speaker platforms, they developed an auditing framework that leverages online advertising. In order to evaluate their framework, they looked at Amazon's smart speaker ecosystem and according to their findings, the privacy policies of Amazon and Skills did not clearly disclose their practices regarding data collection.

Meanwhile, MySudo proposes a way to reduce the risk of data exhaust during online interactions[40]. MySudo offers the ability for users to establish separate profiles, known as Sudo profiles, which each feature a unique telephone number, email, private web browser, and virtual card. These profiles can be assigned specific functions, such as shopping, socializing, classified sales, or booking services. By utilizing Sudo information instead of personal information, users can avoid creating digital footprints that could lead to their identification and reveal their private actions.

As evidenced by the depiction in Figure 6, our analysis reveals the presence of three distinct clusters corresponding to instances of erroneous predictions. Guided by our initial hypothesis, we assert that one of these clusters pertains to what we define as *data exhaust*. Notably, in the figure, Cluster 0 is represented by the red data points and stands out as markedly dissimilar from the other two clusters. This differentiation positions Cluster 0 as an outlier within the context of the data distribution. Further exploration of this cluster's characteristics, particularly pertaining to the properties of its features, reinforces our hypothesis. Consistently across the data points within this cluster, there are discernible patterns in the shared features that lend credence to the validity of our hypothesis.

Chapter 4

Conclusion

Some users are skeptical about VAs due to privacy concerns. They worry that a microphone-based device might be exploited by malicious entities to invade their homes, leading to doubts about the device's purpose and intentions[30].

The adoption of VAs as everyday gadgets highlights a notable power imbalance between users and the corporations behind these technologies. Despite being promoted as convenient tools, VAs operate by relaying user commands to company servers, where extensive data processing occurs, adhering to regulations such as GDPR. While users are aware of data mining, their understanding of how personal data influences online advertising remains limited. Concerns have arisen due to the methods of data collection, emphasizing the necessity for clearer explanations from vendors[7]. However, current data protection laws have limitations in ensuring a comprehensive privacy statement. Users often underestimate privacy risks due to incomplete comprehension, resulting in varying levels of acceptance regarding privacy loss. Furthermore, existing privacy controls are not fully utilized, highlighting the disconnect between user needs and the reliance on VA companies for privacy safeguards.

As evidenced in a study conducted by Bolton et al. [7], there are numerous ambiguous aspects in the provided privacy policies concerning the collection and utilization of user-generated data in almost most of the VAs.

This thesis aims to create a user-oriented monitoring system. It achieves this by training a model to predict network traffic data generated from data exhaust. We delved into a rich dataset comprising network traffic data, aiming to unravel the intricacies of voice command interactions. To start, we implemented a model training process, equipping the system to predict both the commander responsible for initiating the command and the specific voice command associated with each data point. This analysis results in finding instances of misclassifications within the dataset. These misclassifications hinted at a significant issue – there was a possibility that either the command was not genuinely generated by the assigned commander, or the uttered voice command did not accurately correspond to the intended action.

Identifying these misclassifications marked a crucial step in our investigation. To tackle this challenge, we analyzed the discrepancies. In an effort to discern patterns and clusters within these misclassifications, we employed the K-means clustering technique. By grouping these discrepancies into distinct clusters, we aimed to gain a deeper understanding of the underlying issues. One of the key outcomes of our analysis was the identification of a specific cluster, which we designated as the potential data exhaust. This cluster indicated instances where either the commands were not authentic, or the voice commands were mismatched with the intended actions.

This approach can significantly enhance how users perceive transparency, which is crucial in today's user-centric environments. Since many users tend to accept privacy policies without close scrutiny, this solution empowers them to keep a watchful eye on the integrity of transmitted data.

VAs offer convenience, but they also bring up privacy worries due to their always-on microphones[30][4]. These worries can be mitigated using the solution proposed in this thesis.

Using this solution will always help users to know if there is any data transmitted even when they are not calling them.

Smart speaker manufacturers should prioritize data minimization to restrict the volume and sensitivity of the information they collect. It's vital to avoid collecting unnecessary or irrelevant data, as this could jeopardize user privacy. Implementing data exhaust control measures can help mitigate the risks of data leaks or unauthorized access.

4.1 Future Work

In this study, we used a dataset proposed by Barcel et al. There we 105 different voice commands used in this dataset. For future works, we recommend adopting the method proposed by Barcel et al. to collect our unique network traffic data [4]. This involves examining various scenarios, including experimenting with more sensitive voice commands like making a purchase or setting up the experiment in an environment including background sounds.

Additionally, it's essential to compare different Voice Assistants and assess the amount of unwanted data they collect to better understand privacy preservation.

In the study conducted by Bolton et al. [7], an extensive analysis of privacy and policy texts across various brands of VAs was conducted. The researchers meticulously scrutinized these documents, evaluating the transparency levels in each brand's privacy text. This thorough examination allowed them to discern the varying degrees of openness and clarity present in the privacy policies of different VAs.

Building on this groundwork, we can further extend the research scope by applying the proposed solution to detect data exhaust to multiple brands of VAs. By employing the same methodology, we aim to assess the reliability and the extent to which these VAs adhere to their stated privacy policy texts. This comparative analysis becomes invaluable in unravelling the nuances of how different VAs prioritize transparency and user privacy. Through this comprehensive evaluation, we endeavour to shed light on the reliability of these technologies and the authenticity of their privacy claims, contributing to a deeper understanding of the landscape of VA platforms in terms of user data protection and transparency.

In other words, while our current study focused solely on the Alexa Echo Dot, future research may explore other types of VA to provide a broader perspective on data privacy and collection practices.

Appendix A

List of Publications

- Mellaty, Mahdieh; Sampalli, Srini; Zincir-Heywood, Nur; de Snayer, Kevin; Dougall, Terri; A Systematic Review of Data Exhaust in IoT Devices, ACM Computing Surveys, 2023, manuscript under review.

Bibliography

[1] Abbas Acar, Hossein Fereidooni, Tigist Abera, Amit Kumar Sikder, Markus Miettinen, Hidayet Aksu, Mauro Conti, Ahmad-Reza Sadeghi, and Selcuk Uluagac. Peek-a-boo: I see your smart home activities, even encrypted! In *Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, pages 207–218, 2020.

[2] Sadiq Ur Rehman Aqeel-ur Rehman, Iqbal Uddin Khan, Muzaffar Moiz, and Sarmad Hasan. Security and privacy issues in iot. *International Journal of Communication Networks and Information Security (IJCNIS)*, 8(3):147–157, 2016.

[3] Amit Kumar Balyan, Sachin Ahuja, Umesh Kumar Lilhore, Sanjeev Kumar Sharma, Poongodi Manoharan, Abeer D Algarni, Hela Elmannai, and Kaamran Raahemifar. A hybrid intrusion detection model using ega-pso and improved random forest method. *Sensors*, 22(16):5986, 2022.

[4] R. Barceló-Armada, I. Castell-Uroz, and P. Barlet-Ros. Amazon alexa traffic traces. *Computer Networks*, 205:108782, 2022.

[5] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

[6] Adrien Bibal, Valentin Delchevalerie, and Benoît Frénay. Dt-sne: T-sne discrete visualizations as decision tree structures. *Neurocomputing*, 529:101–112, 2023.

[7] Tom Bolton, Tooska Dargahi, Sana Belguith, and Carsten Maple. Privextractor: Toward redressing the imbalance of understanding between virtual assistant users and vendors. *ACM Transactions on Privacy and Security*, 26(3):1–29, 2023.

[8] P. Cheng and U. Roedig. Personal voice assistant security and privacy—a survey. *Proceedings of the IEEE*, 110(4):476–507, 2022.

[9] Peng Cheng, Ibrahim Ethem Bagci, Jeff Yan, and Utz Roedig. Smart speaker privacy control-acoustic tagging for personal voice assistants. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 144–149. IEEE, 2019.

[10] McKay Cunningham. Next generation privacy: The internet of things, data exhaust, and reforming regulation by risk of harm. *Groningen Journal of International Law*, 2, 2014.

[11] Anas Dakkak, Hongyi Zhang, David Issa Mattos, Jan Bosch, and Helena Hölmstrom Olsson. Towards continuous data collection from in-service

products: Exploring the relation between data dimensions and collection challenges. In *2021 28th Asia-Pacific Software Engineering Conference (APSEC)*, pages 243–252. IEEE, 2021.

[12] Jie Ding, Mahyar Nemati, Chathurika Ranaweera, and Jinho Choi. Iot connectivity technologies and applications: A survey. *arXiv preprint arXiv:2002.12646*, 2020.

[13] R. Dubin, A. Dvir, O. Pele, and O. Hadar. I know what you saw last minute—encrypted http adaptive video streaming title classification. *IEEE Transactions on Information Forensics and Security*, 12(12):3039–3049, 2017.

[14] Mehdi Gheisari, Hamid Esmaeili Najafabadi, Jafar A Alzubi, Jiechao Gao, Guojun Wang, Aaqif Afzaal Abbasi, and Aniello Castiglione. Obpp: An ontology-based framework for privacy-preserving in iot-based smart city. *Future Generation Computer Systems*, 123:1–13, 2021.

[15] Fengxian Guo, F Richard Yu, Heli Zhang, Xi Li, Hong Ji, and Victor CM Leung. Enabling massive iot toward 6g: A comprehensive survey. *IEEE Internet of Things Journal*, 8(15):11891–11915, 2021.

[16] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer, 2003.

[17] Manjul Gupta and Joey F George. Toward the development of a big data analytics capability. *Information & Management*, 53(8):1049–1064, 2016.

[18] Nastaran Hajiheydari, Mojtaba Talafidaryani, and SeyedHossein Khabiri. Iot big data value map: how to generate value from iot data. In *Proceedings of the 2019 the 5th international conference on e-society, e-learning and e-technologies*, pages 98–103, 2019.

[19] Muhammad Hamka and Ngatik Ramdhoni. K-means cluster optimization for potentiality student grouping using elbow method. In *AIP Conference Proceedings*, volume 2578. AIP Publishing, 2022.

[20] Matthew B Hoy. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.

[21] Umar Iqbal, Pouneh Nikkhah Bahrami, Rahmadi Trimananda, Hao Cui, Alexander Gamero-Garrido, Daniel Dubois, David Choffnes, Athina Markopoulou, Franziska Roesner, and Zubair Shafiq. Your echos are heard: Tracking, profiling, and ad targeting in the amazon smart speaker ecosystem. *arXiv preprint arXiv:2204.10920*, 2022.

[22] Yinhao Jiang, Ba Dung Le, Tanveer Zia, and Praveen Gauravaram. Privacy concerns raised by pervasive user data collection from cyberspace and their countermeasures. *arXiv preprint arXiv:2202.04313*, 2022.

[23] Meg Leta Jones and Kevin Meurer. Can (and should) hello barbie keep a secret? In *2016 IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS)*, pages 1–6. IEEE, 2016.

[24] W. Kang and B. Shao. The impact of voice assistants' intelligent attributes on consumer well-being: Findings from pls-sem and fsqca. *Journal of Retailing and Consumer Services*, 70:103130, 2023.

[25] Ashwin Karale. The challenges of iot addressing security, ethics, privacy, and laws. *Internet of Things*, 15:100420, 2021.

[26] Wafa'a Kassab and Khalid A Darabkh. A–z survey of internet of things: Architectures, protocols, applications, recent advances, future directions and recommendations. *Journal of Network and Computer Applications*, 163:102663, 2020.

[27] Rajalakshmi Krishnamurthi, Adarsh Kumar, Dhanalekshmi Gopinathan, Anand Nayyar, and Basit Qureshi. An overview of iot sensor data processing, fusion, and analysis techniques. *Sensors*, 20(21):6076, 2020.

[28] Michael Kubach and Heiko Roßnagel. Smart assistants in it security–an approach to addressing the challenge by leveraging assistants' specific features. In *HCI for Cybersecurity, Privacy and Trust: Second International Conference, HCI-CPT 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, pages 575–587. Springer, 2020.

[29] A. H. Lashkari, G. D. Gil, M. S. Mamun, and A. A. Ghorbani. Characterization of tor traffic using time based features. In *International Conference on Information Systems Security and Privacy*, volume 2, pages 253–262. SciTePress, 2017.

[30] J. Lau, B. Zimmerman, and F. Schaub. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–31, 2018.

[31] J. Li, L. Pan, M. R. Azghadi, H. Ghodosi, J. Zhang, et al. Security and privacy problems in voice assistant applications: A survey. *arXiv preprint arXiv:2304.09486*, 2023.

[32] Chencheng Ma, Xuehui Du, and Lifeng Cao. Improved knn algorithm for fine-grained classification of encrypted network flow. *Electronics*, 9(2):324, 2020.

[33] Nicole A Maher, Joeky T Senders, Alexander FC Hulsbergen, Nayan Lamba, Michael Parker, Jukka-Pekka Onnela, Annelien L Bredenoord, Timothy R Smith, and Marike LD Broekman. Passive data collection and use in healthcare: A

systematic review of ethical issues. *International Journal of Medical Informatics*, 129:242–247, 2019.

[34] G. McLean and K. Osei-Frimpong. Hey alexa… examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior*, 99:28–37, 2019.

[35] A. H. Mhaidli, M. K. Venkatesh, Y. Zou, F. Schaub, and M. Kandadai. Listen only when spoken to: Interpersonal communication cues as smart speaker privacy controls. *Proc. Priv. Enhancing Technol.*, 2020(2):251–270, 2020.

[36] Daniele Miorandi, Sabrina Sicari, Francesco De Pellegrini, and Imrich Chlamtac. Internet of things: Vision, applications and research challenges. *Ad hoc networks*, 10(7):1497–1516, 2012.

[37] Subhra Shriti Mishra and Akhtar Rasool. Iot health care monitoring and tracking: A survey. In *2019 3rd international conference on trends in electronics and informatics (ICOEI)*, pages 1052–1057. IEEE, 2019.

[38] Timothy Morey, Theodore Forbath, and Allison Schoop. Customer data: Designing for transparency and trust. *Harvard Business Review*, 93(5):96–105, 2015.

[39] Steve Morgan. The world will store 200 zettabytes of data by 2025. `https://cybersecurityventures.com/the-world-will-store-200-zettabytes-of-data-by-2025/`, June 2020.

[40] mysudo. What is digital exhaust and why does it matter? `https://mysudo.com/2020/08/what-is-digital-exhaust-and-why-does-it-matter/`, August 2020.

[41] P. Nagaraj, S. Selva Birunda, R. Venkatesh, V. Muneeswaran, S. Krishna Narayanan, U. Dhannu Shree, and B. Sunethra. Automatic and adaptive segmentation of customer in r framework using k-means clustering technique. In *2022 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–5. IEEE, 2022.

[42] J Naskath, G Sivakamasundari, and A Alif Siddiqua Begum. A study on different deep learning algorithms used in deep neural nets: Mlp som and dbn. *Wireless Personal Communications*, 128(4):2913–2936, 2023.

[43] D. Navani, S. Jain, and M. S. Nehra. The internet of things (iot): A study of architectural elements. In *2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 473–478. IEEE, 2017.

[44] Ali Yadavar Nikravesh, Samuel A Ajila, Chung-Horng Lung, and Wayne Ding. Mobile network traffic prediction using mlp, mlpwd, and svm. In *2016 IEEE International Congress on Big Data (BigData Congress)*, pages 402–409. IEEE, 2016.

[45] Sandro Nižetić, Petar Šolić, Diego López-de-Ipiña González-de, Luigi Patrono, et al. Internet of things (iot): Opportunities, issues and challenges towards a smart and sustainable future. *Journal of Cleaner Production*, 274:122877, 2020.

[46] Kt Nshimba and Roelien Goede. An architecture approach to a secure home area network. In *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pages 1–6. IEEE, 2022.

[47] Daniel OLeary and Veda C Storey. Data exhaust: Life cycle, framework and a case study of stolen911. com. 2017.

[48] Daniel E O'Leary and Veda C Storey. Discovering and transforming exhaust data to realize managerial value. *Available at SSRN 3746010*, 2020.

[49] D. A. Orr and L. Sanchez. Alexa, did you get that? determining the evidentiary value of data stored by the amazon® echo. *Digit. Investig.*, 24:72–78, 2018.

[50] Debajyoti Pal, Chonlameth Arpnikanondt, Mohammad Abdur Razzaque, and Suree Funilkul. To trust or not-trust: privacy issues with voice assistants. *IT Professional*, 22(5):46–53, 2020.

[51] Urmila A Patil, Mithra Venkatesan, and Shashikant Prasad. An improved wireless network architecture for iot in hospital healthcare. In *2021 IEEE Bombay Section Signature Conference (IBSSC)*, pages 1–6. IEEE, 2021.

[52] James Pierce. Smart home security cameras and shifting lines of creepiness: A design-led inquiry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[53] D. Pins, T. Jakobi, A. Boden, F. Alizadeh, and V. Wulf. Alexa, we need to talk: A data literacy approach on voice assistants. In *Designing Interactive Systems Conference 2021*, pages 495–507, 2021.

[54] R Porkodi and V Bhuvaneswari. The internet of things (iot) applications and communication enabling technology standards: An overview. In *2014 International conference on intelligent computing applications*, pages 324–329. IEEE, 2014.

[55] Junaid Qadir, Anwaar Ali, Andrej Zwitter, Arjuna Sathiaseelan, Jon Crowcroft, et al. Crisis analytics: big data-driven crisis response. *Journal of International Humanitarian Action*, 1(1):1–21, 2016.

[56] Jingjing Ren, Daniel J Dubois, David Choffnes, Anna Maria Mandalari, Roman Kolcun, and Hamed Haddadi. Information exposure from consumer iot devices: A multidimensional, network-informed measurement approach. In *Proceedings of the Internet Measurement Conference*, pages 267–279, 2019.

[57] Richard L Rutledge, Aaron K Massey, and Annie I Antón. Privacy impacts of iot devices: A smarttv case study. In *2016 IEEE 24th International Requirements Engineering Conference Workshops (REW)*, pages 261–270. IEEE, 2016.

[58] Biswa PS Sahoo, Saraju P Mohanty, Deepak Puthal, and Prashant Pillai. Personal internet of things (piot): What is it exactly? *IEEE Consumer Electronics Magazine*, 10(6):58–60, 2021.

[59] S Sujin Issac Samuel. A review of connectivity challenges in iot-smart home. In *2016 3rd MEC International conference on big data and smart city (ICBDSC)*, pages 1–4. IEEE, 2016.

[60] Deepti Sehrawat and Nasib Singh Gill. Smart sensors: Analysis of different types of iot sensors. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 523–528. IEEE, 2019.

[61] Mohamed Seliem, Khalid Elgazzar, and Kasem Khalil. Towards privacy preserving iot environments: a survey. *Wireless Communications and Mobile Computing*, 2018:1–15, 2018.

[62] Yizhou Shen, Shigen Shen, Qi Li, Haiping Zhou, Zongda Wu, and Youyang Qu. Evolutionary privacy-preserving learning strategies for edge-based iot data sharing schemes. *Digital Communications and Networks*, 2022.

[63] Chanyang Shin, Prerit Chandok, Ran Liu, Seth James Nielson, and Timothy R Leschke. Potential forensic analysis of iot data: an overview of the state-of-the-art and future possibilities. 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)*, pages 705–710. IEEE, 2017.

[64] Biljana L Risteska Stojkoska and Kire V Trivodaliev. A review of internet of things for smart home: Challenges and solutions. *Journal of cleaner production*, 140:1454–1464, 2017.

[65] G. Terzopoulos and M. Satratzemi. Voice assistants and smart speakers in everyday life and in education. *Informatics in Education*, 19(3):473–490, 2020.

[66] Sergio Trilles, Alberto González-Pérez, and Joaquín Huerta. An iot platform based on microservices and serverless paradigms for smart farming purposes. *Sensors*, 20(8):2418, 2020.

[67] Lionel Sujay Vailshery. Number of internet of things (iot) connected devices worldwide from 2019 to 2021, with forecasts from 2022 to 2030. `https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/`, August 2022.

[68] C. Valero, J. Pérez, S. Solera-Cotanilla, M. Vega-Barbas, G. Suarez-Tangil, M. Alvarez-Campana, and G. López. Analysis of security and data control in smart personal assistants from the user's perspective. *Future Generation Computer Systems*, 144:12–23, 2023.

[69] Joel Varghese and Thaier Hayajneh. A framework to identify security and privacy issues of smart home devices. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 135–143. IEEE, 2018.

[70] C. Wang, S. Kennedy, H. Li, K. Hudson, G. Atluri, X. Wei, W. Sun, and B. Wang. Fingerprinting encrypted voice traffic on smart speakers with deep learning. In *Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, pages 254–265, 2020.

[71] Carla White and James N Gilmore. Imagining the thoughtful home: Google nest and logics of domestic recording. *Critical Studies in Media Communication*, pages 1–14, 2022.

[72] Takahiro Yabe, Nicholas KW Jones, P Suresh C Rao, Marta C Gonzalez, and Satish V Ukkusuri. Mobile phone location data for disasters: A review from natural hazards and epidemics. *Computers, Environment and Urban Systems*, 94:101777, 2022.

[73] Lijin Yang, Qing Yang, Yonghua Li, and Yuqing Feng. K-nearest neighbor model based short-term traffic flow prediction method. In *2019 18th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)*, pages 27–30. IEEE, 2019.

[74] Naqliyah Zainuddin, Maslina Daud, Sabariah Ahmad, Mayasarah Maslizan, and Syafiqa Anneisa Leng Abdullah. A study on privacy issues in internet of things (iot). In *2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP)*, pages 96–100. IEEE, 2021.

[75] Yubo Zhai and Xianghan Zheng. Random forest based traffic classification method in sdn. In *2018 international conference on cloud computing, big data and blockchain (ICCBB)*, pages 1–5. IEEE, 2018.

[76] Ian Zhou, Imran Makhdoom, Negin Shariati, Muhammad Ahmad Raza, Rasool Keshavarz, Justin Lipman, Mehran Abolhasan, and Abbas Jamalipour. Internet of things 2.0: Concepts, applications, and future directions. *IEEE Access*, 9:70961–71012, 2021.