Computing Surveys

A Systematic Review of Data Exhaust in IoT Devices

Journal:	*Computing Surveys*
Manuscript ID	Draft
Paper:	Short Survey Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Mellaty, Mahdieh; Dalhousie University, Faculty of Computer Science Sampalli, Srini; Dalhousie University, Faculty of Computer Science Zincir-Heywood, Nur; Dalhousie University, Faculty of Computer Science de Snayer, Kevin; Calian Ltd. Dougall, Terri; Calian Ltd.
Computing Classification Systems:	Security and Privacy, Privacy Protection, Information Flow Control, Networks, Sensors and actuators, Sensor Devices and Platforms

SCHOLARONE™
Manuscripts

A Systematic Review of Data Exhaust in IoT Devices

MAHDIEH MELLATY, Dalhousie University, Canada
SRINIVAS SAMPALLI, Dalhousie University, Canada
NUR ZINCIR-HEYWOOD, Dalhousie University, Canada
KEVIN DE SNAYER, Calian Ltd, Canada
TERRI DOUGALL, Calian Ltd, Canada

The rapid expansion of Internet of Things (IoT) technology and the exponential growth of connected devices has resulted in a vast amount of generated data, expected to exceed 200 zettabytes by 2025. Data, both intentionally generated and unintentionally produced, called Data Exhaust, holds substantial value for businesses and third parties. However, the collection and analysis of such data raise privacy concerns, particularly when it contains sensitive information. The study aims to address the gap in understanding data exhaust in various IoT devices and to explore key aspects related to IoT data exhaust, considering the associated privacy and security risks.

CCS Concepts: • **Security and privacy** → **Privacy protections**; **Information flow control**; • **Networks** → *Network components*; • **General and reference** → **Surveys and overviews**; • **Hardware** → **Sensors and actuators**; **Sensor devices and platforms**.

Additional Key Words and Phrases: IoT, data exhaust, data collection, privacy

ACM Reference Format:
Mahdieh Mellaty, Srinivas Sampalli, Nur Zincir-Heywood, Kevin de Snayer, and Terri Dougall. 2022. A Systematic Review of Data Exhaust in IoT Devices. In . ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

For Peer Review

1 INTRODUCTION

Internet of Things (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 [34]. 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 [9]. 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 [22][45]. 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. Unintentionally generated data can provide valuable insights, enabling a comprehensive understanding of users and customers. By analyzing collected data using data mining techniques [11], extract valuable information. However, the data generated, stored, and transmitted by sensors and actuators may contain sensitive personal information, raising privacy concerns[1].

In light of this perspective, it becomes clear that privacy and security risks are a serious concern for IoT devices when they collect user data and store it in the cloud. Ultimately, the process of sensing and collecting data through sensors

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Association for Computing Machinery.
Manuscript submitted to ACM

and actuators, storing it in the cloud, and analyzing it with machine learning methods can result in data breaches and privacy violations.

The goal of this survey paper is to fill a gap in the existing discussion regarding data exhaust in different types of IoT devices. It provides a comprehensive insight into the IoT ecosystem by identifying the different types of data exhaust, with a particular emphasis on personal IoT devices. It considers privacy preserving laws and regulations, and identified potential solutions.

The remainder of this paper is organized as follows: Section 2 describes the most significant contributions to this work. In Section 3, an overview of IoT terminology, IoT ecosystem components and different architecture models for IoT devices, and the data life-cycle in an IoT ecosystem is presented. The concept of Data Exhaust is discussed and expanded within the context of the Internet of Things in section 4. Moreover, in this section, different types of data exhaust will be classified and these categories will be explored for various types of IoT devices. At the end of this section, we will discuss the use of data exhaust. In section 5, we discuss the privacy concerns associated with IoT devices for consumers. In section 6, a discussion of potential solutions to the problem of IoT data exhaust is provided. Towards the end of this document, section 7 provides an insight into future work.

Figure 1 provides a general overview of the paper's flow and may assist the reader with a clearer understanding of the paper's focus.

For P

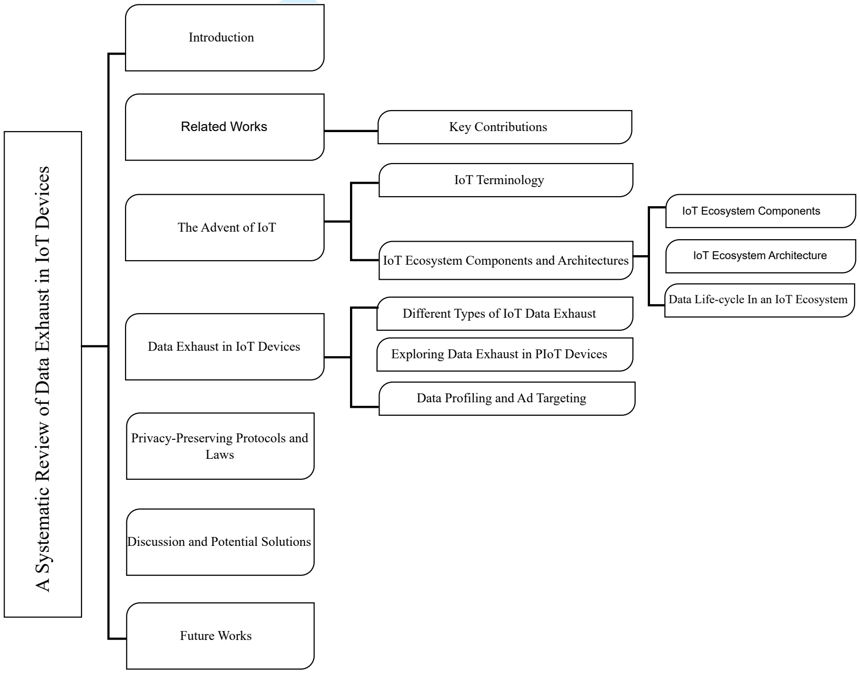

A Systematic Review of Data Exhaust in IoT Devices
Introduction
Related Works
Key Contributions
The Advent of IoT
IoT Terminology
IoT Ecosystem Components and Architectures
IoT Ecosystem Components
IoT Ecosystem Architecture
Data Life-cycle In an IoT Ecosystem
Data Exhaust in IoT Devices
Different Types of IoT Data Exhaust
Exploring Data Exhaust in PIoT Devices
Data Profiling and Ad Targeting
Privacy-Preserving Protocols and Laws
Discussion and Potential Solutions
Future Works

Fig. 1. Flow of the Paper

2 RELATED WORKS

The objective of this survey 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 regards to the data collection in these devices, it is important to understand their architecture models and the general data flow. The papers present 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 [12] 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. [13] 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.[49] 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. [28] 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. [34] 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 behavior from audio and video recording devices. In addition, they identify several cases in which exposure varies depending on the device location.

Pierce et al. [31] 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 [18] also examined ethical concerns as well as solutions related to each ethical concerns such as passive data collection, secondary data use, and storage of passive data.

O'Leary et al. [27] 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 potentials 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 [35] 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 GBRAM goal mining and refinement, the authors identified 293 privacy-related goal statements for a Samsung SmartTV. These

For Peer Review

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. [4] 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.

2.1 Key Contributions

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 table1 you can find a comparison between this work and other related studies.

For

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 1. Comparison of this paper with other papers

3 THE ADVENT OF IOT

There has been a revolution in the world when considering Wireless Sensor Networks (WSNs) and Radio Frequency Identification (RFIDs) since 1999. It was as a result of these innovations that the 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, RFID tags, and readers, the system is able to enable both physical and virtual interactions with the surrounding environment. By 2011, interconnected systems outnumbered humans, and in 2012 there were over nine billion devices that were connected[40]. It is predicted that there will be 55.7 billion connected devices by the end of the next five years, of which 75 percent will be the Internet of Things, as we move towards a smart and connected world[7].

3.1 IoT 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[11].

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[19].

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[8].

Definition 4: Using information-sensing devices, the IoT enables devices to identify, operate, and manage themselves via the internet intelligently[41].

Definition 5: An integrated network of interconnected (physical and virtual) things that provides advanced services with existing and evolving technology that's interoperable[30].

Definition 6: The IoT is a network of connected devices that integrates the cyber and physical worlds[42].

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[32].

3.2 IoT Ecosystem Components and Architecture

In the context of IoT, we refer to a system. As in any other system, IoT systems are surrounded by and influenced by their environment. In a well-designed IoT ecosystem, everyone performs certain functions based on rules to accomplish predetermined goals. The system consists of smart devices, sensors, connectivity, gateways, cloud, and databases. Putting these components together determines IoT architecture. Depending on how complex the IoT ecosystem is, the IoT architecture may vary. A brief description of each element is followed by a description of two architectural models.[25]

3.2.1 IoT ecosystem Components:

Smart devices/sensors: As the main units of an IoT system, smart devices are capable of sensing, monitoring, controlling, and actuating[16]. 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.

For Peer Review

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 connectivity between multiple sensors[32]. IoT devices require specific wireless connectivity technologies based on their type[36]. RFID, NFC, Wi-Fi, ZigBee, Bluetooth, Z-Wave, Thread, and WSN are some of these standards[37]. 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[6].

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 [30].

Cloud: The cloud is the central component of the IoT ecosystem, and it is responsible for accumulating and processing sensor data[43]. 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[3].

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[38].

For Peer Review

3.2.2 *IoT Ecosystem Architectures:*

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[15].

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 labeled as smart home, smart city, smart healthcare, or another type of IoT ecosystem[15].

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.

Components of the Internet of Things 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 [24] [2].

3.3 Data Life-cycle in an IoT Ecosystem

The components listed in section 3.1.1 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

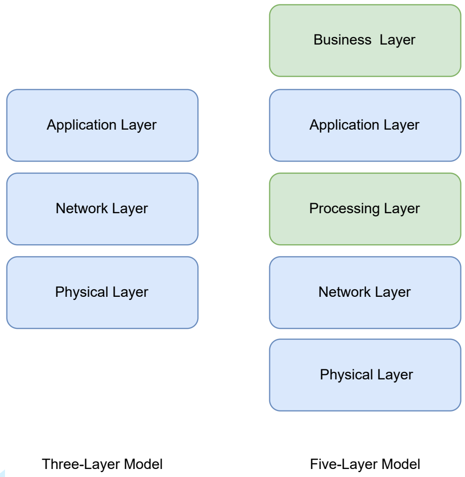


Fig. 2. IoT Architecture Models

For Peer Review

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 Comprehensive Scenario Agnostic Data Lifecycle Model proposes three main blocks namely, Data Acquisition, Data Processing, and Data Preservation. Furthur, a detailed breakdown of each block is provided in more detail in each phase[44]. Data Acquisition is the process of collecting data from a variety of sources, evaluating its quality, and tagging it with additional 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[17]. 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)[34].

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 [44].

4 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[10]. "Big Data" is the term applied to massive sets of largely unstructured data that we are able to collect, process, and analyze[33].

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[5].

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 behaviors. 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[21]. 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.

4.1 Different Types of IoT Data Exhaust

Devices that are part of the Internet of Things 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[27].

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[4].

4.2 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."[36] 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.

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 call detail records (CDRs) collected from mobile

For Peer Review

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[48].

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[14].

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), Virtual Assistants (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* [31][47].

Another example would be iRomba generating maps of the user's house during the vacuuming process[31]. 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[46].

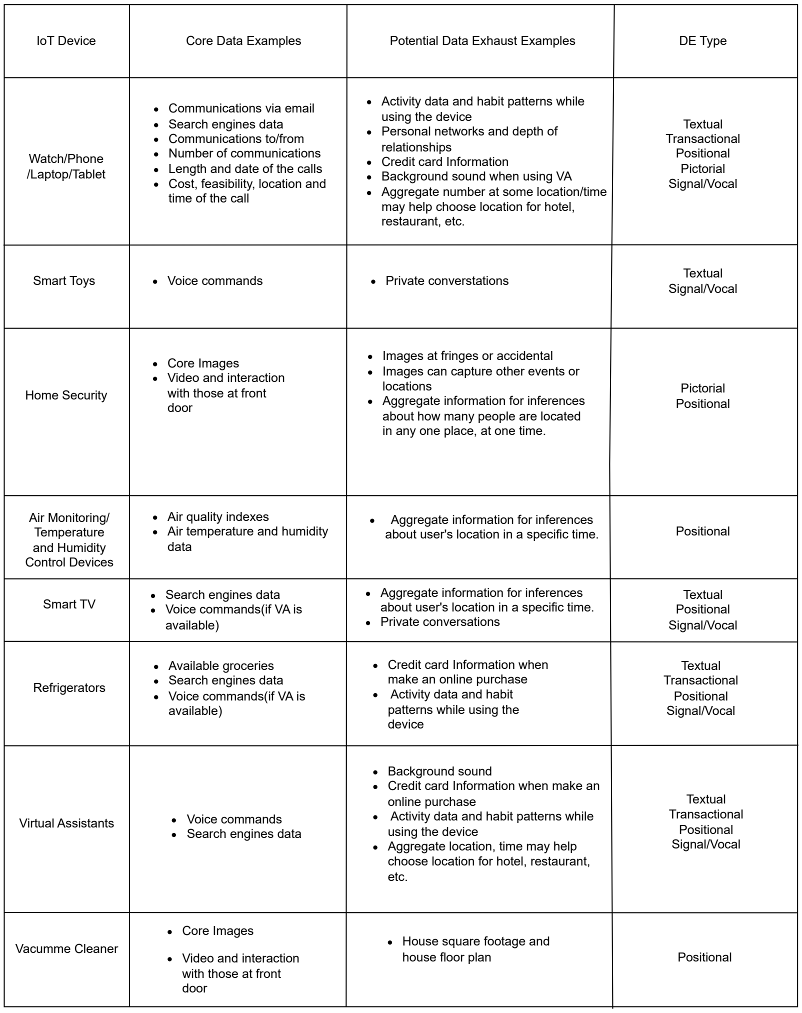
Based on a study conducted by Nshimba et al.[26] smart devices owners are subject to countless privacy risks. A number of privacy issues are addressed by this study based on the use cases of SAMSUNG, LG, and Sony smart televisions. By taking into account the architecture of IoT devices, the types of 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.[12] 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

For Peer Review

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 be collected by the sensor device. Table2 summarizes the core data and potential data exhaust for each IoT device discussed above.



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 make an online purchase • Activity data and habit patterns while using the device • Aggregate location, time may help choose location for hotel, restaurant, etc.	Textual Transactional Positional Signal/Vocal
Vacumme Cleaner	• Core Images • Video and interaction with those at front door	• House square footage and house floor plan	Positional

Table 2. Potential DE in different types of PIoT devices

4.3 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 to the users' personal information in many ways. Consequently, it is possible that users' profiles being developed based on not only their personal information but also long-term activities and behaviors. 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[15]. 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[39]. 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.

5 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[20]. 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[13].

Rutledge et al [34] 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 [35] 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 maybe unclear for the IoT devices users that **who** collect **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 SmartTVs for analysis through the use of goal-oriented techniques which they applied. According to 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 [29] examined different aspects of privacy in terms of voice assistants. According to this work, in a more sophisticate 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

For Peer Review

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[13]. 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[4].

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 Internet of Things (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[4]. 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[18].

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.

6 DISCUSSION AND POTENTIAL SOLUTIONS

As you may 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. Businesses and application depending on the type of the sensors that they may use in their smart device, may generate much data than that you expect. For instance, once a VA has been activated, the VA not only collects data about the person who activated and supposed to command it, but also collects information from background voice conversations like two other people conversation[29]. 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. 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

For Peer Review

necessary data and prevent it from collecting unnecessary raw data. This solution target 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 smart watches, 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 behavior without requiring participants to take part [18]. 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.

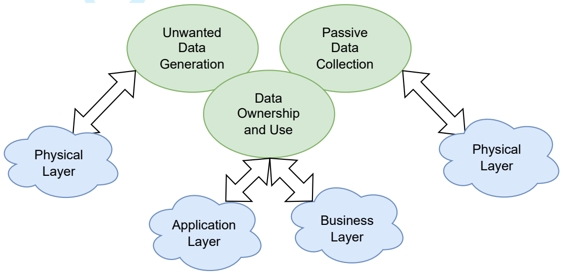


Fig. 3. Data Exhaust Issues In An IoT Ecosystem

In a study conducted by Iqbal et al. [12], 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 develop 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[23]. 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, classifieds 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.

7 FUTURE WORKS

Future research in the area of data exhaust in IoT devices could take two forms:

1- Protection of personal data: In view of the fact that IoT devices generate a large amount of sensitive and personal data, it is imperative that this data be protected from unauthorized access and manipulation. IoT device data exhaust could be handled and processed in a secure, privacy-preserving manner in the future.

For Review

2- Machine learning models for detecting IoT data exhaust: Machine learning models can be developed to detect whether generated data is data exhaust or core data. Future work could focus on developing machine learning techniques tailored to the unique characteristics of data exhaust.

REFERENCES

[1] Acar, A., Fereidooni, H., Abera, T., Sikder, A. K., Miettinen, M., Aksu, H., Conti, M., Sadeghi, A.-R., and Uluagac, S. 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* (2020), pp. 207–218.

[2] Aqeel-ur Rehman, S. U. R., Khan, I. U., Moiz, M., and Hasan, S. Security and privacy issues in iot. *International Journal of Communication Networks and Information Security (IJCNIS) 8*, 3 (2016), 147–157.

[3] Cheng, P., Bagci, I. E., Yan, J., and Roedig, U. Smart speaker privacy control-acoustic tagging for personal voice assistants. In *2019 IEEE Security and Privacy Workshops (SPW)* (2019), IEEE, pp. 144–149.

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

[5] Dakkak, A., Zhang, H., Mattos, D. I., Bosch, J., and Olsson, H. H. 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)* (2021), IEEE, pp. 243–252.

[6] Ding, J., Nemati, M., Ranaweera, C., and Choi, J. Iot connectivity technologies and applications: A survey. *arXiv preprint arXiv:2002.12646* (2020).

[7] eWEEK EDITORS. Human-centric iot: Top priority for business success. https://www.eweek.com/networking/human-centric-iot, May 2022.

[8] Gheisari, M., Najafabadi, H. E., Alzubi, J. A., Gao, J., Wang, G., Abbasi, A. A., and Castiglione, A. Obpp: An ontology-based framework for privacy-preserving in iot-based smart city. *Future Generation Computer Systems 123* (2021), 1–13.

[9] Guo, F., Yu, F. R., Zhang, H., Li, X., Ji, H., and Leung, V. C. Enabling massive iot toward 6g: A comprehensive survey. *IEEE Internet of Things Journal 8*, 15 (2021), 11891–11915.

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

[11] Hajiheydari, N., Talafidaryani, M., and Khabiri, S. 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* (2019), pp. 98–103.

[12] Iqbal, U., Bahrami, P. N., Trimananda, R., Cui, H., Gamero-Garrido, A., Dubois, D., Choffnes, D., Markopoulou, A., Roesner, F., and Shafiq, Z. Your echos are heard: Tracking, profiling, and ad targeting in the amazon smart speaker ecosystem. *arXiv preprint arXiv:2204.10920* (2022).

[13] Jiang, Y., Le, B. D., Zia, T., and Gauravaram, P. Privacy concerns raised by pervasive user data collection from cyberspace and their countermeasures. *arXiv preprint arXiv:2202.04313* (2022).

[14] Jones, M. L., and Meurer, K. Can (and should) hello barbie keep a secret? In *2016 IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS)* (2016), IEEE, pp. 1–6.

[15] Karale, A. The challenges of iot addressing security, ethics, privacy, and laws. *Internet of Things 15* (2021), 100420.

[16] Kassab, W., and Darabkh, K. A. A–z survey of internet of things: Architectures, protocols, applications, recent advances, future directions and recommendations. *Journal of Network and Computer Applications 163* (2020), 102663.

[17] Krishnamurthi, R., Kumar, A., Gopinathan, D., Nayyar, A., and Qureshi, B. An overview of iot sensor data processing, fusion, and analysis techniques. *Sensors 20*, 21 (2020), 6076.

[18] Maher, N. A., Senders, J. T., Hulsbergen, A. F., Lamba, N., Parker, M., Onnela, J.-P., Bredenoord, A. L., Smith, T. R., and Broekman, M. L. Passive data collection and use in healthcare: A systematic review of ethical issues. *International Journal of Medical Informatics 129* (2019), 242–247.

[19] Miorandi, D., Sicari, S., De Pellegrini, F., and Chlamtac, I. Internet of things: Vision, applications and research challenges. *Ad hoc networks 10*, 7 (2012), 1497–1516.

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

[21] Morey, T., Forbath, T., and Schoop, A. Customer data: Designing for transparency and trust. *Harvard Business Review 93*, 5 (2015), 96–105.

[22] Morgan, S. 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.

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

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

[25] Nižetić, S., Šolić, P., González-de, D. L.-d.-I., Patrono, L., et al. Internet of things (iot): Opportunities, issues and challenges towards a smart and sustainable future. *Journal of Cleaner Production 274* (2020), 122877.

[26] Nshimba, K., and Goede, R. An architecture approach to a secure home area network. In *2022 International Conference on Electrical, Computer and*

For Peer Review

Energy Technologies (ICECET) (2022), IEEE, pp. 1–6.

[27] OLeary, D., and Storey, V. C. Data exhaust: Life cycle, framework and a case study of stolen911. com.

[28] O'Leary, D. E., and Storey, V. C. Discovering and transforming exhaust data to realize managerial value. *Available at SSRN 3746010* (2020).

[29] Pal, D., Arpnikanondt, C., Razzaque, M. A., and Funilkul, S. To trust or not-trust: privacy issues with voice assistants. *IT Professional 22*, 5 (2020), 46–53.

[30] Patil, U. A., Venkatesan, M., and Prasad, S. An improved wireless network architecture for iot in hospital healthcare. In *2021 IEEE Bombay Section Signature Conference (IBSSC)* (2021), IEEE, pp. 1–6.

[31] Pierce, J. 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* (2019), pp. 1–14.

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

[33] Qadir, J., Ali, A., Zwitter, A., Sathiaseelan, A., Crowcroft, J., et al. Crisis analytics: big data-driven crisis response. *Journal of International Humanitarian Action 1*, 1 (2016), 1–21.

[34] Ren, J., Dubois, D. J., Choffnes, D., Mandalari, A. M., Kolcun, R., and Haddadi, H. Information exposure from consumer iot devices: A multidimensional, network-informed measurement approach. In *Proceedings of the Internet Measurement Conference* (2019), pp. 267–279.

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

[36] Sahoo, B. P., Mohanty, S. P., Puthal, D., and Pillai, P. Personal internet of things (piot): What is it exactly? *IEEE Consumer Electronics Magazine 10*, 6 (2021), 58–60.

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

[38] Sehrawat, D., and Gill, N. S. Smart sensors: Analysis of different types of iot sensors. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (2019), IEEE, pp. 523–528.

[39] Seliem, M., Elgazzar, K., and Khalil, K. Towards privacy preserving iot environments: a survey. *Wireless Communications and Mobile Computing 2018* (2018), 1–15.

[40] Sfar, A. R., Natalizio, E., Challal, Y., and Chtourou, Z. A roadmap for security challenges in the internet of things. *Digital Communications and Networks 4*, 2 (2018), 118–137.

[41] Shen, Y., Shen, S., Li, Q., Zhou, H., Wu, Z., and Qu, Y. Evolutionary privacy-preserving learning strategies for edge-based iot data sharing schemes. *Digital Communications and Networks* (2022).

[42] Shin, C., Chandok, P., Liu, R., Nielson, S. J., and Leschke, T. R. 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)* (2017), IEEE, pp. 705–710.

[43] Stojkoska, B. L. R., and Trivodaliev, K. V. A review of internet of things for smart home: Challenges and solutions. *Journal of cleaner production 140* (2017), 1454–1464.

[44] Trilles, S., González-Pérez, A., and Huerta, J. An iot platform based on microservices and serverless paradigms for smart farming purposes. *Sensors 20*, 8 (2020), 2418.

[45] Vailshery, L. S. 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/, Aug. 2022.

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

[47] White, C., and Gilmore, J. N. Imagining the thoughtful home: Google nest and logics of domestic recording. *Critical Studies in Media Communication* (2022), 1–14.

[48] Yabe, T., Jones, N. K., Rao, P. S. C., Gonzalez, M. C., and Ukkusuri, S. V. Mobile phone location data for disasters: A review from natural hazards and epidemics. *Computers, Environment and Urban Systems 94* (2022), 101777.

[49] Zainuddin, N., Daud, M., Ahmad, S., Maslizan, M., and Abdullah, S. A. L. A study on privacy issues in internet of things (iot). In *2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP)* (2021), IEEE, pp. 96–100.

For Peer Review