People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

University A/Mira of Bejaia
Faculty of Exact Sciences
Department of Computer Science

MASTER Thesis

In Computer Science

Option

Artificial Intelligence

Theme

Machine Learning Algorithms for Data Fusion in SIoT

Presented by:

- Nesrine BENSEGHIR.
- Tinhinane ALOUACHE.

Defended on 30th June 2024, before a jury composed of:

President	**Mr. M. Atmani**	**U. A/Mira Béjaia.**
Supervisor	**Mr. A. Achroufene**	**U. A/Mira Béjaia.**
Examiner	**Mrs. D. Zamouche**	**U. A/Mira Béjaia.**
Guest	**Mrs. B. Ait Moula**	**U. A/Mira Béjaia.**

Academic Year: 2023/2024

Dedications

I dedicate this thesis to myself.

To my father, the one who never stopped believing in me, even when everyone else had lost hope.Your faith in me has always been my guiding light. During those times when doubt clouded my own vision, your unwavering confidence kept me going. It's because of you that I never gave up on myself.

To my mother, whose love, patience, and unwavering encouragement have illuminated my path. Your selfless sacrifices and nurturing spirit have infused me with the strength to persevere through every challenge.

To my brothers, Sofiane, Bilel, and Oussama, and my sisters, Sabrina, Chahrazed and Kenza, for their steadfast support and love.

To my cousin Linda, for always being there for me, through thick and thin.
To my family and friends.
To my partner Tinhinane.
A special dedication to BDLX group.

Nesrine

Dedications

I dedicate this thesis with gratitude and deep love:

To my parents, your unconditional love, unwavering support, and endless sacrifices have made me the person I am today. Thank you for your precious guidance, your words of encouragement that always resonate in my heart, and for always being there when I needed comfort and support. I am forever indebted to you both, and I strive to make you proud in all that I do.

To my brother Ayoub and my sister Lyna. May this work be a source of inspiration for you. I wish you all the best in your studies. May you always find the strength and determination to pursue your dreams with the same passion and perseverance that you see in me.

To my paternal and maternal families.

To my partner, Nesrine, it was a pleasure working with you. I wish you much success.

Tinhinane

Acknowledgements

Praise and thanks to God, the Almighty, for His blessings throughout our research work.

We would like to express our gratitude to our supervisor, Mr. ACHROUFENE, who has always been attentive and available throughout the realization of this thesis. We thank him for framing, guiding, and advising us, and for the time he has devoted to us.

We express our appreciation to the jury members, Mr. Atmani and Mrs. Zamouche, and our guest of honor, Baya Ait Moula, for the honor they have bestowed upon us by evaluating this work.

Lastly, we are grateful to all those who have contributed in any way to the completion of this thesis.

Nesrine and Tinhinane

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI Artificial Intelligence.

ANN Artificial Neural Network.

API Application Programming Interface.

CERP-IoT Cluster of European Research Projects on IoT.

DPWS Device Profile for Web Services.

DST Dempster-Shafer Theory.

EM Embodied Microblogging.

GPS Global Positioning System.

ID Identification.

IEEE Institute of Electrical and Electronics Engineers.

IoT Internet of Things.

IoT-GSI Internet of Things Global Standards Initiative.

IP Internet Protocol.

LoRa Long Range.

ML Machine Learning.

RESTful Representational State Transfer.

RFID Radio Frequency Identification.

SIoT Social Internet of Things.

SN Social Networks.

UIT Union Internationale des Télécommunications.

WiFi Wireless Fidelity.

WoT Web of Things.

GENERAL INTRODUCTION

Advances in the technology of electronic devices and their miniaturization, the development of networks, particularly wireless technology, and the renewed interest in artificial intelligence in all its branches, have contributed considerably to the evolution of the Internet of Things (IoT). The integration of social relationships between intelligent objects of IoT has led to the emergence of the concept of the Social Internet of Things (SIoT). This innovation allows devices to build relationships and collaborate effectively and independently.

Problematic

SIoT presents a data management challenge due to the massive amount and diverse nature of data from connected devices with social interactions. Effective data fusion methods are crucial to integrate and exploit this complex data. Evaluating machine learning for classifying these complex relationships is difficult, as it requires techniques that handle data variety, dynamism, privacy, security, and high accuracy.

Objective

This study aims to improve the accuracy and reliability of classifying relationships in the Social Internet of Things through a multi-stage data fusion process and the application of advanced machine learning algorithms.

Methodology

This thesis is structured into three parts:

- **Chapter 1**: covers definitions and generalities on the basic notions of the social internet of things and the data fusion.
- **Chapter 2**: is a state of the art of some works in the literature that are related to data fusion in the context of SIoT.
- **Chapter 3**: presents the proposed data fusion approach to classify relationships among SIoT devices using machine learning. It also provides and discusses the obtained result of the performed tests.

Finally, the conclusion recalls the problem, summarizes the objectives stated at the beginning of this thesis, summarizes what has been done in the state of the art, highlights the tests and results obtained and considers interesting perspectives which can improve our work.

CHAPTER 1

SOCIAL INTERNET OF THINGS AND DATA FUSION

1.1 Introduction

Today, the Internet hosts billions of connections and exchanges, making it the most powerful tool for sharing information. In a few decades, it has become the driving force behind profound transformations in the lives of businesses, individuals, and institutions. Continuous communication is facilitated by a wireless network among various objects of daily life that interconnect and interact intelligently and cooperatively. This is the so-called Social Internet of Things (SIoT), a new paradigm combining social network concepts and the Internet of Things (IoT). This paradigm has given rise to an ecosystem in which devices can communicate with each other and users. An essential element of SIoT is the collection and fusion of data generated by connected devices. Data fusion is the process of the combination and integration of several data sources of different types or formats to produce a more complete and exact representation of the phenomenon studied. Using various strategies, such as statistical methodologies, machine learning algorithms, or expert knowledge, enables the extraction of useful and valuable information from the combined data.

In this chapter, we'll present the definition of the IoT, followed by the definition of the SIoT, its architecture, and fields of application. We'll also look at data fusion, its architecture, advantages, challenges, and techniques, and finally review the different types of machine learning.

1.2 Internet of Things

The term "Internet of Things" (IoT) does not yet have a consensus on its definition. Some emphasize its technical aspects, while others focus more on its uses and functionalities.

1.2.1 Definition of IoT

Among the definitions of IoT found in the literature, we cite :

- Internet of Things Global Standards Initiative (IoT-GSI) working group, led by Union Internationale des Télécommunications (UIT), considers IoT as [18]:

  «**A global infrastructure serving the information society to provide advanced services by interconnecting objects (physical and virtual) through the interoperability of existing or evolving information and communication technologies**».

- The Institute of Electrical and Electronics Engineers (IEEE) defines Internet of Things (IoT) as [19]:

  «**A network of elements, each equipped with sensors, that are connected to the Internet**».

- Cluster of European Research Projects on IoT (CERP-IoT) defines the Internet of Things as [20]:

  «**A dynamic infrastructure of a global network. This global network has self-configuring capabilities based on interoperable communication standards and protocols. In this network, physical and virtual objects have identities, physical attributes, virtual personalities, and smart interfaces, seamlessly integrated into the network**».

In summary, IoT can be considered as a collection of connected objects designed to make the real world smarter.

Some of the concepts used in the previous definitions, such as sensor, connected object and smart object, require clear definitions to distinguish them. This is what we present below.

1.2.2 Sensor

A sensor is an electronic device that measures physical, social, or environmental parameters, essential in the context of IoT because it allows the collection of crucial data for the operation of intelligent systems. The data collected by these sensors are then utilized to enhance social interactions, facilitate decision-making, and optimize the user experience within the context of an interconnected environment [21].

In IoT, different types of sensors are used depending on the specific needs and requirements of each use case. Here are some examples of sensors used in these contexts:

- Motion sensors (accelerometers and gyroscopes): integrated into wearable devices, connected vehicles, and surveillance systems to detect movement, orientation, and vibration.
- Position sensors (GPS, RFID): used for geolocation of connected objects, asset tracking, navigation, etc.
- Biometric sensors: utilized for security and authentication in Internet of Things systems, include fingerprint and face recognition sensors.
- Vibration sensors: monitoring the condition of infrastructure, predictive maintenance and equipment safety.
- Proximity and presence sensors: Used in home automation devices to detect object presence, as well as lighting control, security, etc.
- Ambient light sensors: To automatically adjust lighting based on surrounding light conditions, or to monitor lighting quality in wellness applications.
- Temperature and humidity sensors: Used to monitor and control environmental conditions inside buildings, in smart agriculture, or for cold chain management.
- Gas and air quality sensors: Used in industrial environments, smart buildings and smart cities to monitor air pollution and toxic gas levels.

These sensors are essential for collecting real-time data, transmitting it via wireless networks (like WiFi, Bluetooth, Zigbee, LoRa, etc.) to data management and analysis systems. They thus make it possible to make objects more intelligent and create connected environments capable of responding to user needs and preferences proactively.

1.2.3 Connected Object

Connected object is a physical device capable of communicating and exchanging information remotely through different types of connectivity with other devices. These objects are equipped with sensors to collect data from their environment, once the data is collected, the connected objects transmit this information to other devices, and sometimes process it to help make decisions or initiate actions. Connectivity is the main and differentiating element of a connected object, it allows the object to be connected to an infrastructure or to another connected object via a communication channel. The connected object integrates one or more IoT networks, as it shown in figure 1.1 [22].

Figure 1.1: Connected Objects [1].

Connected objects possess five key characteristics [23] [24]:

- **Identification:** each connected object is assigned a unique identifier (barcode, RFID chip, or IP address) for distinct recognition.
- **Sensitivity to the environment:** Capable of perceiving, analyzing, and collecting information from its surroundings.
- **Interactivity:** establishes a network connection either permanently or temporarily, depending on the object and its requirements.

- **Virtual representation:** an electronic signature representing the physical connected object.
- **Autonomy:** the ability of an object to act independently without external intervention, making each object responsible for itself.

In the composition of a connected object, the processing part, carried out by the microprocessor(s), can take an increasingly important part, providing the object with computing and processing power equivalent to a microcomputer. The object thus becomes more and more "smart" [25].

1.2.4 Smart Object

A smart object (or intelligent object) is a connected object that has advanced processing and analysis capabilities through an integrated microprocessor. It can process data autonomously, run complex algorithms, and adapt or learn from the environment using artificial intelligence. Smart objects can not only collect and transmit data, but also make decisions based on this collected data without human intervention. Smart objects connect to other objects or integrate a collection of devices, also called an IoT network [26] [25].

The intensive use of smart objects in recent years has contributed significantly to the rapid evolution of social networks. But what is a social network?

1.3 Social Networks

This section gives the definition of social networks and features that can be sued with IoT to constitute an advanced version of IoT.

1.3.1 Definition

Social Networks (SN) are online platforms that allow users to create public or semi-public profiles, interact with other users, and browse their connections. The nodes in SN refer to individuals and the edges between the nodes describe the relationships between the people. The SN are characterized by the following characteristics [3]:

- Community-driven: Discovering new friends and reconnecting with old ones.

- Interactive: Interacting with events and news.
- User-based: Real-time updates and control of profiles by users.

1.3.2 Component of the Social Networks that can be used with IoT

Social media networks profile users when creating their account, including their personal details. These profiles are accessible via a social graph displaying links between users. To maintain contact, SN use tools such as emails, instant messaging, blogs, discussion forums, telephony and videoconferencing. Service APIs enable the integration of third-party applications and external content. Therefore, the three fundamental features that can be adopted to give a social structure to the IoT are the following ones: [3]

- Identifier management: Assignment of universal IDs to identify all objects, ensuring interoperability of methods for detecting new objects.
- Object profile: Static and dynamic details of objects, classified according to key characteristics, allowing identification based on the services offered or interfaces offered.
- Owner control: Determining the functions and shareable data of objects, with security and access policies for each future operation.

These features facilitate the interaction and effective management of connected objects, integrating third-party services and improving communication between users.

1.4 From IoT to SIoT historical

The evolution from the Internet of Things to the Social Internet of Things (SIoT) represents a major step forward in connecting physical objects with the digital world. In the beginning, IoT systems were developed in isolation, leading to small, isolated groups of smart devices that couldn't easily work together because there was no standard architecture. This fragmentation limited the integration and functionality of these systems and prevented the creation of a unified IoT environment for complex applications.A simple but effective countermeasure to IoT fragmentation is to enable objects to communicate directly with external frameworks using the web protocols and networking paradigms universally accepted by the modern Internet of services [3].

The first innovation was the implementation of what is now widely known as the Web of Things (WoT) [27], which was a crucial development, using web protocols within devices or their gateways to enable direct communication with external systems through widely accepted web protocols. This integration led to the use of Device Profile for Web Services (DPWS) and Representational State Transfer (RESTful) APIs. This helped integrate IoT devices into the broader Internet of Services, but WoT still faced challenges in advertising, discovering, accessing, and using these devices and their services [28] [3].

Another advantageous aspect is the ability of Internet users and providers to sense and interact with the physical world. One approach in this context is to build a platform where objects can be easily discovered, checked, exploited, and composed. This is the case for several recent web solutions, such as SenseWeb (http://www.sensormap.org) and Xively (formerly Pachube— http://xively.com), that provide users with a central forum to exchange their sensor data and implement related applications. The natural evolution of this concept is to encourage sharing smart devices among trusted users without needing to create a new social network or user database on a new web service [3].

Indeed, Holmquist et al. [29] introduced one of the first ideas of pseudo-socialization between objects. Using the so-called Smart-Its Friends technique, users had a very convenient interface to establish temporary friendship relationships on Smart-Its (smart wireless systems that typically combine sensing, computing, and communication functions) depending on the system context.

The so-called Blog-jects, corresponding to the "objects that blog" presented in [30], describe this new approach towards strong interaction with the world, which is considered necessary to be embodied in traditional devices. The leap away from the past is illustrated here by a clear distinction between a "thing" that is merely linked to the Internet and a "thing" that plays an active role in the social network.

The concept of Embodied Microblogging (EM) proposed in [31] goes beyond simply connecting objects via the IoT. It suggests enhancing everyday objects to facilitate human-to-human communication and make daily events more noticeable.

The reference [32] describes the expected IoT network architecture but does not detail potential social features. While this article explores the combination of IoT and social networks and provides useful application examples, it does not propose protocols or architectural solutions for a social IoT.

A significant contribution to describing a social IoT is presented in [32]. This article explores the possibilities of combining IoT and social networks and provides useful application examples. However, it does not discuss potential protocols for establishing social connections between objects or possible architectural solutions for a social IoT. Similarly, the concept of a social IoT is present in numerous strategic study agendas, often as a simple declaration of interest, as illustrated by the Finnish Strategic Agenda for Science [3].

Significant attention is given to exploring the social potential of IoT components in [33]. This research describes an architecture where objects are explicitly considered capable of forming interest groups and taking collective actions. However, the article does not define how to create the desired social network of objects or incorporate the necessary architectures and protocols.

Various studies, evaluate social characteristics by examining social relationships among nodes and describe initial investigation results regarding system characteristics in terms of specific key parameters [34]. The behavior of mobile nodes is also analyzed by applying typical principles of social networks, as detailed in [2]. The following figure 1.2 shows in summary the historical movement from IoT to SIoT.

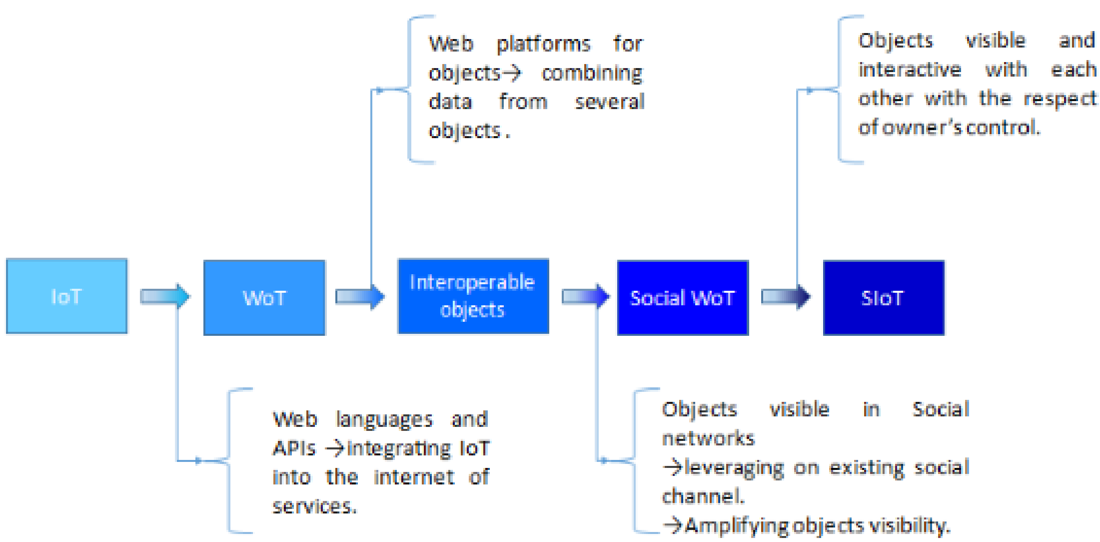


Figure 1.2: The historical development from IoT to SIoT [2] [3].

Consequently, a new generation of social objects with the following capabilities is envisioned [3]:

- can communicate with other objects independently of their owners.
- can efficiently navigate the IoT composed of billions of objects to access services and

information in a trust-oriented manner.

- are capable of advertising their presence to offer services to the rest of the network.

This represents a new vision of an enhanced IoT where principles and technologies characteristic of social networks are extended to the realm of things, enabling resource visibility, service discovery, object reputation estimation, source aggregation, and service composition, akin to advancements made in solving routing challenges in delay-tolerant networks. [3]

1.4.1 Definition of SIoT

The Social Internet of Things (SIoT) is a paradigm that integrates the Internet of Things and social networks as presented in figure 1.3, where every object can establish social relationships with other objects independently with based on the heuristics set by the owner object. The primary goal of SIoT is to handle the vast number of interconnected devices, especially when confronted with challenges related to information and service discovery. Unlike traditional IoT, which focuses on sensing and networking, SIoT emphasizes service discovery and composition to facilitate autonomous interactions among objects, thereby improving the overall user experience [35].

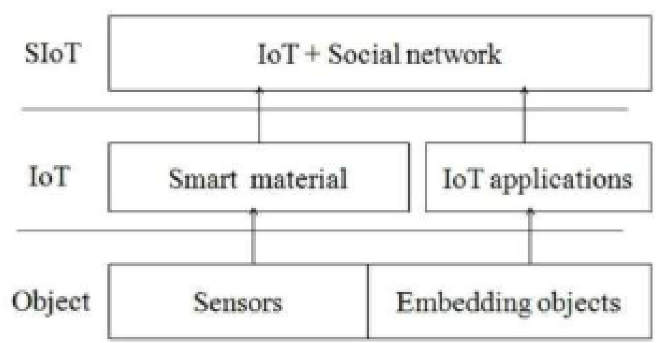


Figure 1.3: Combination of social network and IoT [3]

.

1.4.2 SIoT Architecture

To design SIoT systems, several architecture were presented. The breakdown of the SIoT architecture shown in figure 1.4 into different layers helps in understanding its flow and functionalities.

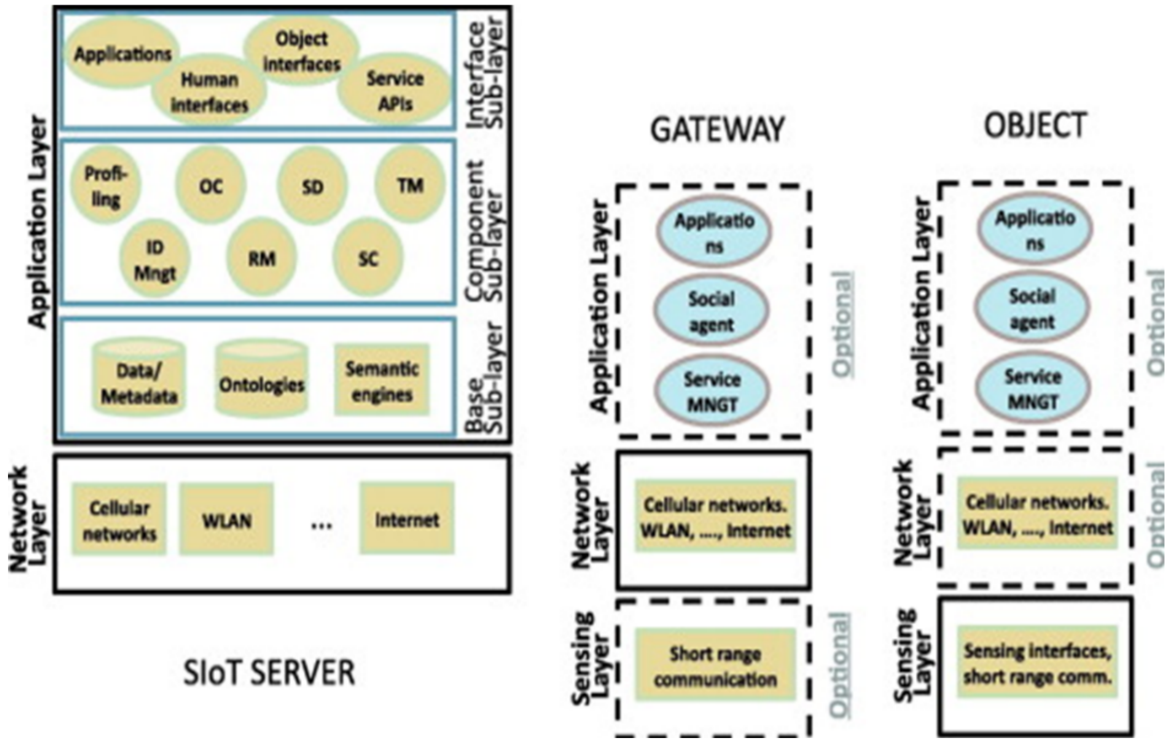


Figure 1.4: Architecture of SIoT.

Let's delve deeper into each layer [36] :

1. **Sensing Layer:**
   - In this layer, sensors and devices are responsible for collecting data from the physical environment. These devices are typically equipped with various sensors (e.g., temperature, motion, light) to gather relevant information.
   - The collected data might include environmental conditions, user activities, or any other relevant parameters.
2. **Network Layer:**
   - The network layer establishes the communication infrastructure that connects IoT devices to each other and to the internet.
   - Wired and wireless communication protocols such as WiFi, Bluetooth, Zigbee, etc., facilitate the exchange of data between devices and enable connectivity.
3. **Application Layer:**
   - The application layer encompasses the specific applications developed to address various domains, such as healthcare, transportation, or smart homes.

- Applications leverage the capabilities of the lower layers to provide services and functionality tailored to the needs of users and specific use cases.
- The layer is where users directly interact with SIoT applications to monitor, control, and receive insights from their IoT devices.

Furthermore, the SIoT system is composed of three basic elements which are :

1. **SIoT server:** includes the network and application layers. The latter layer is composed of three sub-layers: the base sub-layer that contains the database for storing and managing data and relevant descriptors, the social relation management sub-layer that manages the profiles of social members and their relationships, as well as the activities of objects in the real and virtual worlds, and the service management sub-layer that manages the provision of services.
2. **Gateway:** is optional, it is used to connect objects to the network layer.
3. **Objects:** is a physical element that is connected to the SIoT and can be sensors, actuators, or processing devices.

The success of SIoT depends on effective integration and collaboration between these layers and addressing challenges such as security, privacy, and interoperability to ensure a seamless and trustworthy user experience.

1.4.3 SIoT Application Domains

Social IoT has applications across various industries according to [4] as shown in the figure 1.5, including the following:

- **Traffic Management:** is the process of cars exchanging data about traffic conditions to assist drivers in selecting the most efficient routes that would get them to their destination faster. An example of this application can be found in [37].
- **Healthcare:** popular applications such as [14] where devices make it easier to find specialist doctors by leveraging co-location or social connections and tracking patient health information.
- **Education:** you can just message other devices in a social network to find solutions to mathematical equations [38].

Figure 1.5: SIoT application domains [4].

- **Industry:** the cooperation of industrial devices can facilitate the resolution of technical problems and improve manufacturing processes, as described in [39].
- **Logistics and supply chain management:** in [40] devices can track the movement of goods and optimize logistics operations.
- **Retail management:** connected devices can help automate billing processes and improve the customer experience, as well as help manage inventory [41].
- **Farming and agriculture:** for instance [42], with the use of devices, farmers may track crop growth, keep an eye on the weather, and improve their methods. .

1.4.4 Social Relationships between Objects

According to [5] [4], there are five types of relationships that can be established between the objects, as illustrated in the figure 1.6 :

- **Co-location Object Relationship**: this relationship is established among objects which are located in the same place.

- **Co-work Object Relationship**: this relationship exists whenever objects collaborate together to offer a common IoT application.
- **Parental Object Relationship**: it is established between objects belonging to the same production batch (e.g., the same model or the same manufacturer).
- **Social Object Relationship**: it occurs between objects when they come in contact with each other through social relationships, such as a relationship between a sensor and objects belonging to friends in a social network.
- **Co-owner Object Relationship**: it is established between heterogeneous objects that belong to the same user.

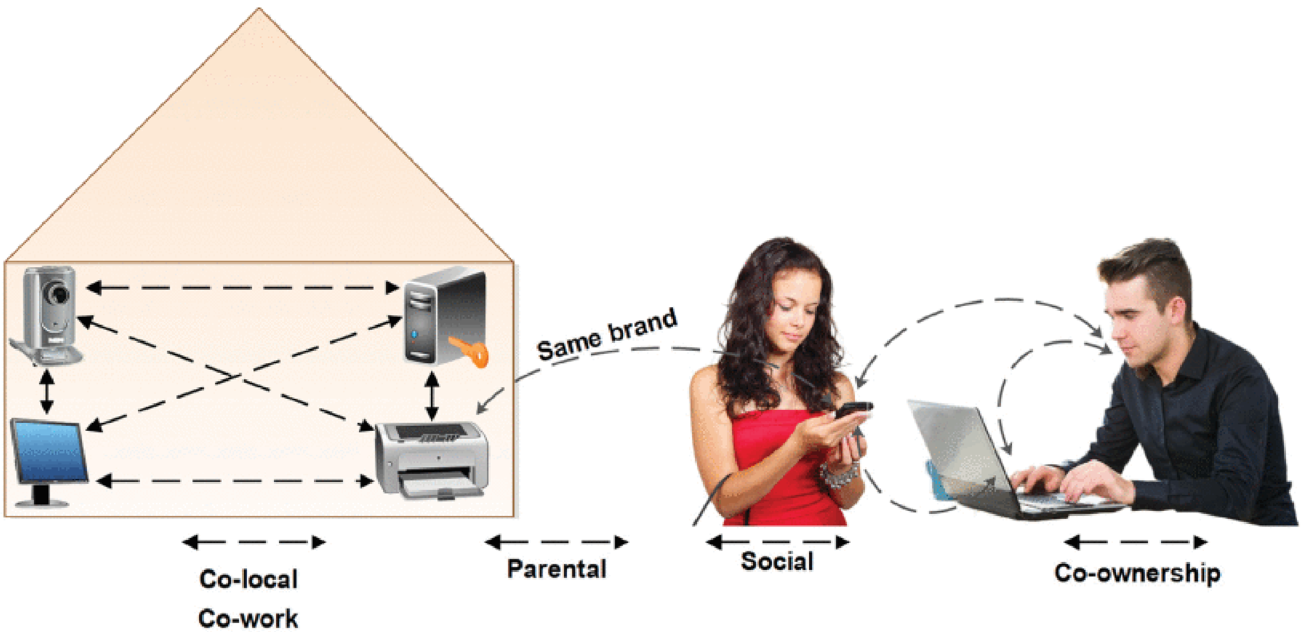


Figure 1.6: Types of social relationships among objects in the context of SIoT [5].

Sociological studies show that the type of relationship influences the level of trust and interactions between individuals. Similarly, interactions between objects are influenced by the type of relationship, affecting trust and reliability of services. In a hierarchy of relationships between objects, the parental relationship (POR) is the strongest in terms of reliability and trust, because it is characterized by a high level of support and trust, similar to a family relationship. Next comes the ownership relationship (OOR), where objects belong to the same owner, implying a special connection but with a lower level of trust than POR. The cooperative relationship (CWOR) follows, where objects collaborate for common work, requiring a certain mutual trust. The social relationship (SOR) is based on opportunistic or planned meetings of the owners of the objects, with a lower level of trust due to the uncertainty of the parameters. Finally, the communication layer relationship (CLOR) is the most implicit and has the lowest level of trust, because it is simply based on the objects being in the same network or communication environment. Thus, the proposed hierarchy is: CLOR < SOR < CWOR < OOR < POR. [43]

Like the Internet of Things, the Social Internet of Things generates a huge amount of data, which is difficult to manage in terms of collection, processing and transmission. One of the solutions proposed for these problems in the IoT is the use of data fusion, which can also be applied to the SIoT.

1.5 Data Fusion

In the context of the Internet of Things, the constant flow of communication between smart objects produces vast volumes of data often containing imperfections such as imprecision, uncertainty, and conflict. The main challenge lies in the effective management of this data, including its analysis, manipulation and transfer. To solve this problem, the data fusion process is essential. Fusion makes it possible to combine massive, multi-source, heterogeneous and sparse data sets to produce more reliable information for better decision-making. In what follows, we will define data fusion, see its architecture, its benefits, its techniques and its challenges.

1.5.1 Definition of Data Fusion

Depending on the application [6] [44] [45], data fusion can be defined as the combination of diverse data, knowledge, and information from different heterogeneous sources. This process aims to complement, validate, and enrich finding, resulting in more reliable, precise, accurate, and insightful information than what can be achievable from any single source.

1.5.2 Data Fusion Architecture Overview

The fusion process can be summarized in four major steps as shown in figure 1.7:

1. **Modeling:** it consists of choosing the knowledge representation formalism that can be guided by additional information. This step is crucial, as it determines a function (distribution, cost, etc.) for each piece of information from any source.

2. **Estimation:** it depends on the modeling. This step is not systematic but it is often necessary. For example, it involves estimating probability distributions.

3. **Combination:** it is the actual fusion step, which combines the data by selecting an operator from among the various proposed according to their basic properties (associativity, commutativity, idempotency, and adaptability).

4. **Decision:** this is the final step, in which a decision criterion is used to determine the result of the fusion. The choice of criterion depends on the modeling and combination steps.

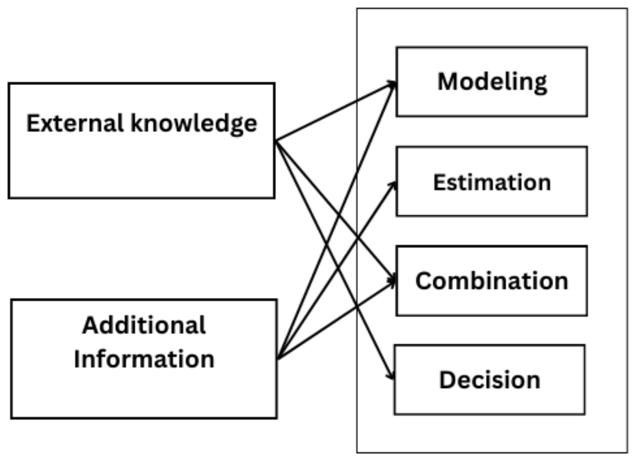


Figure 1.7: Representation of the Fusion block [6]

1.5.3 Data Fusion Challenges

Data fusion still faces several challenges to overcome to maximize its benefits, despite the various fusion models proposed to meet specific needs in many real-world applications. Most of these challenges are due to the complexity of the environments in which sensors operate, as well as the diversity of data to be combined, among other factors. According to [46] [47], here are some key challenges:

- **Data Imperfection**: this problem is common and a major challenge that all data fusion methods must address. The data collected by sensors is often imprecise, uncertain, ambiguous, vague and incomplete. To improve data quality, it is usually necessary to model these imperfections and use other available information as well as powerful mathematical tools. If data fusion fails to extract accurate and useful information, the imperfection of the data will seriously compromise the quality of the fusion.

- **Data inconsistency**: uncertainties can arise due to noise inherent in measurements, sensors and environments. These noises lead to data disorder or confusion, collectively called data inconsistency. Data inconsistency can have extremely detrimental effects on data fusion if the fusion model fails to identify the sources of this noise. Data fusion techniques should solve this problem by eliminating the influence of data inconsistency. Additionally, some erroneous data resulting from sustained or dynamic failures are difficult to model and predict in traditional ways.
- **Data conflict**: this problem often arises in systems using belief functions or Dempster-Shafer Theory (DST) . When issues that should be treated independently are mistakenly integrated, representational errors occur.
- **Data Alignment/Registration and Correlation**: data captured by different sensors with distinct frames must be aligned into a common frame before it can be merged, this process is called data alignment or registration. Mistakes in this process can result in over/under confidence. Another challenge is data correlation, which mainly occurs in a distributed environment when the same data set is calculated or merged multiple times, often because of cyclical loops in the topology, a phenomenon called data incest. Correlated data can seriously bias an estimate in a fusion system if it is not properly removed by data fusion algorithms.
- **Data type heterogeneity**: data captured by sensors in various environments can be of very different types. Just as people's eyes, noses, and mouths have distinct functions, sensors also serve varied purposes. Data fusion methods must be able to integrate different types of data to describe the complete state of an object.
- **Fusion location**: this is a major challenge in wireless sensor networks and other distributed fusion environments. Data can be merged into a central or local node. The first method consumes more bandwidth and time. The second method reduces the communication overhead, but may compromise data accuracy due to information loss from local fusion. Finding a balance between cost and quality of fusion is a complex problem.
- **Dynamic fusion**: the complexity of data fusion depends not only on the type of data and the collection environment, but also on its timeliness. To estimate the state of a system, especially a system evolving over time, data may only be meaningful for a limited period. This challenge must be well addressed in a

real-time application environment. The merge node must be able to distinguish the correct order of the data and its validity.

1.5.4 Benefits of Data Fusion

Data fusion offers various advantages, including [47]:

- **Enhanced Information:** Data Fusion combines data from multiple sources to make information more intelligent, decisive, sensible and precise than any single source can provide.
- **Statistical Advantage:** by calculating the several independent observations, one can predict that the data are fused in an optimal way.
- **Energy Efficiency:** the fusion of data obtained from low-power sensors with low accuracy makes it possible to create highly accurate information. Which overcomes limitations of high-power, high-accuracy sensors in IoT applications.
- **Big Data Handling:** Data Fusion helps manage the deluge of data in IoT by transforming it into more concise and accurate information.
- **Improved Security:** Data fusion can help conceal sensitive information or data semantics, enhancing security and privacy.

1.5.5 Data Fusion Levels and Strategies

The data fusion approaches have three types [7] as shown in the figure 1.8 :

- **Early Fusion** : a raw data from various methods is combined at the input level before feeding it to the model.
- **Late Fusion** : the data from each method is processed independently through separate models, and the results of these models are then combined at a later stage.
- **Hybrid Fusion** : it combines different fusion strategies to achieve the desired results.

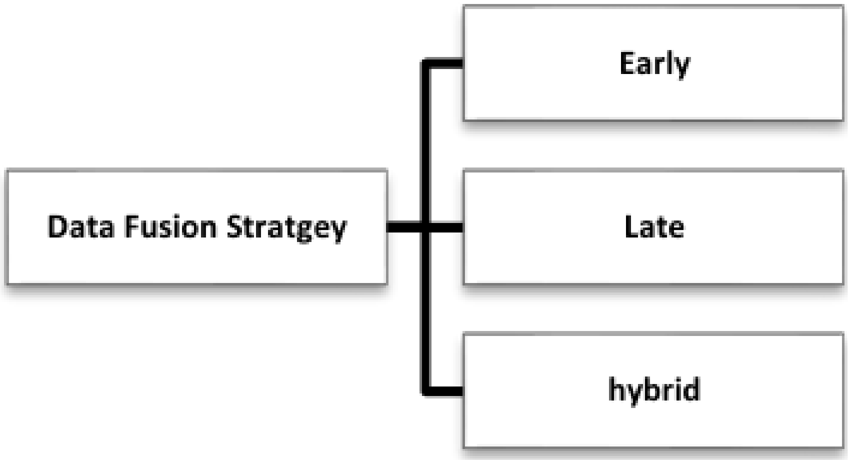


Figure 1.8: Data Fusion Strategies Levels [7].

1.5.6 Data Fusion Techniques

Data fusion techniques can be classified into three categories [47]:

- **Probability-based methods** including Bayesian analysis, statistics, and recursive operators.
- **Artificial Intelligence (AI) based techniques** including classical machine learning, fuzzy Logic, Artificial neural networks (ANNs) and genetic evolution.
- **Theory of Evidence** including Dempster–Shafer theory (DST).

Here is a comparative Table 1.1 of these techniques according to [47].

Approach	Techniques	Strengths	Weaknesses
Probabilistic	Bayesian inference, Hidden Markov Models.	Simple, Less complex, Widely accepted.	Can result in low accuracy. complexity increases with non-monotonic logic.
Artificial Intelligence Approaches	Supervised Machine Learning Neural networks fuzzy logic.	Highly accurate Handle non-linear relationships and uncertainties.	Complex Computationally expensive.
Theory of Evidence	DST.	Efficiently handle conflicting and missing data.	Complexity in computation.

Table 1.1: Comparison of data fusion approaches.

In the following section we introduce machine learning widely in the framework of data fusion.

1.6 Machine Learning

Referencing [48] [8] [49], broadly speaking, Machine Learning (ML) is a field of artificial intelligence (AI) that enables computers to learn from data and make decisions or predictions without having been explicitly programmed to do so. The aim is to create and implement algorithms that facilitate these decisions and predictions. These algorithms are designed to improve their performance over time.

In traditional programming, a computer follows a predefined set of instructions to complete a task. In machine learning, on the other hand, the computer is given a set of examples (data) and a task to perform, but it is up to the computer to determine how to perform the task based on the examples given to it.

Machine Learning techniques can be classified into three types, as shown in figure 1.9, based on the nature of the learning system and the data available: supervised learning, unsupervised learning, and reinforcement learning.

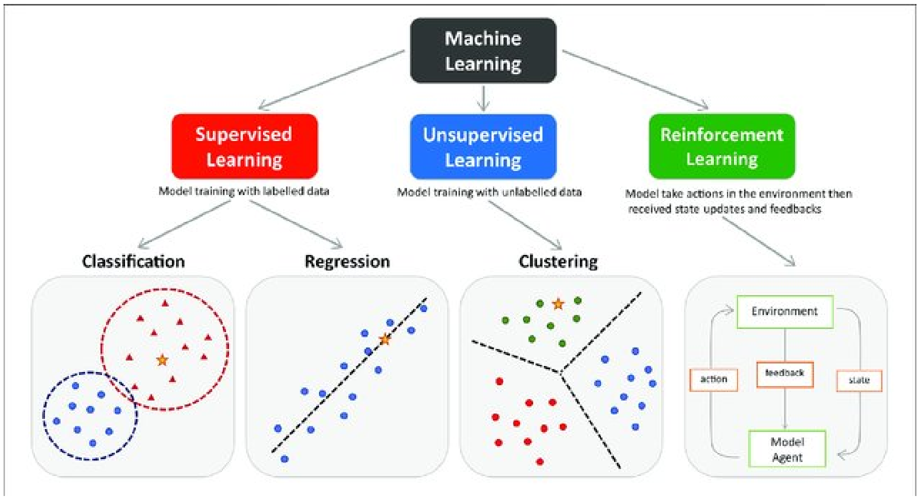


Figure 1.9: The main types of Machine Learning [8].

Supervised Learning

In this approach, the data is accompanied by a label that the model is trying to predict. This could be anything from a category label to a real-valued number. During the training phase, the model learns the mapping between inputs (features) and outputs (labels). Once trained, the model can predict the output for new, unkown data. It is used for classification and regression [48] [49].

Unsupervised Learning

The model is trained on an unlabeled dataset. The model is left to find patterns and relationships in the data itself. It is often used for clustering and dimensionality reduction [48] [49].

Reinforcement Learning

Reinforcement learning is a type of machine learning where an agent learns to make decisions by interacting with its environment. The agent is rewarded or penalized (with points) for the actions it takes, and its goal is to maximize the total reward. It is particularly suited to problems where the data is sequential, and the decision made at each step can affect future outcomes [8] [49].

1.7 Conclusion

Through this chapter, we've introduced the generalities and a few important aspects of IoT, SIoT, Data Fusion and Machine Learning. In the next chapter we will present the state of art of Machine Learning Algorithms for Data Fusion in SIoT.

CHAPTER 2

A STATE OF THE ART

2.1 Introduction

The rapid development and integration of the Social Internet of Things (SIoT) has led to a massive increase in data generated by smart devices. It is essential to implement data fusion techniques to improve the communication, analysis and decision-making capabilities of intelligent systems.

In this chapter, we begin by reviewing the main works on data fusion techniques in SIoT, evaluating the effectiveness of machine learning algorithms for data aggregation and relationship classification. Following this, a comparative analysis of these approaches will be presented to identify the strengths and limitations of existing methods.

2.2 Related work

In this section, we will present 12 articles which were chosen according to the following criteria:

- **Thematic Relevance**: Each article addresses crucial aspects of data fusion or relationship classification in the context of the Social Internet of Things (SIoT). Each has been chosen specifically for its direct link to the central themes of our research theme.
- **Diversity of Approaches**: A variety of methodological and technical approaches

have been included to provide a holistic perspective. This diversity makes it possible to compare the methods with each other and to evaluate their respective advantages as well as their limits.

- **Currentness**: Particular attention has been paid to recent articles in order to reflect the most current advances in the field. However, older articles were also considered if they have had a significant impact or are fundamental to understanding the evolution of knowledge in this area.

2.2.1 Performance comparison of machine learning algorithms for data aggregation in social internet of things

- **Keywords**: Data Aggregation, Internet of Things, Machine Learning, Performance evaluation, Social Internet of Things.
- **Authors**: "Meghana J", "Hanumanthappa J", "Shiva Prakash SP"
- **Abstract**:

  In this paper [9], the authors have proposed a method based on data aggregation in Social internet of things (SIoT) according to the object's profile, identifying conditions to establish relationships between devices, and evaluating the performance of machine learning algorithms for relationship classification.

  This approach is divided into four major parts as shown in the figure 2.1 and described below:

  - **Data Aggregation**: in this step the data generated by SIoT devices is aggregated based on the object profile.
  - **Relationship Conditions**: the method identifies the conditions to establish relationships between devices based on their features, such as device type, device brand, protocols, etc.
  - **Training the Machine Learning Algorithms**: the Machine Learning algorithms, such as Decision Tree, K-Nearest Neighbors, Naive Bayes, and Artificial Neural Network, are trained using the aggregated data and the relationship conditions established in the preceding steps.

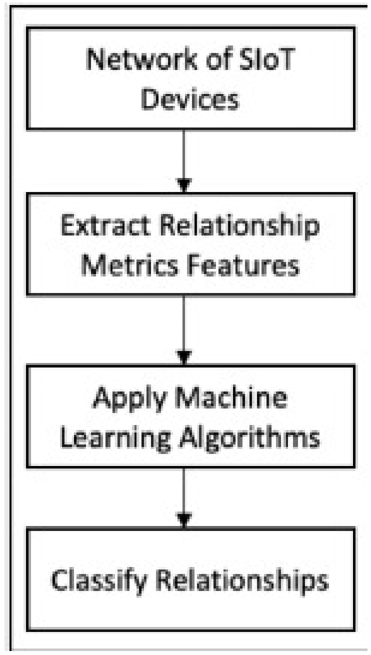


Figure 2.1: Performance comparison methodology in SIoT [9].

- **Classification of Relationships**: once the algorithms are trained, they are applied to classify the relationships between devices in the SIoT. The algorithm's performance was evaluated using metrics like accuracy, recall, and F1 score.

The results indicated that the decision tree outperformed others in terms of precision, recall, and F1 score. The study highlights several limitations including potential bias in data aggregation and the need for additional testing across a wider range of devices and environments to effectively generalize the results. In addition, there are constraints related to the scalability issues with increasing number of devices and data volume, and limited consideration of dynamic changes in device states and relationships.

2.2.2 DFIOT : Data Fusion for Internet of Things

- **Keywords**: Data Aggregation, Internet of Things, Machine Learning, Performance evaluation, Social Internet of Things.
- **Authors**: "Sahar Boulkaboul" and "Djamel Djenouri".
- **Abstract**:

  In this paper [10], the authors have proposed a new data fusion method for the Internet of Things, called DFIOT, which considers the reliability and conflicts among each device in the network and combines rules based on Basic Probability Assignment (BPA) to represent uncertain information or quantify the similarity between two belief sets. It is based on Dempster-Shafer theory and an adaptive weighted algorithm that assigns a weight to each data source describing its level of confidence, considering the information's lifespan, the distance between sensors and entities, and computation reduction to ensure maximum reliability, accuracy, and conflict management. The steps are illustrated in the figure 2.2

Figure 2.2: The steps of DFIOT method [10].

DFIOT achieved up to 99.18% accuracy on artificial datasets and 98.87% on real datasets, with a conflict rate of 0.58% and energy savings up to 90%. However, the method may still be computationally intensive for larger networks and, while effective in smart building testbeds, its applicability to other IoT domains requires further validation.

2.2.3 Effective Features to Classify Big Data Using Social Internet of Things

- **Keywords**: : Internet of Things, social Internet of Things, machine Learning, big data, feature selection.
- **Authors**: "Lakshmanaprabu S. K", "Shankar.K", "Khanna Ashish", "Gupta Deepak", "Rodrigues Joel. JPC", "Pinheiro Plácido. R" and "De. Albuquerque, Victor Hugo. C".
- **Abstract**:

  LAKSHMANAPRABU et al [11] propose in this article an approach which aims to extract relevant features and classify Big Data using Social Internet of Things (SIoT). The method is divided into five steps as shown in the figure 2.3.

  - **SIoT data collection**: data is collected from various sensors and connected objects within the Social Internet of Things.

– **Data filtering**: Gabor filter is used to remove noise and undesirable data in the raw data. It is the key of the proposed approach.

– **Data base reduction**: Hadoop's MapReduce framework is used in the filtered data to reduce the database by fixing a threshold value.

– **Feature selection**: the most important features are selected using Elephant Herd Optimization (EHO) to improve classification.

– **Classification**: the optimal feature attributes are given as an input to a linear kernel support vector machine (SVM), which classifies this data.

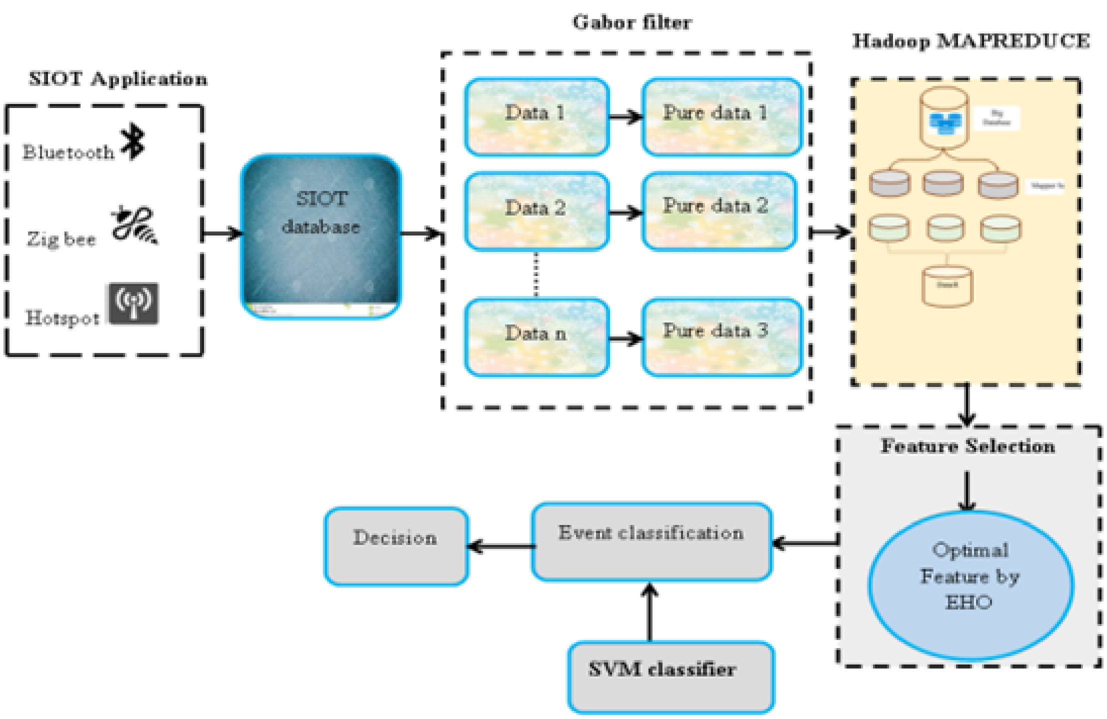


Figure 2.3: Elephant Herd Optimization method [11].

The proposed method achieved a maximum accuracy of 98.86%, demonstrating its effectiveness compared to existing methods in terms of accuracy, processing time, and memory usage. However, the method may have limitations in handling the heterogeneous nature of SIoT, high memory consumption, and processing power requirements, which can affect scalability and real-time performance.

2.2.4 Service Oriented R-ANN Knowledge Model for Social Internet of Things

- **Keywords**: : SIoT, objects, ANN (Artificial Neural Network), AI, predictive modeling.
- **Authors**: "Mohana.S.D", "S.P.Shiva Prakash" and "Kirill Krinkin".
- **Abstract**:

  The aim of this paper [12] is to develop a service-oriented knowledge model for the Social Internet of Things (SIoT) using an Artificial Relational Neural Network (R-ANN) to establish relationships between objects and services. The proposed method is divided into 5 steps:

  - **Data collection**: Data is collected from public and private SIoT objects, which is distributed randomly in the SIoT environment.
  - **Pre-processing of sensor data**: the Gaussian technique is used to normalize data, as shown in the figure 2.4.
  - **Feature selection**: the relevant features are selected for services using semantic rules.
  - **Use of R-ANN**: to establish relationships between objects and services based on semantic rules and defined conditions.
  - **Classification of services**: according to their relationship with objects and users.

  The performance of the proposed R-ANN model was evaluated using metrics such as accuracy, precision, and recall, which demonstrated the model's effectiveness in classifying services and establishing meaningful relationships between objects and services in the SIoT environment. The model showed high accuracy in predicting and classifying services, with the use of semantic rules and feature selection methods contributing to its robustness. However, the model faces challenges with handling heterogeneous data from various SIoT objects, scalability in larger and more complex environments, and real-time data processing.

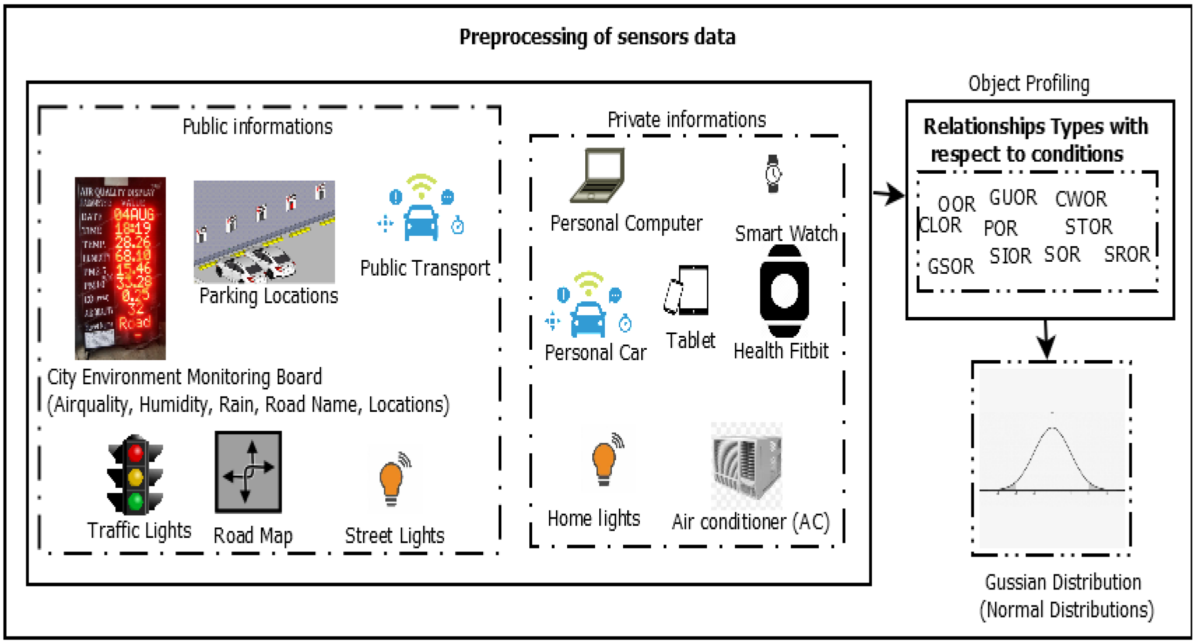


Figure 2.4: Sensor data preprocessing from random to Gaussian distribution [12].

2.2.5 An Artificial Intelligence Approach for Enhancing Trust Between Social IoT Devices in a Network

- **Keywords**: : SIoT, Trust, Social networks, Transitivity, Artificial intelligence.
- **Authors**: “J. Senthil Kumar”, “G. Sivasankar ”, “S. Selva Nidhyananthan”.
- **Abstract**:

  In this paper [50], the authors have proposed a method to enhance security and trust in Social Internet of Things (SIoT) networks using an approach integrating blockchain, cryptography and artificial intelligence (AI). To achieve this objective, the authors propose a several-step method:

  – **Integration of DeepChain**: DeepChain is a platform that combines blockchain, cryptography and AI to secure communication and decision-making in SIoT networks.

  – **Secure Communication**: DeepChain is used to facilitate secure communication between SIoT devices, ensuring authenticity and confidentiality of data exchanges.

  – **Threat Detection**: analyze the behavior of SIoT devices with AI algorithms integrated into DeepChain to detect threats or abnormal activities.

  – **Data Protection**: the use of cryptographic techniques specifically the Paillier algorithm combined with DeepChain, to protect the confidentiality of data exchanged between SIoT devices, thereby ensuring their security.

This approach was evaluated through different social networks (Facebook, Quora, Twitter) using the CC3200 SimpleLink Wi-Fi module from Texas Instruments and MATLAB simulations of extra nodes. Trust transitivity was calculated using aggressive, conservative, and traditional approaches, with the best trust levels going to the aggressive approach. The results showed that Quora achieved a better net profit. However, the approach encounters several challenges and limitations, including complex search strategies, resource constraints, the necessity for dynamic configurations, and susceptibility to various malicious trust attacks.

In conclusion, the study showed how AI can improve SIoT trust despite limitations and challenges, as well as the need for effective algorithms and approaches to solve them.

2.2.6 A Social IoT-based platform for the deployment of a smart parking solution

- **Keywords**: : Internet of Things, Vehicle detection, Smart Parking.
- **Authors**: “Alessandro Florisa,b”, “Simone Porcua,b”, “Luigi Atzoria,b”, “Roberto Girauc”.
- **Abstract**:

The authors propose in this paper [13], a new smart parking solution based on Social IoT that aims to provide information on the status of parking spaces provided in street parking lots. The platform uses the Social Internet of Things (SIoT) paradigm to create virtual entities of real objects involved in smart parking systems, solving issues such as scalability, interoperability, low energy consumption and timely prediction of parking space availability. The system also uses magnetometer sensors, concentrators, control dashboards, Android apps and smart payment services to improve the efficiency and user experience of smart parking solutions.

The Smart Parking (SP) system architecture is based on the Lysis architecture, which is known for its scalability and flexibility. It consists of four layers as shown in figure 2.5:

– **Hardware layer**: including physical sensing and data transmission devices such as vehicle identification cards and Bluetooth beacons.
– **Virtualization layer**: create Social Virtual Object (SVO) parking units for each parking space, implementing the Social Internet of Things (SIoT) paradigm.
– **Aggregation layer**: Use microengines (ME) to improve functionality and perform data analysis on aggregated data from SVO.
– **Application layer**: includes the administrator management platform and Android applications, allowing citizens to receive notifications, track parking space occupancy, and access other services such as electronic payments.

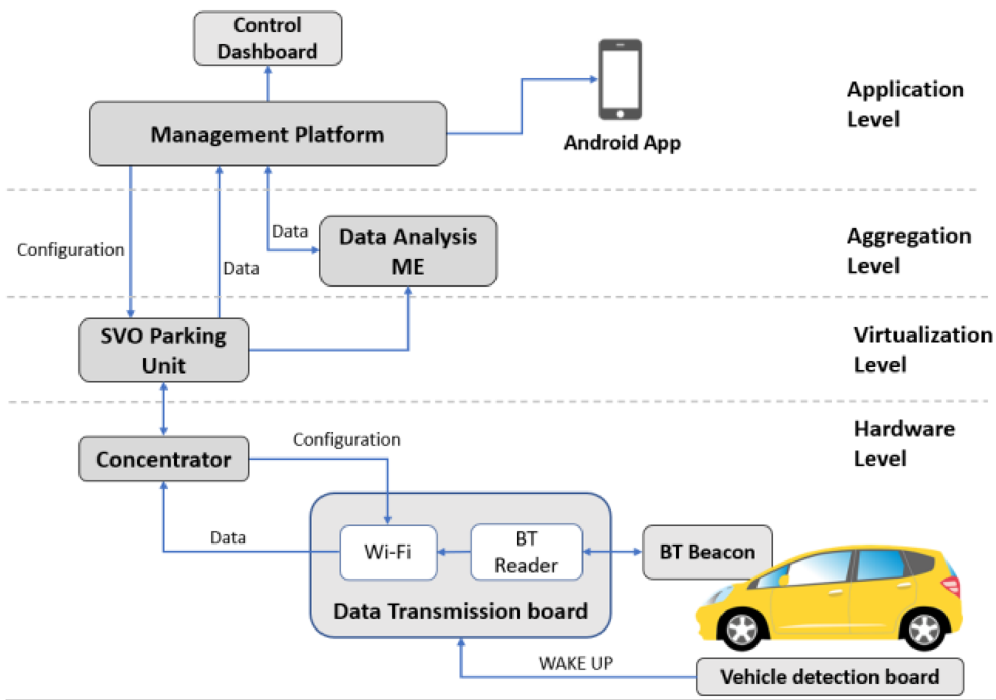


Figure 2.5: Architecture of the proposed Smart Parking system [13].

The results showed efficient and fast detection, high accuracy, good scalability, low power consumption and excellent interoperability. However, some areas are identified as weaknesses such as potential issues related to sensor accuracy, reliance on continuous data transmission, and the need for widespread adoption of the app and beacons.

In summary, this paper presents a comprehensive IoT-based smart parking solution, leveraging the Social Internet of Things, delivering significant improvements in efficiency, user experience and sustainability, recognizing the areas for improvement.

2.2.7 Multi-Modal Social Networks with IoT-Enabled Wearable Devices for Healthcare

- **Keywords**: : Social Networking, Wearable Devices, Healthcare, Chronic disease, Generalization, Convolution, and Sequential Neural Networks.
- **Authors**: “OM PRAKASH”, “RAJEEV KUMAR”.
- **Abstract**:

  In this paper [14], the authors proposed a unified ML architecture for generating alerts and monitoring chronic diseases in the context of SIoT as shown in figure 2.6. The proposed framework is composed of four main sub-blocks:

  – **Patient-side signal generation**: collection of patient health data, such as sleep patterns and vital signs, using wearable devices equipped with sensors for data collection, such as smartwatches.

  – **Information processing**: extract relevant information from the collected data, such as key characteristics.To achieve this, machine learning algorithms, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are implemented for feature extraction and data analysis.

  – **Transmission of data to healthcare providers**: the transmission of processed information to healthcare professionals, allowing them to monitor the health status of patients in real time.

  – **Alert-Based Actions**: this block involves taking actions in response to system-generated alerts, such as treatment recommendations or emergency medical interventions.

    Various classifiers, including Random Forest, Support Vector Machines, XGBoost, and Logistic Regression, were used to evaluate the framework on depression and ECG datasets. Logistic Regression achieved the best accuracy. The evaluation metrics included accuracy, precision, recall, and F1 score, demonstrate that the system is effective. Despite some challenges such as scalability, data security, interoperability, and user engagement, the proposed approach demonstrates considerable potential for enhancing patient care through real-time data processing and alert generation.

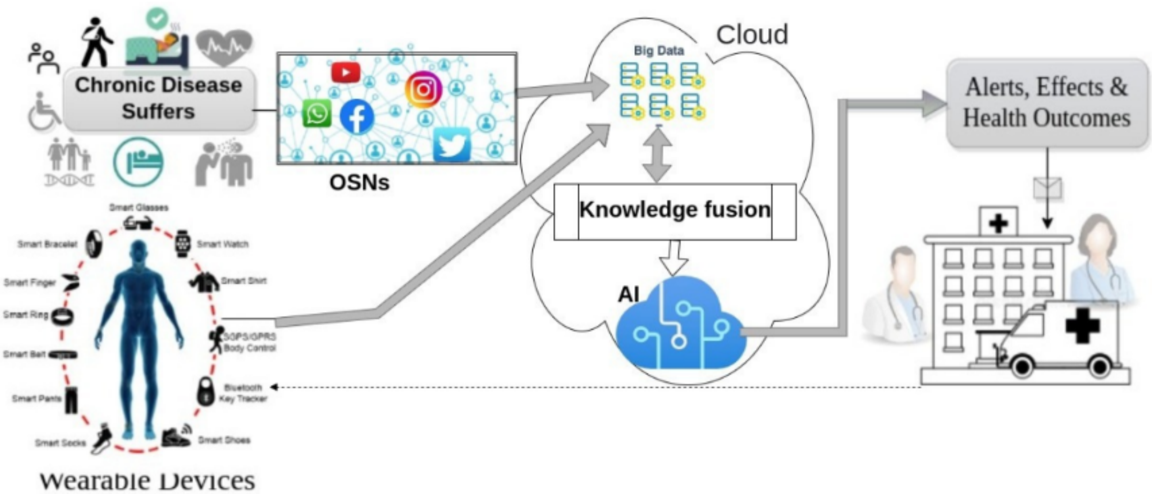


Figure 2.6: A Unified SIoT ML Architecture for Healthcare [14].

2.2.8 An Adaptive and Late Multimodal Fusion Framework in Contextual Representation based Evidential Deep Learning Dempster-Shafer Theory

- **Keywords**: : multimodal data fusion, modality, context-ware, late fusion, deep learning, uncertainty.
- **Authors**: “Doaa Mohey Eldin”, “Aboul Ella Hassanein”, “Ehab E Hassanien”.
- **Abstract**:

  In this paper [7], the authors proposed a framework for adaptive and late multimodal fusion in contextual representation, using the Dempster-Shafer theory of evidential deep learning. This framework addresses the challenges associated with integrating multiple decision-making and control modalities in intelligent systems, aiming to enhance decision-making and classification accuracy by efficiently combining information from various data sources in different contextual representations. The approach consists of two main fusion levels: Model-Based Fusion and Feature-Based Fusion, each with several layers as shown in figure 2.7.

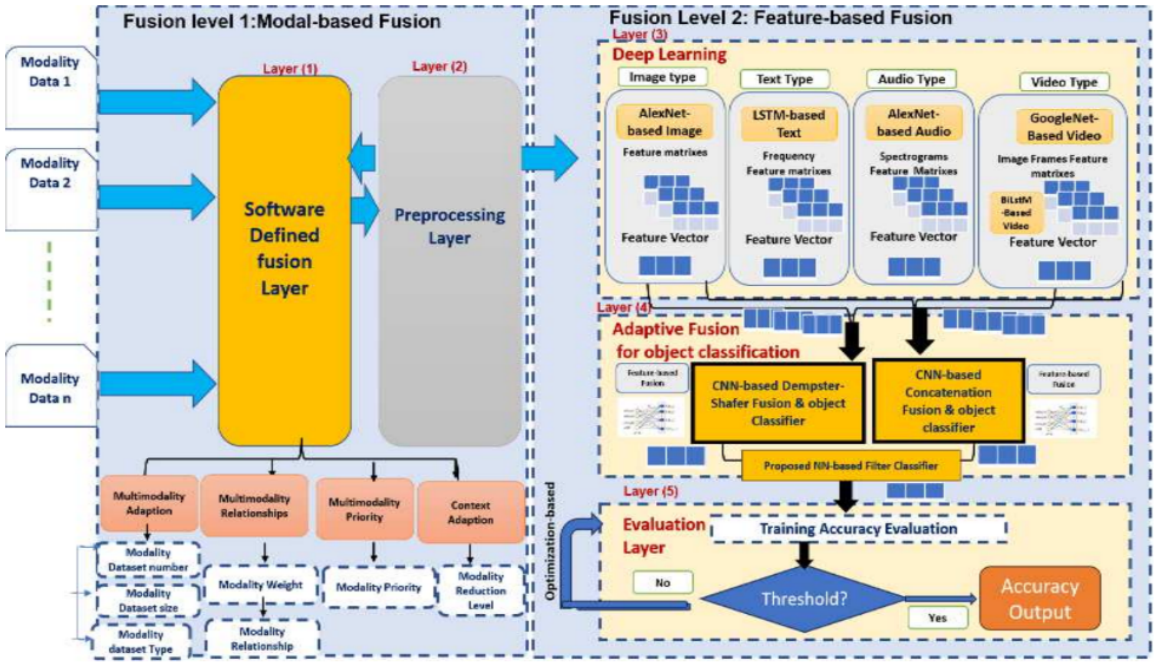


Figure 2.7: An Adaptive Multimodal Fusion Framework in Contextual Representation based on Late Fusion Level using MultiFusion Learning Model and Improved Evidential Deep Learning Dempster-Shafer [7].

Model-Based Fusion

- **Software-defined fusion layer**: this layer is responsible for extracting relationships and deducing weights for classification accuracy.
- **Preprocessing layer**: modalities are processed dynamicallyin this layer to prepare them for further analysis.

Feature-Based Fusion

- **Dynamic classification layer**: this layer involves creating deep learning models adapted for different input modalities.Techniques used include AlexNet for image data, LSTM for text data, and GoogleNet combined with BiLSTM for video data.
- **Adaptive fusion layer**: enhances the Dempster-Shafer fusion theory by integrating adaptive fusion techniques, improving the overall fusion process.
- **Evaluation layer**: assesses the performance of the fusion model in diverse contexts, demonstrating its effectiveness in improving decision-making processes and control in smart systems. Performance measures include accuracy, optimization through swarm techniques, and adaptivity to various input types and contexts.

The performance of the proposed system was evaluated using several metrics, including accuracy, optimization, and adaptivity. The system demonstrated high accuracy in classifying multiple modalities and contexts. The adaptive fusion model effectively handles data from different sources, providing a unified classification output. The system achieves significant improvements in decision-making processes across various smart applications.

Despite its strengths, the proposed framework has some limitations. The system's complexity requires substantial computational resources and sophisticated implementation strategies. Additionally, the framework's performance is optimal in certain predefined contexts, which may limit its generalizability to completely novel situations.

2.2.9 Multi-channel data flow software fault detection for social internet of things with system assurance concerns

- **Keywords**: : Cloud computing, Social internet of things, Data fow software, Fault detection, Ubiquitous clouds, System assurance.
- **Authors**: “Ling You”.
- **Abstract**:

  In this paper [15], the authors proposed a novel algorithm for fault detection using multi-channel data flow analysis and cloud computing technologies to ensure early detection of software failures in distributed systems, particularly within the Social Internet of Things (SIoT) as shown in figure 2.8. This method involves several key steps:

  - **Cloud Computing Technology**: utilizes virtualization to run multiple virtual machines on physical hosts, ensuring resource monitoring and sharing. Multiple copies of data are stored in distributed file and database systems to improve availability and reliability. Resource management simplifies the interface for developers, allowing them to focus on software without worrying about the underlying architecture.
  - **Data Flow Related Software Failure Model**: employs data flow analysis to monitor variable states and operations, detecting undefined or unused values that could cause failures. Data fusion technology integrates data from multiple sensors to identify degradation patterns and faults.

– **Fault Detection Technology**: uses clustering and data fusion models to enhance fault detection characteristics through numerical verification. Fault tolerance and error correction schemes are integrated to improve system robustness.

The proposed method achieved a fault detection rate of 92%, compared to 85% for traditional methods. The false positive rate was reduced to 3% from 7%. The average detection time was reduced from 15 seconds to 8 seconds, enabling faster response to faults and minimizing downtime. Graphs and simulations further illustrate the efficiency of the algorithm, showing consistent outperformance over traditional methods in various scenarios. On the other hand, this approach presents notable challenges such as the complexity of the proposed method and the resource intensity required to integrate cloud computing and data fusion models.

In conclusion, this paper presents a promising approach to software fault detection in distributed systems, leveraging cloud computing and multi-channel data stream analysis to improve detection accuracy and robustness. Despite the challenges, the results indicate significant potential for improving early fault detection.

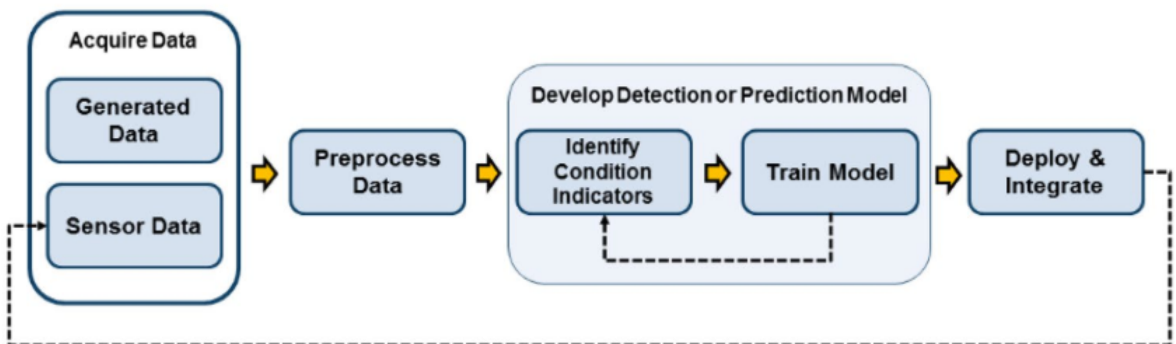


Figure 2.8: the fnalized framework for the detection process [15].

2.2.10 The Impact of AI Applications on Smart Decision Making in Smart Cities as Mediated by the Internet of Things and Smart Governance

- **Keywords**: Artificial intelligence, Internet of Things, big data, smart governance, smart decisionmaking, parallel-sequential multi-mediating effect.
- **Authors**: "SYED ASAD ABBAS BOKHARI","SEUNGHWAN MYEONG".

- **Abstract**:

  In this article [16], the authors proposed a comprehensive study on the impact of artificial intelligence (AI) applications on smart decision-making in smart cities, mediated by the Internet of Things (IoT) and smart governance, as shown in figure 2.9 . This study examines how AI, assisted by IoT and smart governance, influences intelligent decision-making in urban environments. this proposed method involves several key steps:

  - **Data Collection**: primary data was collected in South Korea from a diverse demographic.
  - **Data screening**: after filtering out incomplete responses, 516 usable samples were obtained.
  - **Analysis Tools**: SmartPLS version 4 was used in the study to analyze the relationships between IoT systems, smart governance, smart decision making and AI applications.
  - **Hypotheses Testing**: a number of hypotheses were tested, including the direct impact of AI on decision-making and the mediating roles of IoT and smart governance.

  the study found that AI applications have a positive direct impact on intelligent decision-making, with IoT systems and smart governance playing an important mediating role. Overall, the integration of AI, IoT and smart governance leads to a notable improvement in intelligent decision-making. The study has some limitations, such as limiting the geographic scope to South Korea, the cross-sectional nature of the data, and potential biases of self-reported data. Overall, the research provides valuable insights into how AI, IoT and smart governance collectively improve decision-making processes in smart cities.

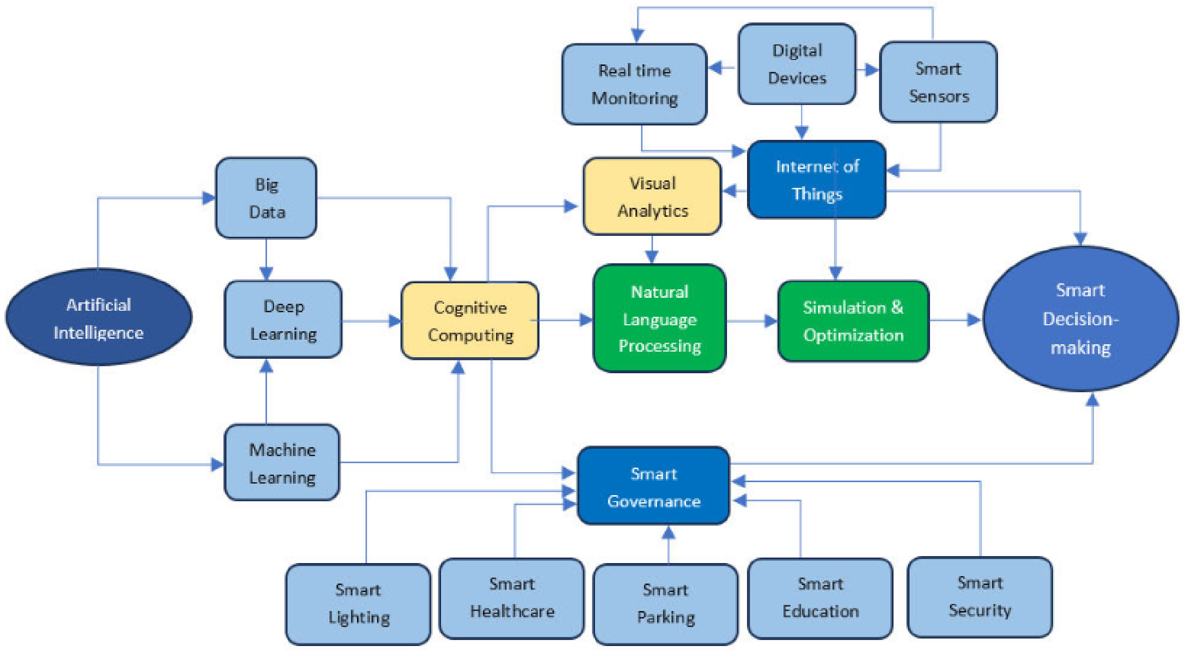


Figure 2.9: Conceptual framework of impacts by AI applications, the IoT, and smart governance on smart decision-making [16].

2.2.11 Boosted Barnacles Algorithm Optimizer: Comprehensive Analysis for Social IoT Applications

- **Keywords**: Social IoT, Barnacles Mating Optimizer, triangular mutation, opposition-based learning.
- **Authors**: "MOHAMMED A. A. AL-QANESS", "AHMED A. EWEES", "MOHAMED ABD ELAZIZ", "ABDELGHANI DAHOU", "MOHAMMED AZMI AL-BETAR", "AHMAD O. ASEERI", "DALIA YOUSRI", " REHAB ALI IBRAHIM".
- **Abstract**:

  The authors proposed in this paper [51], an improved method to address the challenges of high-dimensional data in Social Internet of Things (SIoT) applications. The proposed Dynamic Barnacles Mating Optimizer (DBMT) is an enhancement of the traditional Barnacles Mating Optimizer (BMO). It integrates two main techniques: Triangular mutation and dynamic Opposition-based learning (OBL), this method involves three main steps:

- **Data Collection**: the collection of datasets comes from various SIoT applications, including sensor data and social interaction data.
- **Preprocessing**: normalize and prepare data for analysis to ensure compatibility with the optimization algorithm.
- **Algorithm Application**:
  * **Barnacles Mating Optimizer (BMO)**: to mimic barnacle mating behavior for global optimization.
  * **Triangular Mutation**: to increase the diversity of solutions and prevent the algorithm from getting stuck in local optima.
  * **dynamic Opposition-based Learning**: to evaluate opposite solutions during the search process to increase the chances of finding the global optimum and further enhance the exploration capabilities of the algorithm.

DBMT was tested using datasets from the UCI Machine Learning Repository and SIoT-related datasets. The evaluation focused on predicting social-related datasets within the IoT environment. The results showed that it performed better than other existing algorithms in terms of accuracy, efficiency, and scalability.

Despite its promising results, the study highlights several limitations:

- **Data Dependency**: The performance of DBMT depends on the characteristics and quality of the input datasets, and the effectiveness of the method can be greatly impacted by poor data quality.
- **Computational Complexity**: Although DBMT improves efficiency, it still requires substantial computational resources, especially when dealing with large-scale datasets.
- **Generality**: The current evaluation may not fully generalized to all SIoT applications.

2.2.12 SocialNet of Things: A Ubiquitous Relationship Network Inspired by Social Space

- **Authors**: "Huansheng Ning", "Wenxi Wang", "Fadi Farha", "Jinsheng Xie", "Mahmoud Daneshmand".
- **Abstract**:

The article [17] introduces the concept of SocialNet of Things (SoT), which merges traditional and online social spaces through the integration of Internet of

Things (IoT) and social networking principles to create a ubiquitous relational network as shown in figure 2.10. This framework employs technologies such as cloud computing, edge computing, social computing and AI, to improve interactions and relationships between entities (things and people).

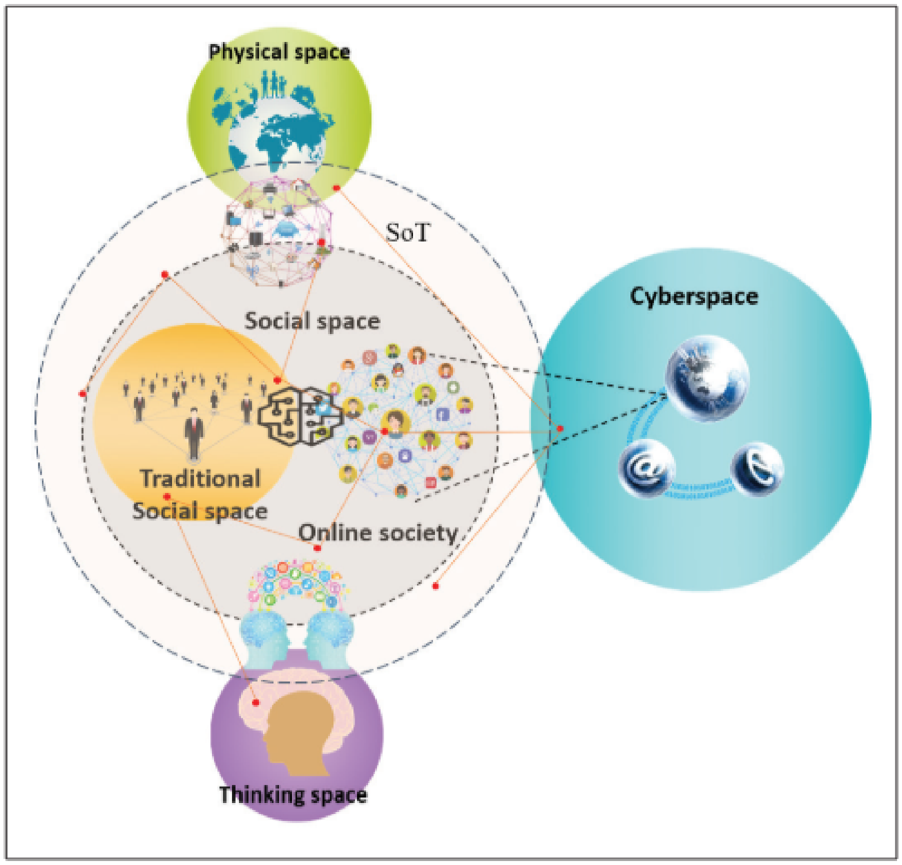


Figure 2.10: SoT determined by SocialNet and things (the shaded part is the space covered by SoT). [17].

The SoT framework involves several steps:

- **Ontology Modeling**: creating a structured framework to define and categorize entities and their relationships within the SoT. Ontology engineering is used to create a structured presentation of information in a specific field, which includes the objects, ideas, and connections among them.
- **Semantic Annotation**: involves adding metadata to data points in order to provide them with significance and context, this ensures compatibility between various systems through the utilization of semantic web technologies such as RDF (Resource Description Framework) and OWL (Web Ontology Language).

- **Relational Reasoning**: logical machine learning algorithms analyze patterns and predict outcomes to discover new relationships and insights in data.
- **Relationship Verification**: Validation processes and consistency checking algorithms are used to cross-reference relationships with known reliable data sources to ensure the reliability and validity of the inferred relationships.
- **Relational Management**: the use of database management systems (DBMS) and dynamic network analysis, to manage dynamic relationships within the SoT.
- **Service Discovery**: It is the process of recognizing and utilizing the services present in the network in order to respond effectively to user requests. This step utilizes service discovery protocols like Universal Plug and Play (UPnP) and Service Location Protocol (SLP), in addition to context-aware computing.

Two main scenarios were examined in order to assess the SoT framework's application in managing the COVID-19 pandemic. In resource allocation, intelligent algorithms predict the demand for medical supplies and integrate logistics information for efficient distribution. However, a unified network balances the distribution of medical personnel across regions using national data on doctors and patients. For close contact tracking, a contact tracing scheme is developed to monitor and predict the spread of COVID-19, integrating data from confirmed cases to optimize response strategies. Evaluation focuses on how accurate predictions are, the effectiveness of allocating resources like medical supplies on time and in sufficient quantities, and the ability of the system to adapt to changes and manage large quantities of data.

the article identifies three main limitations of SoT. First, scalability is a major challenge because SoT must be able to handle the dynamic increase in the number of interconnected entities and integrate various networks to provide intelligent and optimized services. Second, data management is more and more complex due to the explosion of data volumes from many sources, requiring good management and efficient fusion of heterogeneous data. Finally, privacy protection poses a crucial challenge, as SoT's interactions with traditional and online social spaces can lead to privacy violations. It is particularly difficult to anonymize data while maintaining its utility and integrity.

2.3 Comparative synthesis

We compare the different works mentioned previously based on some parameters that found below. Table 2.1 shows the results of this comparison.

- **Objective**: it defines the objective of the document.
- **Techniques**: techniques used in the approach proposed in the paper.
- **Metrics**: metrics used to assess the performance of the proposed models.
- **Results**: the results of the proposed approaches.
- **Limitations**: weaknesses or shortcomings identified in the document.

Table 2.1: Comparative Table of Some Approaches on Data Fusion in SIoT

Article & Year	**Objective**	**Techniques**	**Metrics**	**Results**	**Limitations**
Meghana et al. [9], (2021).	- Classification of Relationships.	- KNN, ANN, Decision Tree, Naïve Bayes.	- Accuracy, Precision, Recall, F1-score.	- Decision Tree performed well for all the device types with respect to accuracy and Precision.	- Potential scalability issues with increasing number of devices and data volume, Limited consideration of dynamic changes in device states and relationships.
Sahar Boulkaboul et al. [10], (2020).	- Develop a new data fusion method to handle conflicting and uncertain data.	- Dempster-Shafer theory, Adaptive weighted fusion algorithm, DFIOT.	- Accuracy, Efficiency, Robustness.	- DFIoT method effectively reduces conflicts and improves accuracy compared to existing approaches.	- High computational complexity, needs validation in other IoT domains.

Continued on next page

Authors & Year	**Objective**	**Techniques**	**Evaluation criteria**	**Results**	**Limitations**
Lakshmanaprabu et al. [11], (2018).	- Big Data Classification.	- Gabor filter, Hadoop Map-Reduce, EHO, LK-SVM	- Accuracy, Sensitivity.	- The proposed model attains a maximum accuracy of 98.86% compared to other existing approaches.	- Handling heterogeneous nature of SIoT, high memory consumption, processing power requirements.
SD et al. [12], (2022).	- Classification of services.	- R-ANN	- Accuracy, Precision, Recall, F1 Score, RMSE, MSE, MAE, R	- R-ANN model offers efficient and scalable knowledge management and service provision within SIoT networks.	- Challenges with data heterogeneity, scalability, and real-time processing.
Senthil Kumar et al. [50], (2020).	- Enhance security and trust in Social Internet of Things (SIoT) networks	- DeepChain (deep learning and blockchain)	- Trust transitivity (aggressive, conservative, traditional approaches), network performance (net profit).	- Quora emerges as the optimal network for fostering trust among SIoT nodes, boasting superior performance in success rate, node availability, and trustee potential.	- Complex search strategies, resource constraints, need for dynamic configurations, vulnerability to malicious trust attacks.
Floris et al. [13], (2022).	- Develop a Smart Parking system that optimizes space usage.	- Decision tree, Long Short-Term Memory recurrent neural network (LSTM RNN)	- Timeliness of Detection, Accuracy of Detection, Scalability, Energy Efficiency, Interoperability, User Experience.	- The LSTM RNN model is more suitable and reliable for predicting parking spot occupancy in a real smart parking scenario.	- Sensor accuracy issues due to environmental factors, dependency on continuous data transmission, need for widespread adoption of Android app and Bluetooth beacons.

Continued on next page

Authors & Year	Objective	Techniques	Evaluation criteria	Results	Limitations
Prakash et al. [14], (2023).	- Generate alerts and monitor chronic diseases.	- Residual neural networks (ResNet), XGBoost, Random Forest, Logistic Regression, Support Vector Machines(SVM)	- Accuracy, Precision, Recall, F1-score, AUC-ROC Curve.	- The logistic regression model achieves the highest precision at 84%.	- Scalability - Data security - Interoperability - User engagement.
Eldin et al. [7], (2023).	- Integrate various data modalities to improve classification accuracy in various contexts.	- Dempster-Shafer fusion theory, adaptive fusion techniques, deep learning models (AlexNet, LSTM, GoogleNet, BiLSTM).	- Accuracy, precision, recall, F1-score.	The proposed multimodal fusion framework achieves an average accuracy of 97.45%.	- Computational complexity, need for large datasets, optimal performance in predefined contexts.
You et al. [15], (2023).	- enhance early detection of faults in distributed systems, within the Social Internet of Things (IoT).	- Data fusion, Clustering, Numerical verification	- Reliability, Efficiency, Accuracy	- A significant improvement in the robustness and efficiency of distributed systems in cloud environments.	- Complexity of the proposed method, resource intensity, and scalability.

Continued on next page

Authors & Year	**Objective**	**Techniques**	**Evaluation criteria**	**Results**	**Limitations**
Bokhari et al. [16], (2023).	- Analyze the impact of AI applications on smart decision-making in smart cities, mediated by IoT and smart governance.	- AI, IoT, Smart governance, SmartPLS version 4	- Beta coefficients (β), t-statistics, p-values.	- AI has a positive impact on decision-making, IoT systems and smart governance. - IoT and governance also positively influence decision-making.	- Geographic scope confined to South Korea, cross-sectional design, potential biases from self-reported data
Al-qaness et al. [51], (2023).	- Evaluate the effectiveness of the DBMT approach for feature selection in IoT applications.	- Barnacles Algorithm Optimizer (BMO), Triangular Mutation (TM), Dynamic Opposition-based Learning (DOL).	- Accuracy, Sensitivity, Specificity, Average Fitness Value (Fitb).	- Best average fitness performance in 4/11 benchmark datasets. - Best minimum fitness values in 4/11 datasets. - Best classification accuracy in 36% of datasets.	- Potential computational complexity, risk of local optima,Data Dependency, Generality.
Ning et al. [17], (2022).	- Introduce the SocialNet of Things (SoT) integrating traditional and online social spaces with IoT and social networking principles.	- Cloud computing, edge computing, social computing, artificial intelligence (AI).	- Accuracy of predictions and the effectiveness of allocating medical resources and the system's ability to adapt to changes and manage large volumes of data.	- Accelerates the connection and convergence of spaces, application examples include resource allocation and close contact tracking in the fight against COVID-19.	- Scalability, data management, privacy protection.

2.4 Conclusion

In this state of the art chapter, we presented some existing work in the literature related to the research in SIoT based on data fusion. This study revealed that the application of data fusion in SIoT strengthens the capabilities of intelligent systems in terms of communication, analysis, and decision-making. We also note the intensive use of several machine and deep learning algorithms to boost the proposed approach. Our proposed approach, detailed in Chapter 3, aims to advance this state of the art by introducing a multi-stage data fusion methodology, which allows for more precise and reliable data aggregation, thereby improving the quality of the fused data and the performance of machine learning algorithms. Additionally, we employ ensemble algorithms such as Random Forest, Gradient Boosting and XGBoost for data classification, which have demonstrated superior performance in terms of precision, recall, and F1 score. Our comparative analysis shows a significant improvement in performance metrics compared to existing approaches, highlighting the effectiveness of our method. Furthermore, our approach proposes theoretical advancements in how data is fused and classified in the context of SIoT, providing a solid foundation for future research and more sophisticated applications. In summary, our proposed approach addresses several limitations identified in the current state of the art, making a significant contribution to the field of SIoT.

In the following chapter, we will explain and discuss in detail the proposed data fusion approach, to classify relationships between devices using machine learning techniques.

CHAPTER 3

PROPOSAL AND VALIDATION OF A DATA FUSION APPROACH IN SIOT ENVIRONMENT

3.1 Introduction

As we saw in the previous chapter, relationships between devices, such as trust, in the SIoT are crucial to building communities and working together. Determining this relationship is therefore a key issue. To this end, researchers have relied on ML algorithms for relationship classification based on device characteristics, such as device type, device brand, protocols, etc., by data aggregation. In addition, we need to evaluate the performance of these algorithms in order to choose the one best suited to a given application.

This chapter begins presenting the problematic and the proposed solution, then gives a description of the dataset considered in this study. Following this, we detail the proposed methodology. We then provide an overview of the hardware and software development tools and environment employed for our approach. Finally,we evaluate the results and the performance of different machine learning models in relationship classification.

3.2 Problematic

The development of the Social Internet of Things (SIoT) is leading to massive and diverse generation of data due to the increase in connected devices and their social interactions. These devices, whether they are environmental sensors, wearables, smart home systems or connected vehicles, produce data in large quantities and in various formats. The rapid evolution of SIoT amplifies this data production, making their management even more complex.

To integrate and exploit this data, efficient fusion methods are crucial. However, evaluating machine learning techniques for classifying complex relationships within SIoT is challenging. This requires methods capable of handling the variety of data formats, their dynamism, while guaranteeing the confidentiality and security of information, as well as high precision in the results obtained.

3.3 Proposed solution

SIoT generates a vast volume of heterogeneous data from diverse sources which presents a significant challenge in managing and analyzing this data. This complexity hinders the effective classification of relationships between connected devices (objects). Objects form the network, and relationship management determines the relationship between objects.

To address these challenges, we propose an approach that combines data fusion techniques with Machine Learning algorithms to effectively classify SIoT relationships. The process of data fusion will occur in two stages which have main advantages such as reduced redundancy, minimized traffic load, energy savings and information accuracy.

Before detailing the proposed fusion approach, we present the SIoT dataset we have relied on in this work to design the solution and make it easier for the reader to understand.

3.4 Dataset Description

The dataset used in this study was created by Marche et al [52] and consists of multiple CSV files describing real IoT objects in a Smart City environment. The main files are described below:

- **objects_description**: this file contains information on 16216 devices, of which 14600 belong to private users and 1616 to public services. The columns in this file include:
  - **id_device**: identifier of each device.

  - **id_user**: identifier of the device owner (0 for municipalities).
  - **device_type**: category associated with the device, represented by a numerical code from 1 to 16.
  - **device_brand**: device brand, represented by a numerical code from 1 to 12.
  - **device_model**: device model, represented by a numerical code from 1 to 24.

- **objects_profile**: describes the services offered and the applications required by each device category:
  - **device_type**: category of the device.
  - **id_off_service**: list of service IDs offered by the device, ranging from 1 to 18.
  - **id_req_app**: list of application IDs required by the device, ranging from 1 to 28.

- **private_static_devices** and **private_static_devices**: describe static devices with the following columns:
  - **id_device**: device ID.
  - **x**: X coordinate of the device's location.
  - **y**: Y coordinate of the device's location.

- **private_mobile_devices** and **public_mobile_devices**: describe mobile devices and their movement data. The columns include:
  - **timestamp_start**: start time of the idle state.
  - **timestamp_stop**: Eend time of the idle state.
  - **id_user/id_device**: user ID for private devices, or device ID for public devices.
  - **x**: X coordinate of the user's location.
  - **y**: Y coordinate of the user's location.

- **Adjacency Matrices**: these are the adjacency matrices that represent the relationships (OOR, C_LOR, SOR, SOR2) in an SIoT network.

3.5 Proposed Approach Methodology

Our approach is divided into three main steps, as it shown in the figure 3.1.

We begin with the initial data fusion process, followed by processing and integrating adjacency matrices representing these relationships into the dataset generated from the first fusion step, forming the second stage of data fusion. Finally, we apply various machine learning algorithms for effective relationship classification.

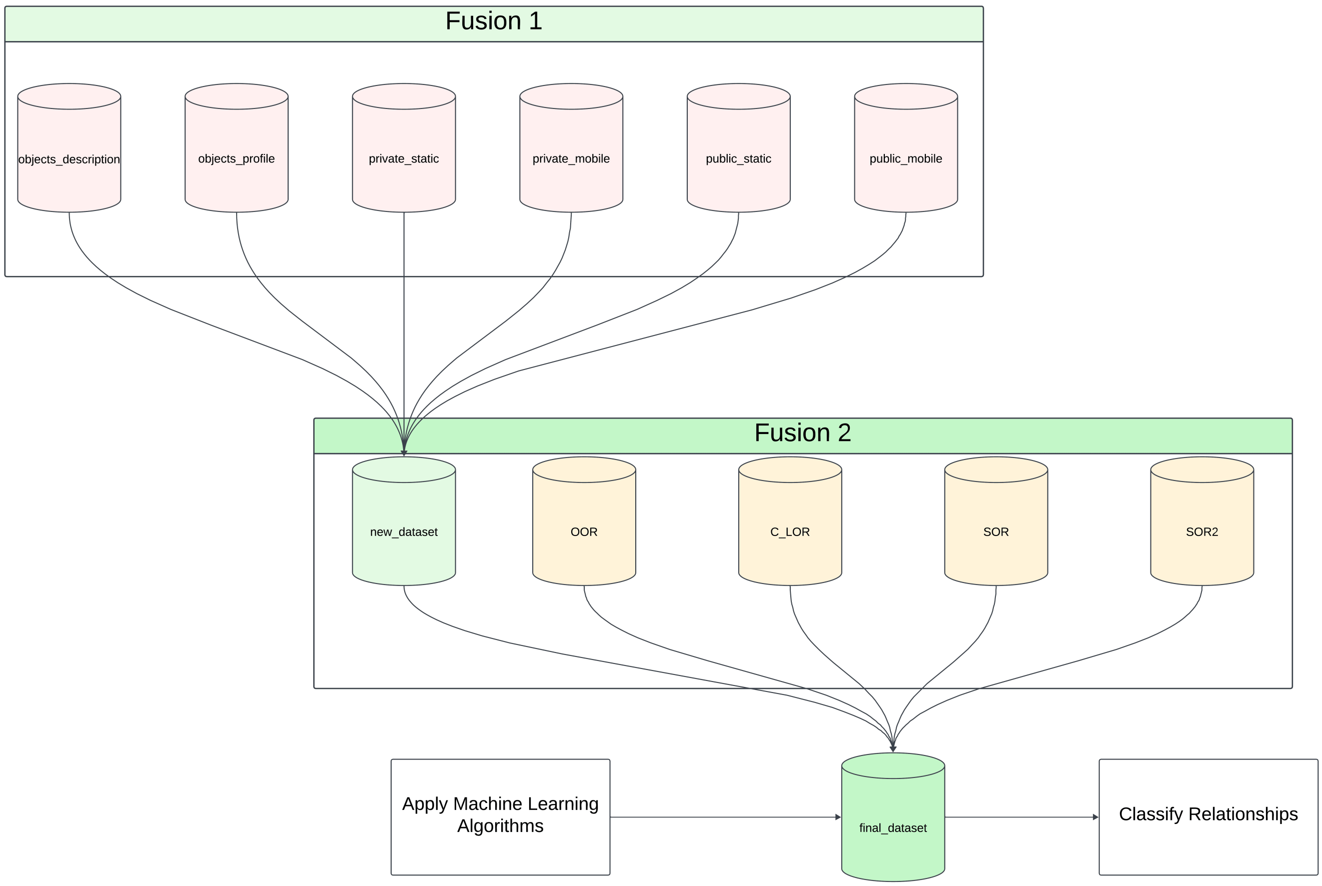


Figure 3.1: Overview of the Main Approach Steps.

3.5.1 Initial Data Fusion process

We drew inspiration from fuzzy logic principles to develop a max_operator function that calculates the maximum score by combining device brand and model attributes. This function was applied to the entire dataset, resulting in the creation of a new column named relevance_score. We used this metric to filter the data based on a defined threshold, keeping only devices that exceeded this score. Additionally, we filtered the data to retain only two instances per id_device, each having different device_brand values to ensure diversity. Next, we merged and concatenated different datasets to form a comprehensive data structure, as shown in the figure 3.2:

1. We merged the **objects_description** dataset with **objects_profile** on the **device_type** attribute to create a unified dataset referred to as **data**.

2. We concatenated **public static** and **mobile devices** and then merged them with **data** based on the **id_device** attribute, resulting in **public_df**.

3. We similarly merged **private static devices** with **data** and then concatenated the result with **public_df** to form the dataframe **df**.

4. We merged **private mobile devices** with **data** based on the **id_user** attribute and concatenated the result with **df** to generate a **new_dataset**.

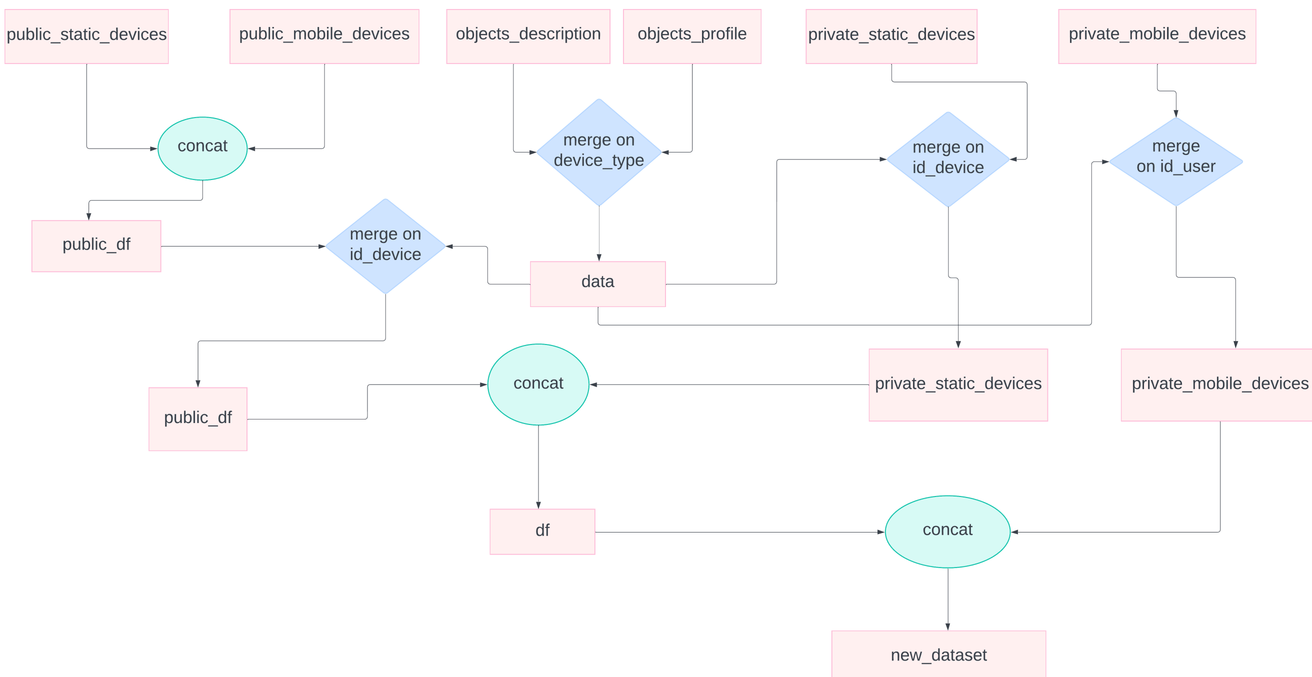


Figure 3.2: Overview of the Initial Data Fusion process.

3.5.2 Second Stage of Data Fusion

We represented relationships in the Social Internet of Things (SIoT) using adjacency matrices. Four types of relationships are considered in this dataset [52] :

- **OOR** (Ownership Object Relationship): type of relationship defined for objects owned by the same user. About public static devices, objects will create a relation only if they are in the communication range of each other. Public mobile objects don't create this type of relation.

- **C-LOR** (Co-Location Object Relationship): it is established between static devices (public or private) and private mobile located in the same place, and the number of meetings is more than 13.
- **SOR** (Social Object Relationship): this relationship is based on three parameters, that are the number of meetings (N = 3), the meeting duration (TM = 30 minutes) and the interval between two consecutive meetings (TI = 6 hours). The relation is created between private mobile devices.
- **SOR2** (Social Object Relationship): a variant of the SOR called SOR2 is created to connect the public mobile devices. In particular the relation is between public mobile devices and users' mobile objects. The parameters, as in the SOR, is set as follow: N = 3, TM = 1 minute and TI = 1 hour.

We processed adjacency matrix files in chunks to convert them into a long format suitable for merging. This approach allowed us to handle large datasets efficiently by breaking them down into smaller, more manageable pieces. Within each chunk, we reshaped the data using the **melt** function, which transformed the matrix into a long-format DataFrame with columns representing id_device_1, id_device_2, and their corresponding relationship value.

Next, we filtered the rows to include only those where the relationship value is 1, indicating the presence of a significant relationship between devices. We then updated the filtered DataFrame to retain only the relevant columns (**id_device_1**, **id_device_2**) and added a new column with the relation name, setting its value to 1. This step ensured that we clearly marked the presence of each specific relationship.

We concatenated these filtered relations to an accumulating DataFrame named **all_relations**. This allowed us to compile all significant relationships from the different chunks into a single comprehensive DataFrame. Finally, we merged chunks of data from the adjacency matrices with the new_dataset based on id_device, as it shown in the figure 3.1. We removed duplicate columns to ensure a clean and consistent final_dataset.

3.5.3 Machine Learning for Relationship Classification

Finally, we applied machine learning algorithms to classify the relationships between devices, following the steps shown in the figure 3.3:

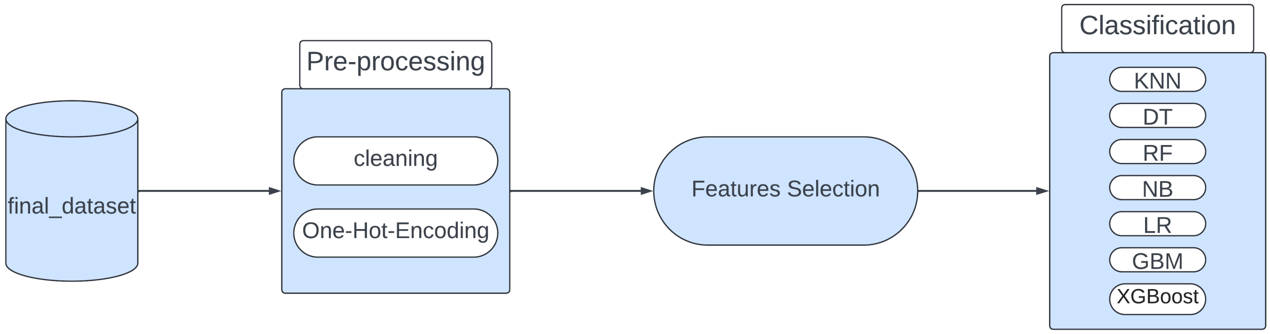


Figure 3.3: Machine Learning Workflow.

- **Pre-Processing**: we cleaned the dataset by removing missing values, then converted categorical data to numerical using one-hot encoding.
- **Features Selection**: we selected features, on which we trained the different models, including coordinate, id-device, id-user, device-type, device-brand, device-model, offered services and applications. The relationships (OOR, C_LOR, SOR, SOR2) served as the target variables for our classification task. We extracted these target columns into a separate DataFrame named targets, isolating the labels from the feature set to facilitate the training of the models.

  We then split the features into training and testing sets using an 80-20 split, ensuring that 80% of the data was used for training and 20% for testing.
- **Classification**: we normalized the data to ensure that all features were on a similar scale. Specifically, we used the **StandardScaler** for this purpose. After normalization, we applied seven Machine Learning algorithms for classification: KNN, Random Forest, Decision Tree, Naive Bayes, Logistic Regression, Gradient Boosting, and XGBoost. Each model was trained on the training set and tested on the test set.

  To handle the multi-label classification problem, we used the **MultiOutputClassifier** wrapper for each model. This approach allowed each model to predict multiple target labels simultaneously. After training, we evaluated each model's performance by making predictions on the test set. The evaluation was conducted using several metrics, including accuracy, recall, precision, and F1 score. Additionally, we analyzed the confusion matrices to gain deeper insights into the performance and misclassifications of each model.

3.6 Development Environment

3.6.1 Hardware Environment

Characteristics	Machine
Model	HP Pavilion Laptop 15-eg0xxx
Processor	Intel(R) Core(TM) i5-1135G7 @ 2.40GHz 2.42 GHz
RAM	8,00 Go
Operating System	Windows 10

Table 3.1: Characteristics of the machine used.

3.6.2 Software Environment

- **Anaconda**: Anaconda is an open-source distribution of the Python and R programming languages for data science that aims to simplify package management and deployment. It comes with over 250 packages automatically installed. Over 7500 additional open-source packages can be installed from PyPI. Anaconda also includes a GUI (graphical user interface) named Anaconda Navigator. It allows users to launch applications and manage conda packages, environments and channels without using command-line commands [53].

- **Spyder**: Spyder is a free and open source scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It features a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package [54].

- **Google Colaboratory**: Colab is a hosted Jupyter Notebook service that requires no setup to use and provides free access to computing resources, including GPUs and TPUs. Colab is especially well suited to machine learning, data science, and education [55].

- **Lucidchart**: Lucidchart is a web-based diagramming application that allows users to visually collaborate on drawing, revising and sharing charts and diagrams, and improve processes, systems, and organizational structures [56].

- **Python**: Python is an interpreted, object-oriented, high-level programming language, appreciated for its simplicity and readability. It offers built-in data structures, promotes modularity, and code reuse through modules and packages. Python can be used for rapid application development and as a scripting language. Its interpreter and standard library are free and available on various platforms [57].

- **NumPy**: NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more [58].

- **Pandas**: Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language [59].

- **Scikit-learn**: Scikit-learn, is an open-source, machine learning and data modeling library for Python. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python libraries, NumPy and SciPy [60].

- **Matplotlib**: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible [61].

3.7 Results and Discussion

In this section, we present the results of our approach on classifying relationships between SIoT devices using different machine learning models. Each model was evaluated in terms of accuracy, recall, precision, F1 Score and Confusion Matrix in order to compare their performances. The results are shown in the table 3.2

Model	Accuracy	Recall	Precision	F1 Score
KNN	0.9225	0.9532	0.9659	0.9595
Decision Tree	0.9026	0.9419	0.9429	0.9424
Random Forest	0.9384	0.9594	0.9732	0.9662
Naive Bayes	0.7501	0.9942	0.7529	0.8446
Logistic Regression	0.9210	0.9551	0.9651	0.9600
Gradient Boosting	0.9300	0.9569	0.9620	0.9594
XGBoost	0.9319	0.9565	0.9697	0.9629

Table 3.2: Classification Results

3.7.1 Model Performance Analysis

Through Table 3.2 and Figure 3.10, we examine the performance metrics of various machine learning models.

- **KNN** demonstrated a high level of accuracy (92.25%), with impressive recall (95.32%), precision (96.59%), and F1 score (95.95%).

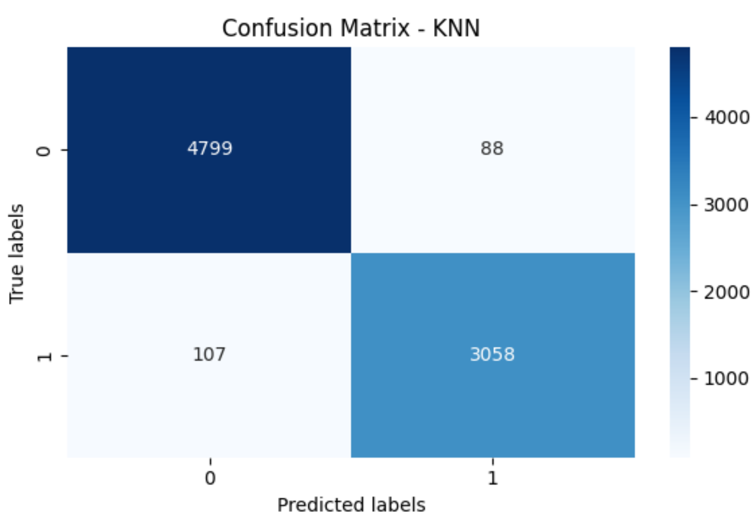


Figure 3.4: Confusion Matrix for KNN Model.

The confusion matrix shown in the figure 3.4 further supports its efficiency in correctly classifying positive and negative cases, with a minor fraction of misclassifications.

- **Decision Tree** showed a slightly lower performance compared to KNN, with an accuracy of 90.26%, recall of 94.19%, precision of 94.29%, and F1 score of 94.24%.

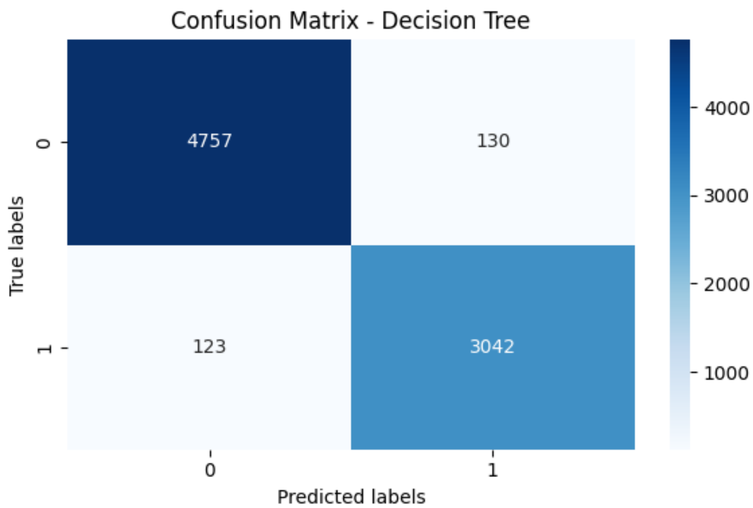


Figure 3.5: Confusion Matrix for Decision Tree Model.

The confusion matrix shown in the figure 3.5 indicates a balanced misclassification rate between false positives and false negatives.

- **Random Forest** outperformed other models in accuracy (93.84%), recall (95.94%), precision (97.32%), and F1 score (96.62%).

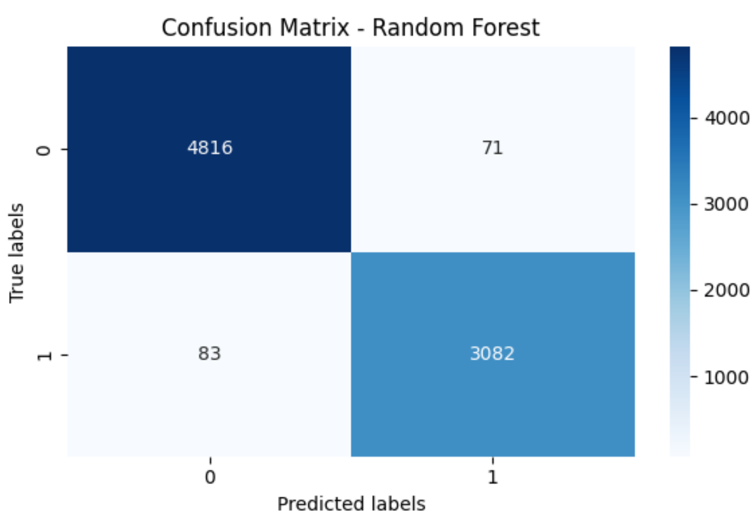


Figure 3.6: Confusion Matrix for Random Forest Model.

Its confusion matrix shown in the figure 3.6 presents the lowest rates of false positives and false negatives, highlighting its robustness in handling the classification task.

- **Naive Bayes** had the lowest accuracy (75.01%) but the highest recall (99.42%), with a precision of 75.29% and an F1 score of 84.46%.

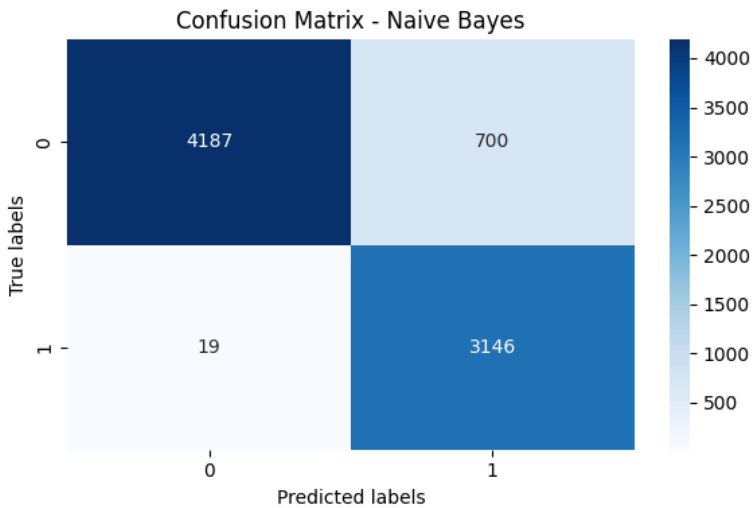


Figure 3.7: Confusion Matrix for Naive Bayes Model.

The confusion matrix shown in the figure 3.7 reveals a significant number of false positives, which negatively impacts its precision.

- **Logistic Regression** offered competitive results with an accuracy of 92.10%, recall of 95.51%, precision of 96.51%, and F1 score of 96.00%.

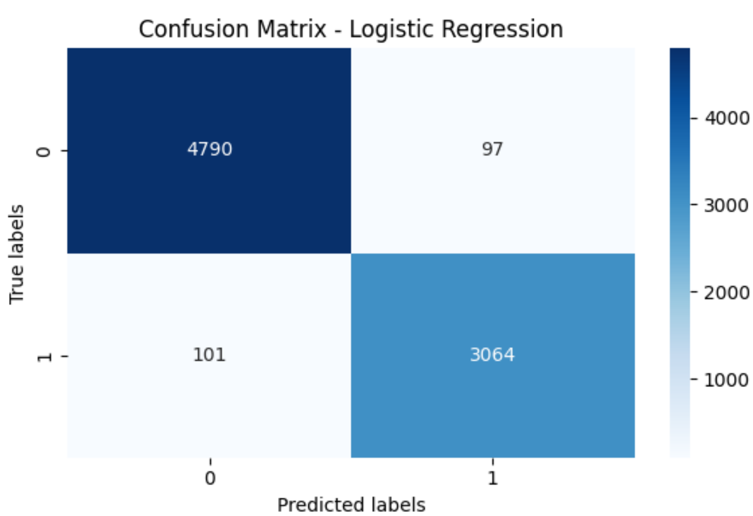
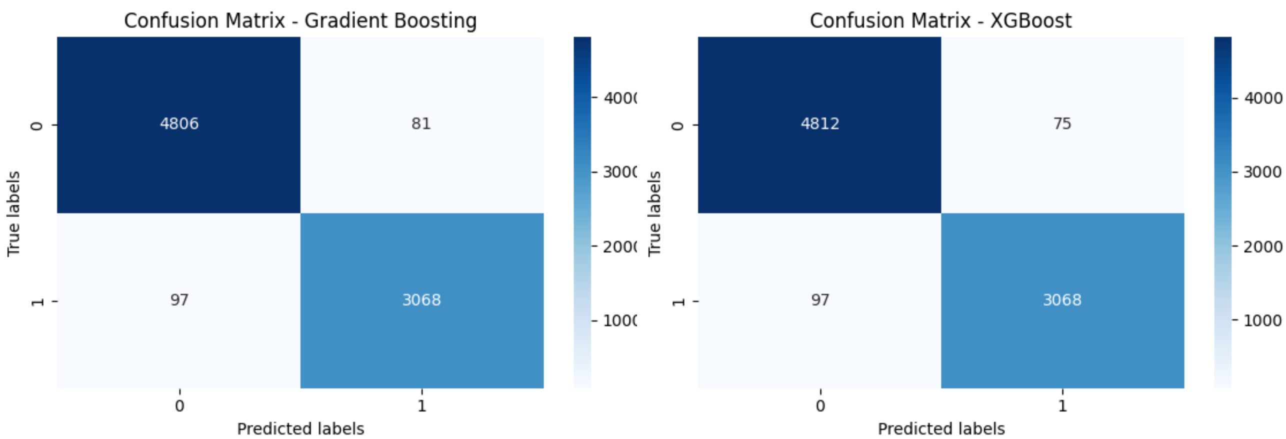


Figure 3.8: Confusion Matrix for Logestic Regression Model.

Its confusion matrix shown in the figure 3.8 shows a balanced distribution of misclassifications, similar to KNN.

- **Gradient Boosting** and **XGBoost** presented very close performances, with Gradient Boosting achieving an accuracy of 93.00%, recall of 95.69%, precision of 96.20%, and F1 score of 95.94%. XGBoost slightly improved on these metrics, indicating its efficiency in the classification task.



(a) Confusion Matrix for GBM Model

(b) Confusion Matrix for XGBsoot Model

Figure 3.9: Confusion Matrix for GBM and XGBoot Models

The confusion matrices of both models, shown in Figure 3.9 indicate few misclassifications, confirming their high accuracy rates.

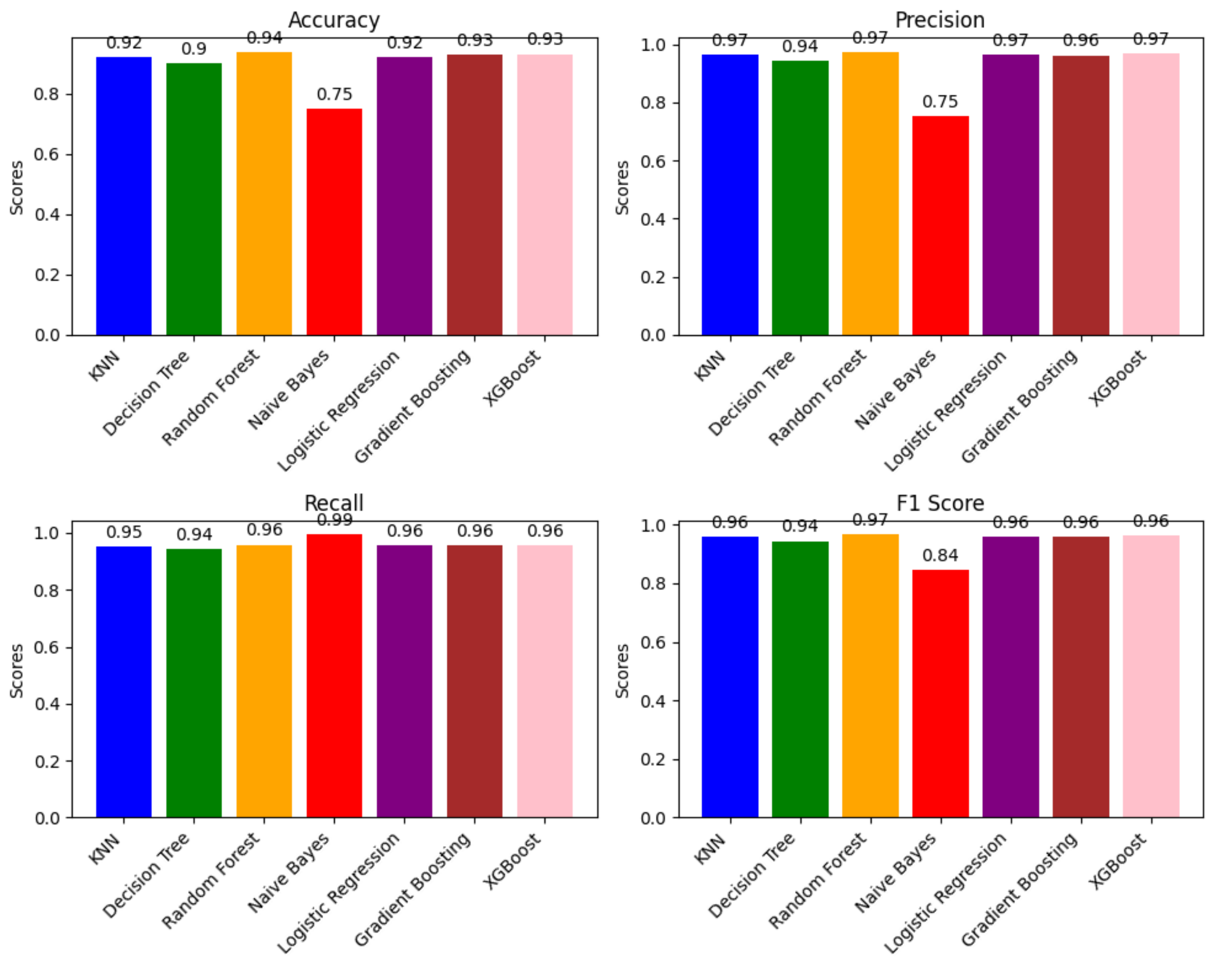


Figure 3.10: Metrics Evaluation.

3.7.2 Discussion

The results indicate that ensemble methods, particularly Random Forest and XGBoost, offer superior performance in terms of accuracy, precision, recall, and F1 score. Among these, Random Forest stands out as the top-performing model.

Naive Bayes, despite its high recall, suffers from a significant number of false positives, leading to the lowest precision and accuracy among the models tested.

Decision Trees provide a good balance between recall and precision but do not reach the performance levels of their ensemble methods.

KNN and Logistic Regression show strong performances. However, they may not always achieve the high levels of accuracy and balance between precision and recall that ensemble methods can.

3.7.3 Analysis of Relationships Frequency

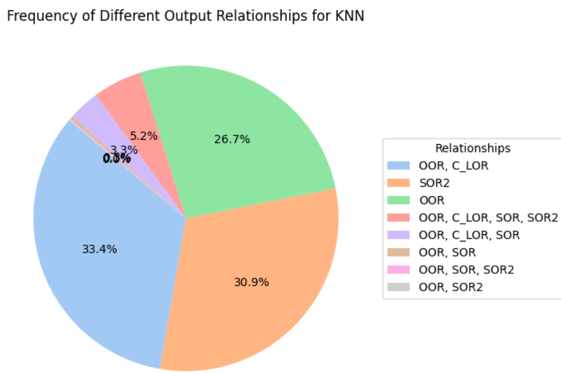
The classification of relationships among SIoT devices reveals interesting patterns across different machine learning models .

- **Dominant Relationships for KNN, DT, RF, LR, GBM, and XGBoost**: (OOR, C_LOR), (SOR2), and (OOR), as it shown in the figure 3.11.
- **Dominant Relationships for Naive Bayes**: (OOR, C_LOR), (SOR2), and (OOR, C_LOR, SOR, SOR2), as it shown in the figure 3.12.

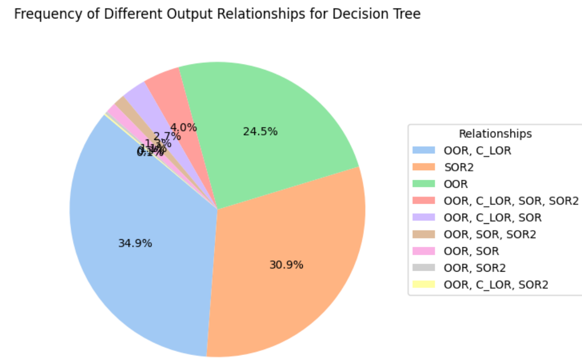
The frequent identification of **OOR (Ownership Object Relationship)** and C-LOR **(Co-Location Object Relationship)** relationships across KNN, Decision Tree, Random Forest, Logistic Regression, Gradient Boosting, and XGBoost models can be attributed to the inherent structure of the dataset. These models effectively recognize patterns of ownership and repeated interactions, which are common in smart city environments where devices frequently interact within their communication ranges.

The SOR2 (Social Object Relationship variant) is also commonly identified by these models, suggesting their capability to detect social interactions between public and user mobile devices, even with short meeting durations and intervals.

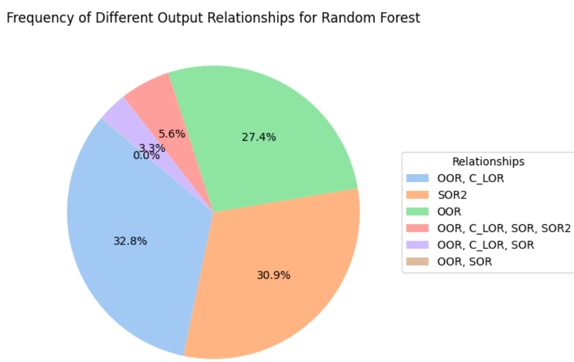
Naive Bayes, however, identifies a broader range of relationships, including **SOR (Social Object Relationship)**, in addition to those detected by other models. This is likely due to its probabilistic nature, which makes it sensitive to capturing patterns in repeated interactions over time. Although this sensitivity may contribute to a higher rate of false positives, it allows Naive Bayes to detect additional relationships that other models might miss.



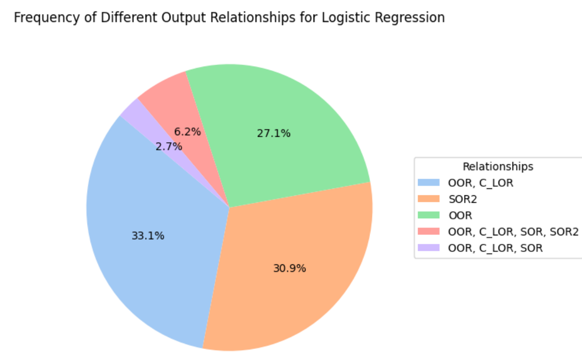
(a) KNN



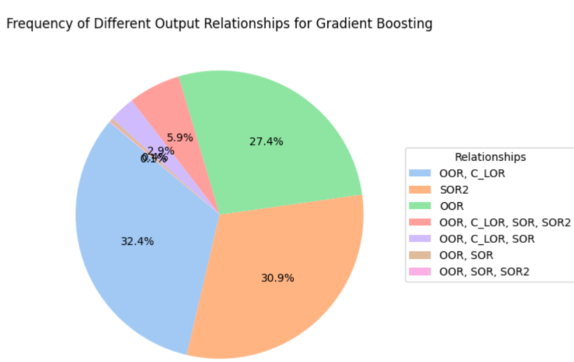
(b) Decision Tree



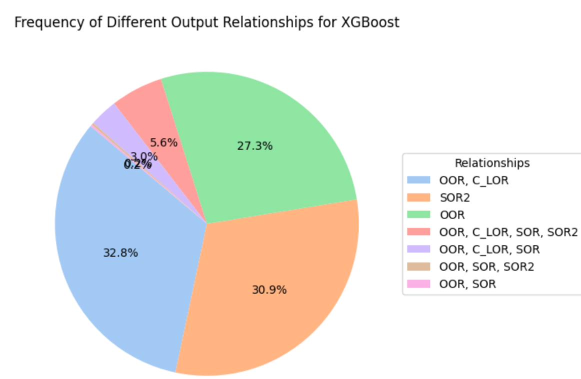
(c) Random Forest



(d) Logistic Regression



(e) Gradient Boosting



(f) XGBoost

Figure 3.11: Frequency of Different Output Relationships for ML models

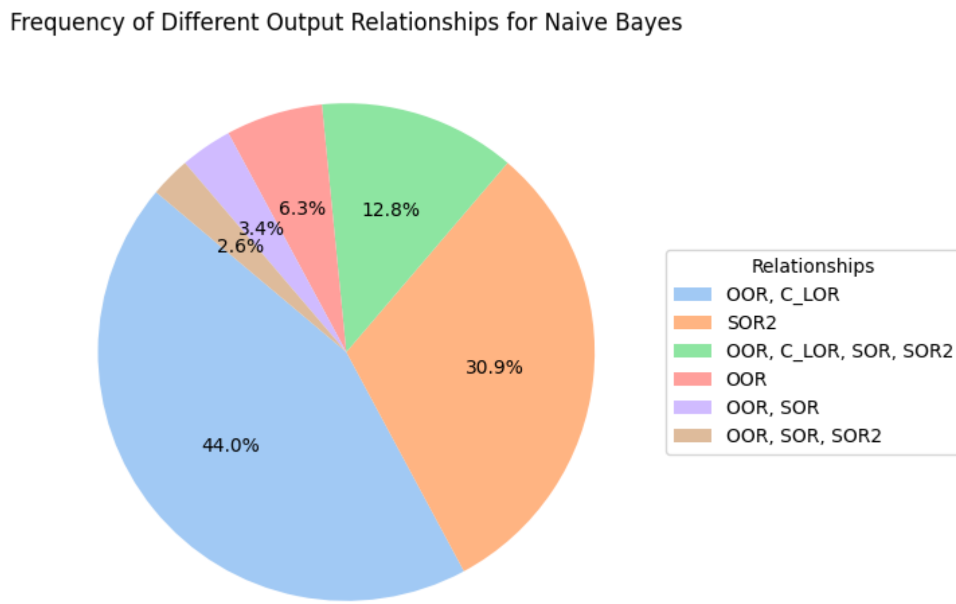


Figure 3.12: Frequency of Different Output Relationships for Naive Bayes model.

3.7.4 Comparison

For a more significant evaluation, we compared our approach with the methodology proposed by Meghana et al (2021) [9], who aggregated the dataset based on device types into public and private devices. The performance metrics are summarized in Table 3.3 below.

3.7.4.1 Analysis of Results

KNN: Our approach achieved higher accuracy (0.9225) compared to the private dataset approach (0.8456) but slightly lower accuracy than the public dataset approach (0.9967). The recall for our approach (0.9532) outperformed both the private and the public dataset approach.

Decision Tree: Our approach demonstrated strong performance across all metrics, particularly in recall, precision, and F1-score. However, in terms of accuracy, our approach achieved a score of 0.9026, which exceeded the private dataset's accuracy of 0.8283 and was comparable to the public dataset's accuracy of 0.9967.

Random Forest: Our approach surpassed both the private and public dataset approach in terms of precision, recall, and F1 Score. This suggests that our Random

Model	Metrics Evaluation	Our Approach	Private	Public
KNN	Accuracy	0.9225	0.8456	0.9967
	Recall	0.9532	0.8841	0.7500
	Precision	0.9659	0.8851	0.7473
	F1 Score	0.9595	0.8840	0.7487
Decision Tree	Accuracy	0.9026	0.8283	0.9967
	Recall	0.9419	0.8676	0.7491
	Precision	0.9429	0.8634	0.7482
	F1 Score	0.9424	0.8654	0.7487
Random Forest	Accuracy	0.9384	0.8683	0.9978
	Recall	0.9594	0.8887	0.7500
	Precision	0.9732	0.8977	0.7482
	F1 Score	0.9662	0.8930	0.7491
Naive Bayes	Accuracy	0.7501	0.5649	0.9858
	Recall	0.9942	0.9965	0.7500
	Precision	0.7529	0.6171	0.7389
	F1 Score	0.8446	0.7309	0.7443

Table 3.3: Performance Comparison

Forest model generalized better to unknown data than the existing approach.

Naive Bayes: While our approach demonstrated impressive recall at 0.9942, the existing approach with the private dataset showed overall higher recall at 0.9965, while the public dataset approach had a recall of 0.7500. Nevertheless, our approach achieved a higher F1 Score than the public dataset approach, indicating better balance between precision and recall in our approach.

3.7.5 Conclusion

This chapter introduced the data fusion classification approach based on machine learning technique. It presented the obtained results for the different algorithms. The comparison of our results with those [9] shows clearly that our approach offers significant advantages, particularly in better generalization and overall performance for KNN, Decision Tree, and Random Forest models.

CONCLUSION AND PERSPECTIVES

The aim of this thesis was to propose a data fusion method for classifying relationships between SIoT objects using machine learning algorithms.

We therefore first reviewed the concepts of SIoT, data fusion and machine learning techniques. Next, we looked at recent work in the literature combining the previous concepts. This study revealed the problem of classifying relationships between SIoT objects.

To deal with this latter issue, we proposed an data fusion approach based on machine learning algorithms as follows. Initially, we developed a max_operator function inspired by fuzzy logic principles to merge various datasets, forming a comprehensive data structure. The second stage involves processing adjacency matrices to represent different relationship types within the SIoT, followed by integrating these matrices into the dataset. Finally, we apply seven machine learning algorithms to classify these relationships. The results demonstrate that Random Forest outperforms other ML models. The high accuracy, precision, recall, and F1 scores achieved by these models highlight their effectiveness in handling the complexity of SIoT relationship classification.

Future work can focus on several perspectives to enhance this approach further. First, exploring deep learning techniques could provide new insights and potentially better results. Second, expanding the dataset to include more diverse types of devices and relationships can help generalize the models and improve their applicability to various SIoT scenarios. Finally, creating autonomous data fusion systems capable of self-managing, self-learning, and self-optimizing without human intervention. These systems can adapt to changing environments and data sources independently.

BIBLIOGRAPHY

[1] https://futureofiot.wordpress.com/2020/12/23/future-of-iot/, (Consulted on 12 October 2023). v, 6

[2] L. Atzori, A. Iera, and G. Morabito, "From" smart objects" to" social objects": The next evolutionary step of the internet of things," *IEEE Communications Magazine*, vol. 52, no. 1, pp. 97–105, 2014. v, 10

[3] B. H. Laallam Fatima Zohra, "Contribution à l'amélioration des stratégies de l'entreprise en utilisant social internet des objets," Mémoire en Intelligence artificielle, University Kasdi Merbah Ouargla, 2022. v, 7, 8, 9, 10, 11

[4] R. K. Chahal, N. Kumar, and S. Batra, "Trust management in social internet of things: A taxonomy, open issues, and challenges," *Computer Communications*, vol. 150, pp. 13–46, 2020. v, 13, 14

[5] F. Amin, A. Majeed, A. Mateen, R. Abbasi, and S. O. Hwang, "A systematic survey on the recent advancements in the social internet of things," *IEEE Access*, vol. 10, pp. 63 867–63 884, 2022. v, 14, 15

[6] A. Martin, *La fusion d'informations*. Polycopié de cours ENSIETA-Réf, 2005, vol. 1484. v, 16, 17

[7] D. M. Eldin, A. E. Hassanein, and E. E. Hassanien, "An adaptive and late multi-modal fusion framework in contextual representation based evidential deep learning dempster-shafer theory," 2023. v, vi, 19, 20, 34, 35, 45

[8] J. Peng, E. Jury, P. Dönnes, and C. Ciurtin, "Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases:

Applications and challenges," *Frontiers in Pharmacology*, vol. 12, 09 2021. v, 21, 22, 23

[9] J. Meghana, J. Hanumanthappa, and S. S. Prakash, "Performance comparison of machine learning algorithms for data aggregation in social internet of things," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 212–219, 2021. v, 25, 26, 43, 64, 65

[10] S. Boulkaboul and D. Djenouri, "Dfiot: Data fusion for internet of things," *Journal of Network and Systems Management*, vol. 28, p. 1136–1160, 2020. [Online]. Available: https://doi.org/10.1007/s10922-020-09519-y v, 26, 27, 43

[11] S. Lakshmanaprabu, K. Shankar, A. Khanna, D. Gupta, J. J. Rodrigues, P. R. Pinheiro, and V. H. C. De Albuquerque, "Effective features to classify big data using social internet of things," *IEEE access*, vol. 6, pp. 24 196–24 204, 2018. v, 27, 28, 44

[12] M. SD, S. S. Prakash, and K. Krinkin, "Service oriented r-ann knowledge model for social internet of things," *Big Data and Cognitive Computing*, vol. 6, no. 1, p. 32, 2022. v, 29, 30, 44

[13] A. Floris, S. Porcu, L. Atzori, and R. Girau, "A social iot-based platform for the deployment of a smart parking solution," *Computer Networks*, vol. 205, p. 108756, 2022. v, 31, 32, 44

[14] O. Prakash and R. Kumar, "Multi-modal social networks with iot-enabled wearable devices for healthcare," *INFOCOMP Journal of Computer Science*, vol. 22, no. 1, 2023. vi, 13, 33, 34, 45

[15] L. You, "Multi-channel data flow software fault detection for social internet of things with system assurance concerns," *International Journal of System Assurance Engineering and Management*, vol. 14, no. Suppl 1, pp. 472–482, 2023. vi, 36, 37, 45

[16] S. A. A. Bokhari and S. Myeong, "The impact of ai applications on smart decision-making in smart cities as mediated by the internet of things and smart governance," *IEEE Access*, vol. 11, pp. 120 827–120 844, 2023. vi, 38, 39, 46

[17] H. Ning, W. Wang, F. Farha, J. Xie, and M. Daneshmand, "Socialnet of things: a ubiquitous relationship network inspired by social space," *IEEE Network*, vol. 36, no. 3, pp. 197–203, 2022. vi, 40, 41, 46

[18] W. Hadjadj, "L'utilisation de n-version de programmation pour la prise en charge des fautes dans un environnement iot: Étude de cas sur un système médical domotique

contrôlé par un sma," Mémoire en Architecture Distribuée, Université Larbi Ben M'hidi Oum El Bouaghi, 2018. 4

[19] https://iot.ieee.org/definition.html, (Consulted on 14 October 2023). 4

[20] Y. Challal, "Sécurité de l'internet des objets : vers une approche cognitive et systémique," Thèse de doctorat en Informatique, Université de Technologie de Compiègne, 2012. 4

[21] https://fr.wikipedia.org/wiki/Capteur/, (Consulted on 15 October 2023). 5

[22] https://iotjourney.orange.com/en/explore/solutions-iot/connected-object/, (Consulted on 2 June 2024). 6

[23] https://www.maxirobots.com/blog/linvasion-objets-connectes/, (Consulted on 23 November 2023). 6

[24] F. Hidouci, "Réalisation et implémentation d'une application a base de protocole mqtt dans iot," Mémoire en Informatique, Université Mohamed Khider de Biskra, 2019. 6

[25] https://iotjourney.orange.com/en/explore/solutions-iot/smart-objects/, (Consulted on 2 Juin 2024). 7

[26] S. Sagar, "Gestion intelligente de réseaux de capteurs, intégrés à des vêtements sportifs instrumentés," Thèse de doctorat en Informatique, Ecole nationale supérieure Mines-Télécom Atlantique, 2019. 7

[27] S. Duquennoy, G. Grimaud, and J.-J. Vandewalle, "The web of things: interconnecting devices with high usability and performance," in *2009 International Conference on Embedded Software and Systems.* IEEE, 2009, pp. 323–330. 9

[28] J. He, Y. Zhang, G. Huang, and J. Cao, "A smart web service based on the context of things," *ACM Transactions on Internet Technology (TOIT)*, vol. 11, no. 3, pp. 1–23, 2012. 9

[29] L. E. Holmquist, F. Mattern, B. Schiele, P. Alahuhta, M. Beigl5, and H.-W. Gellersen, "Smart-its friends: A technique for users to easily establish connections between smart artefacts," in *Ubicomp 2001: Ubiquitous Computing: International Conference Atlanta Georgia, USA, September 30–October 2, 2001 Proceedings 3.* Springer, 2001, pp. 116–122. 9

[30] D. M. Black, *Why things matter: The place of values in science, psychoanalysis and religion.* Routledge, 2011. 9

[31] E. Nazzi and T. Sokoler, "Walky for embodied microblogging: sharing mundane activities through augmented everyday objects," in *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, 2011, pp. 563–568. 9

[32] M. Kranz, L. Roalter, and F. Michahelles, "Things that twitter: social networks and the internet of things," in *What can the Internet of Things do for the Citizen (CIoT) workshop at the eighth international conference on pervasive computing (Pervasive 2010)*, 2010, pp. 1–10. 9

[33] K. Evangelos A, T. Nikolaos D, and B. Anthony C, "Integrating rfids and smart objects into a unifiedinternet of things architecture," *Advances in Internet of Things*, vol. 2011, 2011. 10

[34] J. An, X. Gui, W. Zhang, and J. Jiang, "Nodes social relations cognition for mobility-aware in the internet of things," in *2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing*. IEEE, 2011, pp. 687–691. 10

[35] M. Roopa, S. Pattar, R. Buyya, K. R. Venugopal, S. Iyengar, and L. Patnaik, "Social internet of things (siot): Foundations, thrust areas, systematic review and future directions," *Computer Communications*, vol. 139, pp. 32–57, 2019. 11

[36] L. Atzori, A. Iera, G. Morabito, and M. Nitti, "The social internet of things (siot)–when social networks meet the internet of things: Concept, architecture and network characterization," *Computer networks*, vol. 56, no. 16, pp. 3594–3608, 2012. 12

[37] L. Xing, P. Zhao, J. Gao, H. Wu, and H. Ma, "A survey of the social internet of vehicles: Secure data issues, solutions, and federated learning," *IEEE Intelligent Transportation Systems Magazine*, vol. 15, no. 2, pp. 70–84, 2022. 13

[38] X. Liu, "Blockchain-enabled collaborative edge computing for intelligent education systems using social iot," *International Journal of Distributed Systems and Technologies (IJDST)*, vol. 13, no. 7, pp. 1–19, 2022. 13

[39] X. Yao, N. Ma, J. Zhang, K. Wang, E. Yang, and M. Faccio, "Enhancing wisdom manufacturing as industrial metaverse for industry and society 5.0," *Journal of Intelligent Manufacturing*, vol. 35, no. 1, pp. 235–255, 2024. 14

[40] M. Cimperman, A. Dimitriou, K. Kalaboukas, A. S. Mousas, and S. Quattropani, "Siot for cognitive logistics: Leveragung the social graph of digital twins for effective

operations on real-time events," *ITU J. Future Evol. Technol*, vol. 2, pp. 69–79, 2021. 14

[41] R. de Kervenoael, A. Schwob, R. Hasan, and E. Psylla, "Siot robots and consumer experiences in retail: Unpacking repeat purchase intention drivers leveraging computers are social actors (casa) paradigm," *Journal of Retailing and Consumer Services*, vol. 76, p. 103589, 2024. 14

[42] C. K. Panda and R. Bhatnagar, "Social internet of things in agriculture: an overview and future scope," *Toward Social Internet of Things (SIoT): Enabling Technologies, Architectures and Applications: Emerging Technologies for Connected and Smart Social Objects*, pp. 317–334, 2020. 14

[43] B. Eddy and H. Oussama, "Social relationship paradigm applied to object interactions in industrial iot," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1391–1396, 2018. 16

[44] F. E. White, *Data Fusion Lexicon*. Californie: Data Fusion Subpanel of the Joint Directos of Laboratories Technical Panel for C3, 1991. 16

[45] I. Bloch, *Fusion d'informations en traitement du signal et des images*. Hermes Science Publication, 2003. 16

[46] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Information Fusion*, vol. 57, pp. 115–129, 2020. 17

[47] F. Alam, R. Mehmood, I. Katib, N. N. Albogami, and A. Albeshri, "Data fusion and iot for smart ubiquitous environments: A survey," *Ieee Access*, vol. 5, pp. 9533–9554, 2017. 17, 19, 20

[48] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Information Fusion*, vol. 57, pp. 115–129, 2020. 21, 22

[49] https://www.datacamp.com/blog/what-is-machine-learning, (Consulted on 28 February 2024). 21, 22, 23

[50] J. Senthil Kumar, G. Sivasankar, and S. Selva Nidhyananthan, "An artificial intelligence approach for enhancing trust between social iot devices in a network," *Toward Social Internet of Things (SIoT): Enabling Technologies, Architectures and Applications: Emerging Technologies for Connected and Smart Social Objects*, pp. 183–196, 2020. 30, 44

[51] M. A. Al-qaness, A. A. Ewees, M. Abd Elaziz, A. Dahou, M. A. Al-Betar, A. O. Aseeri, D. Yousri, and R. A. Ibrahim, "Boosted barnacles algorithm optimizer: Comprehensive analysis for social iot applications," *IEEE Access*, 2023. 39, 46

[52] C. Marche, L. Atzori, V. Pilloni, and M. Nitti, "How to exploit the social internet of things: Query generation model and device profiles' dataset," *Computer Networks*, vol. 174, p. 107248, 2020. 49, 52

[53] https://domino.ai/data-science-dictionary/anaconda/, (Consulted on 20 June 2024). 55

[54] https://www.spyder-ide.org/, (Consulted on 20 June 2024). 56

[55] https://colab.google/, (Consulted on 20 June 2024). 56

[56] https://www.lucidchart.com/pages/fr, (Consulted on 20 June 2024). 56

[57] https://www.python.org/doc/essays/blurb/, (Consulted on 20 June 2024). 56

[58] https://numpy.org/devdocs/, (Consulted on 20 June 2024). 56

[59] https://pandas.pydata.org/docs/, (Consulted on 20 June 2024). 56

[60] https://domino.ai/data-science-dictionary/sklearn, (Consulted on 20 June 2024). 56

[61] https://matplotlib.org/, (Consulted on 20 June 2024). 56

[62] https://www.nemesis-studio.com/objets-connectes-un-futur-prometteur, (Consulted on 13 November 2023).

[63] J. Al-Hage, "Fusion de données tolérante aux défaillances : Application à la surveillance de l'intégrité d'un système de localisation," Thèse de doctorat en Automatique et Informatique Industrielle, Université de Lille, Octobre 2016.

[64] https://primabord.eduscol.education.fr/qu-est-ce-qu-un-objet-connecte, (Consulted on 23 November 2023).

[65] Z. Baloch, F. K. Shaikh, and M. A. Unar, "A context-aware data fusion approach for health-iot," *International Journal of Information Technology*, vol. 10, p. 241–245, 2018. [Online]. Available: https://doi.org/10.1007/s41870-018-0116-1

ABSTRACT

Managing data in the Social Internet of Things (SIoT) is challenging due to the large and varied data from connected devices. The social interactions between devices add extra complexity. This study focuses on both data fusion and the classification of relationships between devices using machine learning techniques.

We developed a multi-stage data fusion method and applied machine learning algorithms to classify the relationships accurately. We tested algorithms like KNN, Random Forest, Decision Tree, Naive Bayes, Logistic Regression, Gradient Boosting, and XGBoost.

The experimental results demonstrate that the Random Forest and XGBoost algorithms perform well compared to other machine learning algorithms in terms of accuracy, recall, precision, and F1 score.

Keywords: SIoT, Data Fusion, Machine Learning.

RÉSUMÉ

La gestion des données dans l'Internet social des objets (SIoT) est un défi en raison des données volumineuses et variées provenant des appareils connectés. Les interactions sociales entre appareils ajoutent une complexité supplémentaire. Cette étude se concentre à la fois sur la fusion de données et sur la classification des relations entre les appareils à l'aide de techniques d'apprentissage automatique.

Nous avons développé une méthode de fusion de données en plusieurs étapes et appliqué des algorithmes d'apprentissage automatique pour classer les relations avec précision. Nous avons testé des algorithmes tels que KNN, Random Forest, Decision Tree, Naive Bayes, Logistic Regression, Gradient Boosting et XGBoost.

Les résultats expérimentaux démontrent que les algorithmes Random Forest et XGBoost fonctionnent bien par rapport à d'autres algorithmes d'apprentissage automatique en termes d'exactitude, de rappel, de précision et de score F1.

Mots clés: SIoT, Fusion de Données, Apprentissage Automatique.