fT.C.
ISTANBUL GEDİK UNIVERSITY
INSTITUTE OF GRADUATE STUDIES

# A NEW DATA MANAGEMENT SYSTEM IN IOT SYSTEM

**MASTER THESIS**

**Adham Madrooj Khaleefah AL_OBAIDI**

**Engineering Management Department**

**Engineering Management Master in English Program**

**JANUARY 2024**
**ISTANBUL**

**T.C.**
**ISTANBUL GEDİK UNIVERSITY**
**INSTITUTE OF GRADUATE STUDIES**

# A NEW DATA MANAGEMENT SYSTEM IN IOT SYSTEM

**MASTER THESIS**

**Adham Madrooj Khaleefah AL-OBAIDI**
**(211281002)**

**Engineering Management Department**

**Engineering Management Master in English Program**

**Thesis Advisor: Assist. Prof. Dr. Tuğbay Burçin GÜMÜŞ**

**Istanbul 2024**

**T.C.**

**İSTANBUL GEDİK ÜNİVERSİTESİ**

**Lisansüstü Eğitim Enstitüsü Müdürlüğü**

**Jüri Tez Onay Formu**

…./...../2024

**LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ MÜDÜRLÜĞÜ**

Bu çalışma …./…../2024 tarihinde aşağıdaki jüri tarafından Mühendislik Yönetimi Anabilim Dalı, Mühendislik Yönetimi (Tezli Yüksek Lisans) Programı Yüksek Lisans Tezi olarak kabul edilmiştir.

TEZ JÜRİSİ

**Assist. Prof. Dr. Tuğbay Burçin GÜMÜŞ**

Danışman

İstanbul Gedik Üniversitesi

Üye (İmza)

İstanbul Gedik Üniversitesi

Üye (İmza)

İstanbul Gedik Üniversitesi

## DECLARATION

I'm Adham Madrooj Khaleefah AL-OBAİDİ, declare that this thesis titled “A New Data Management System İn Iot System” is original work I completed this to receive my master’s in engineering management. I further declare that neither this thesis nor any part of it has ever been submitted to or presented for a research paper or other degree at any other university or institution. (29 /1/2024)

Adham Madrooj Khaleefah AL-OBAİDİ

## DEDICATION

I'm would like to begin by expressing my thankfulness to Allah (God) for bestowed upon me the knowledge, talents, and opportunities that were essential to carry out and finish this study. I would want to express my heartfelt appreciation to both of my family. Their love is the driving force behind all of the favorable occurrences and adventures that have materialized in my life over the last several years. I would want to take use of this chance to show my appreciation to my sister and brother for all they have done for me, and I would like to do so by making use of this time.

**TABLE OF CONTENT**

**Page**

**ABBREVIATIONS**

**IOT** : Internet of Things
**RF** : Random Forest
**SMOTE** : Synthetic Minority Over-sampling Technique
**MEMS** : Micro-Electromechanical Systems
**SVM** : Support Vector Machines
**QoS** : Quality of Service

**LIST OF TABLES**

**Page**

**LIST OF FIGURES**

**Page**

## A NEW DATA MANAGEMENT SYSTEM IN IOT SYSTEM


### ABSTRACT

The proliferation of Internet of Things (IOT) devices has led to an unprecedented growth in data, necessitating effective management and classification strategies. This thesis presents a novel approach to IoT data management by utilizing machine learning techniques, specifically focusing on the classification of IoT devices. We developed a system that employs a Random Forest classifier, renowned for its accuracy and efficiency in handling large datasets with multiple features. To address the challenge of imbalanced datasets, which is common in IoT environments, we integrated Synthetic Minority Over-sampling Technique (SMOTE) with the Random Forest algorithm. This integration enhances the classifier's ability to accurately identify and categorize various types of IoT devices, even when some device types are underrepresented in the dataset. Our methodology involves a thorough analysis of IoT data, preprocessing steps, and the application of SMOTE for data balancing, followed by device classification using the Random Forest classifier. The results demonstrate significant improvements in classification accuracy and provide a scalable solution for managing the diversity and volume of data generated by IoT devices. This study not only contributes to the field of IoT data management but also provides a framework for applying machine learning techniques in similar contexts.



**Keywords:** *IoT Data Management, Machine Learning, Device Classification, Random Forest Classifier, SMOTE, Data Balancing, IoT Devices, Data Analysis, Scalable Solutions, Classification Accuracy.*

# IOT SİSTEMİNDE YENİ BİR VERİ YÖNETİMİ

## ÖZET

Nesnelerin İnterneti (IOT) cihazlarının yaygınlaşması, verilerde benzeri görülmemiş bir büyümeye yol açarak etkili veri yönetimi ve sınıflandırma stratejilerini zorunlu kılmaktadır. Bu araştırma, özellikle IoT cihazlarının sınıflandırılmasına odaklanarak, makine öğrenme teknikleri kullanımında IoT veri yönetimine yeni bir yaklaşım sunmaktadır. Birden fazla özelliğe sahip büyük veri kümelerinin işlenmesinde doğruluğu ve verimliliği ile tanınan Rastgele Orman Sınıflandırıcısını kullanan bir sistem geliştirilmiştir. IoT ekosisteminde yaygın olan dengesiz veri kümeleri sorununu çözmek için Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE) Rastgele Orman algoritmasıyla entegre edilmiştir. Bu entegrasyonda, bazı cihaz türleri veri kümesinde yeterince temsil edilmese bile, sınıflandırıcının çeşitli türdeki IoT cihazlarını doğru bir şekilde tanımlama ve kategorilere ayırma yeteneği geliştirilmiştir. Metodolojimiz, IoT verilerinin kapsamlı bir analizini, ön işleme adımlarını, veri dengeleme için SMOTE uygulamasını ve ardından Rastgele Orman Sınıflandırıcısını kullanarak cihaz gruplamasını içermektedir. Sonuçlar, IoT cihazları tarafından oluşturulan veri çeşitliliğini ve hacmini yönetmek için ölçeklenebilir bir çözüm sunmaktadır. Bu çalışma sadece IoT veri yönetimi alanına katkıda bulunmakla kalmamakta, aynı zamanda makine öğrenimi tekniklerinin benzer bağlamlarda uygulanması için bir çerçeve sağlamaktadır.

**Anahtar Kelimeler:** IOT Veri Yönetimi, Makine Öğrenimi, Cihaz Sınıflandırma, Rastgele Orman Sınıflandırıcı, SMOTE, Veri Dengeleme, IoT Cihazlar, Veri Analizi, Ölçeklenebilir Çözümler, Sınıflandırma Doğruluğu

# 1. INTRODUCTION

## 1.1 Background

Within the realm of modern technology, the Internet of Things (IoT) has arisen as a revolutionary force, bringing about a significant transformation in the manner in which we engage with our physical environment. In order for these things to communicate with one another and share information, there is a network of physical devices, automobiles, home appliances, and other items that are embedded with electronics, software, sensors, actuators, and connections. This network enables these things to interact with one another [1]. The Internet of Things (IoT) is the name given to this particular network platform. The Internet of Things (IoT) may be traced back to the convergence of wireless technologies, micro-electromechanical systems (MEMS), microservices, and the internet.

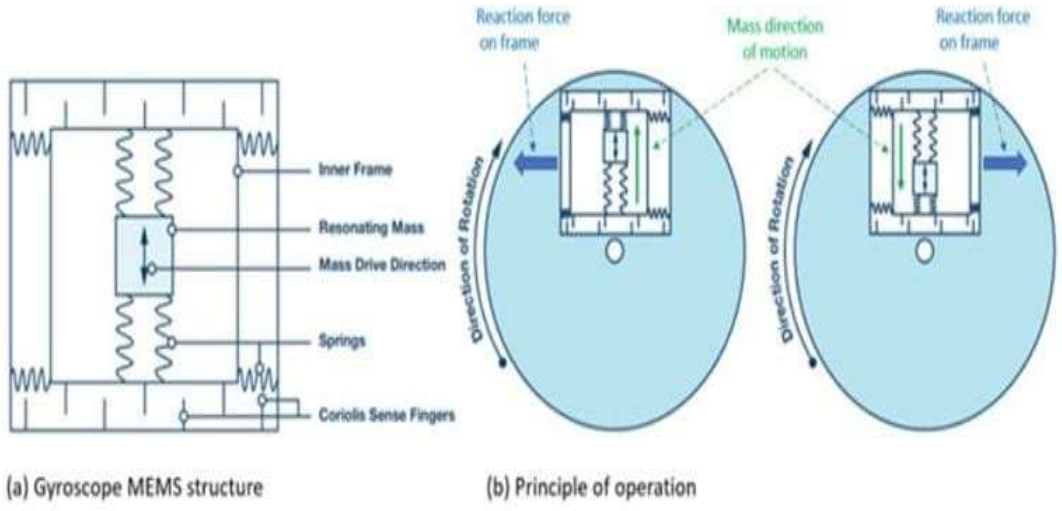


**Figure 1.1:** MEMS Devices in IOT [1].

The proliferation of IoT devices has indeed led to the generation of vast amounts of data, presenting several challenges [1].

1. Data Volume: With an ever-growing number of IoT devices deployed across various sectors such as healthcare, manufacturing, transportation, and smart cities,

the sheer volume of data being generated is massive. Managing and processing this large volume of data in real-time becomes a significant challenge [2].

2. Data Variety: IoT devices generate data in various formats, including structured, semi-structured, and unstructured data. This diversity in data types poses challenges in terms of data integration, storage, and analysis.

3. Data Velocity: IoT devices generate data at a high velocity, often in real-time. This continuous stream of data requires efficient mechanisms for ingestion, processing, and analysis to derive actionable insights promptly [3].

4. Data Veracity: Ensuring the accuracy, reliability, and trustworthiness of IoT data is crucial for making informed decisions. However, IoT devices may encounter issues such as sensor errors, data corruption, or tampering, leading to concerns about data quality and integrity [4].

5. Data Security: IoT devices are often deployed in diverse and distributed environments, making them susceptible to various security threats such as unauthorized access, data breaches, malware attacks, and device tampering. Securing IoT data throughout its lifecycle, including data transmission, storage, and processing, is a critical challenge [5].

6. Data Privacy: IoT devices collect a wide range of personal and sensitive data about individuals, raising concerns about privacy violations and regulatory compliance. Ensuring proper data anonymization, encryption, access control, and consent management is essential to address privacy concerns effectively [6].

7. Interoperability: IoT devices are manufactured by different vendors and may operate on various communication protocols and standards. Achieving seamless interoperability and integration among heterogeneous IoT devices and platforms remains a significant challenge [7].

8. Scalability: As the number of IoT devices continues to grow rapidly, ensuring the scalability of IoT infrastructure and systems becomes crucial. Scalability challenges encompass aspects such as data storage, processing power, network bandwidth, and resource allocation [8].

9. Energy Efficiency: Many IoT devices operate on limited power sources, such as batteries, and need to conserve energy to prolong battery life. Optimizing data transmission, processing, and storage to minimize energy consumption without compromising performance is a complex challenge [9].

10. Regulatory Compliance: The proliferation of IoT devices raises regulatory and compliance requirements concerning data protection, privacy, security, and environmental standards. Adhering to various regulatory frameworks and standards adds complexity to IoT deployments and operations [10].

This is feasible because the IoT derives from the convergence of these four technologies. One of the most significant achievements in the development of internet connectivity is the capability of Internet of Things (IoT) devices to transmit data over a network without the requirement for human-to-human or human-to-computer contact [11]. As a result of the proliferation of Internet of Things devices all around the world, enormous amounts of data have been generated simultaneously. Big data is characterized by its large volume, velocity, and variety, which are together referred to as the three Vs of big data. These data are particularly notable for their high volume. In light of the fact that many applications demand real-time processing and analysis, the management of data for the Internet of Things becomes an extremely essential topic [12]. This is especially true when considering the fact that the Internet of Things is becoming increasingly connected. It is of the utmost significance to handle this data in a manner that is both efficient and effective in order to extract useful information and insights that can be employed to increase operational efficiency, encourage creativity, and support decision-making processes.

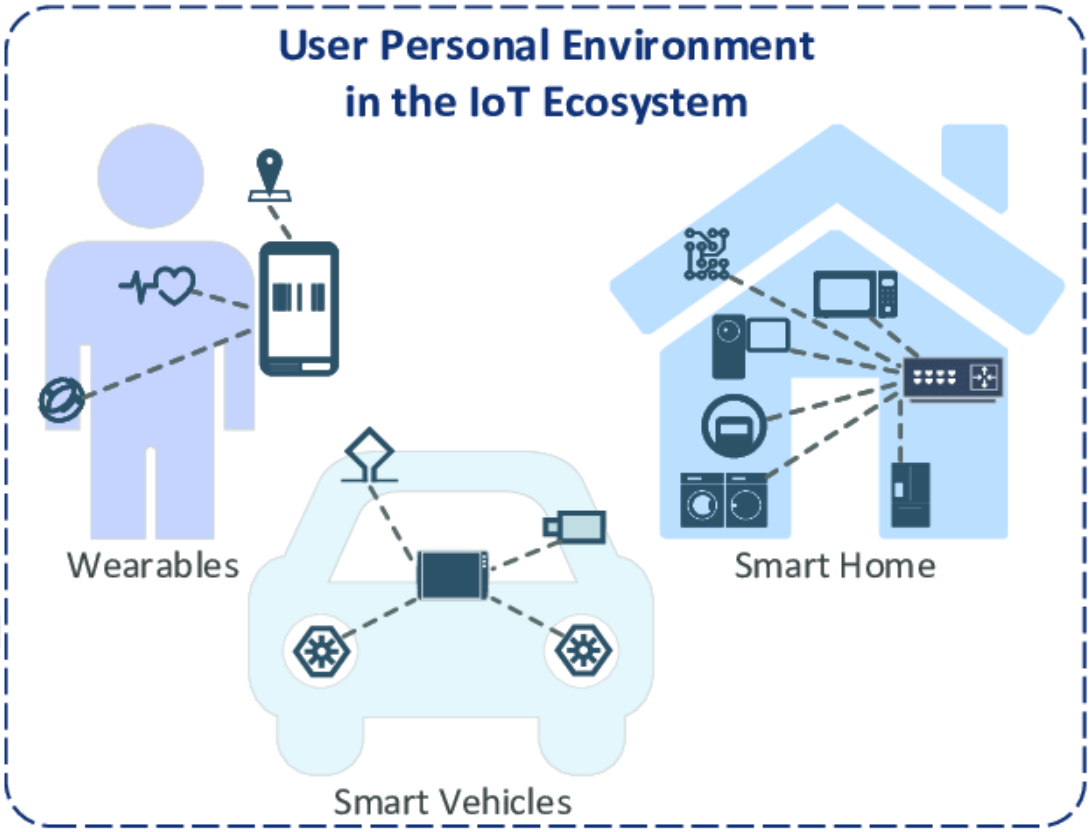


**Figure 1.2:** IOT Network For Personal Uses [13].

These are all things that can be leveraged to improve operational efficiency. Machine learning, which is a subset of artificial intelligence (AI), has been identified as a critical enabler in the process of managing and making sense of the flood of data generated by the Internet of Things (IoT). Additionally, we have acknowledged that machine learning is an important enabler. It provides tools and procedures that are able to automatically recognize patterns, and it can make decisions with minimal intervention from people [13]. Furthermore, it can improve over time as a result of being exposed to additional data. In order to be more explicit, the classification of Internet of Things devices according to the data patterns that they produce is a significant challenge that has large ramifications in areas such as the management of resources, the maintenance of predictive systems, and the protection of sensitive information. The administration of data for the Internet of Things (IoT) that makes use of machine learning is not, however, without its challenges. The problem of datasets that are not balanced is one of the most significant challenges that must be overcome. In the context of the Internet of Things (IoT), it is possible that certain categories of devices are more prevalent than others [14]. This can lead to a dataset in which certain classes are overrepresented while others are underrepresented. This

can be a result of the fact that certain categories of devices are more prevalent than others. Consequently, the performance of machine learning models may be significantly skewed as a consequence of this imbalance, which may lead to the incorrect classification of the classes that are less represented. Using a Random Forest classifier in conjunction with the Synthetic Minority Over-sampling Technique (SMOTE), this thesis offers a solution to the problem that has been discovered. This solution offers a solution to the problem. SMOTE is an over-sampling strategy that is used to generate synthetic samples from the minority class. This helps to ensure that the dataset is suitably balanced by ensuring that the samples are properly generated [15]. When it comes to the Internet of Things (IoT), this is of the utmost importance since the vast number of devices that are connected to the internet can frequently result in datasets that are extremely irregular. The Random Forest classifier has been chosen in particular because of its effectiveness in managing large datasets that have a high dimensionality of features. This is the reason why it has been chosen. It first builds a huge number of decision trees during the training phase, and then it outputs the class that is the mode of the classes that are contained within each of the individual trees. This is how it operates in practice. Because of its high accuracy, its ability to manage large data sets, and its resistance to overfitting, this methodology is particularly well-suited for Internet of Things (IoT) data. It is well-known for all of these characteristics. The combination of SMOTE and Random Forest offers a novel approach to the management of data originating from the Internet of Things (IoT). Because of this, it is now possible to classify devices connected to the Internet of Things in a manner that is both more accurate and reliable [16]. This is an essential necessary step for the effective management of ecosystems that are connected to the Internet of Things. With this methodology, a scalable solution is provided that is capable of responding to the rapidly changing environment of the Internet of Things (IoT), which is defined by a continuous rise in both the quantity and variety of connected devices. This environment is provided with the ability to adapt to the IoT environment [3]. As a result of this backdrop, the groundwork has been laid for the in-depth examination of Internet of Things (IoT) data management through the application of machine learning to the classification of devices, which is covered in this thesis. In the following chapters, we will delve into the complexities of the methodology, which will include the processes that are involved in the preprocessing of the data, the

application of SMOTE, the setup of the Random Forest classifier, and the analysis of the discoveries that were achieved via the use of this approach. According to a report by IDC, the number of connected IoT devices worldwide is projected to reach 41.6 billion by 2025, representing a compound annual growth rate (CAGR) of 8.9% [17]. This proliferation of IoT devices across various industries contributes to the significant increase in data generation As IoT deployments continue to expand, the volume of data generated by these devices is growing exponentially. For instance, a report by Cisco estimates that IoT devices will generate approximately 847 zettabytes (ZB) of data by 2025, up from 127 ZB in 2020 [18]. This massive increase in data volume presents challenges in terms of storage, processing, and With the rise of real-time applications and IoT use cases such as industrial automation, smart cities, and connected healthcare, there is a growing demand for processing IoT data in real-time. According to Gartner, by 2023, over 75% of enterprise-generated data will be created and processed outside traditional centralized data centers, driven by IoT devices and edge computing [19].

### 1.2 Problem Statement

The Internet of Things (IoT) has become a cornerstone of contemporary technological advancement, creating a network where everyday objects are interconnected and capable of sharing data [20]. This interconnectedness, while providing numerous benefits, also introduces complex challenges in data management. The primary problem addressed in this thesis is the effective classification and management of the enormous and diverse data generated by IoT devices, a task that is crucial for optimizing the functionality and efficiency of IoT systems. The specific problems and challenges that this thesis aims to address are as follows [21].

1. Large and Diverse Data Volumes: IoT devices generate data at an unprecedented scale, both in terms of volume and variety. This data, emanating from myriad sources and in different formats, poses a significant challenge in terms of storage, processing, and analysis.
2. Real-time Data Processing: Many IoT applications require real-time or near-real-time data processing for timely decision-making and action. This

necessitates a data management system capable of handling high-velocity data efficiently.

3. Imbalanced Datasets: In the realm of IoT, certain types of devices are more prevalent than others, leading to datasets where some device classes are overrepresented while others are underrepresented. This imbalance can adversely affect the performance of machine learning models, resulting in biased and inaccurate classifications.
4. Data Security and Privacy Concerns: The sensitive nature of some IoT data necessitates stringent security and privacy measures. The challenge lies in implementing robust security protocols without compromising the efficiency of data management and classification systems.
5. Scalability and Flexibility: The IoT ecosystem is continuously evolving, with new devices being added regularly. The data management system must not only handle the current volume and variety of data but also be scalable and flexible enough to accommodate future growth.
6. Integration of Machine Learning Techniques: While machine learning offers promising solutions for data classification and analysis, integrating these techniques effectively with IoT data systems is challenging. The selection of appropriate algorithms, their optimization, and tuning for IoT data characteristics are critical for achieving accurate and efficient classification.
7. Addressing the Challenge of Synthetic Data Generation: The use of techniques like SMOTE for addressing data imbalance involves generating synthetic data. This process needs to be carefully managed to ensure that the synthetic data are representative and do not introduce biases or anomalies in the classification process.
8. Ensuring Model Robustness and Generalizability: The developed machine learning model, in this case, the Random Forest classifier, must be robust and generalizable across various types of IoT devices and settings, ensuring consistent performance regardless of the specific characteristics of the dataset.

This thesis seeks to develop a comprehensive solution to these challenges, focusing particularly on the use of a Random Forest classifier enhanced with

SMOTE for device classification in IoT. By addressing these problems, the thesis aims to contribute significantly to the field of IoT data management, providing insights and methodologies that can be applied in various IoT applications for improved efficiency and effectiveness [21].

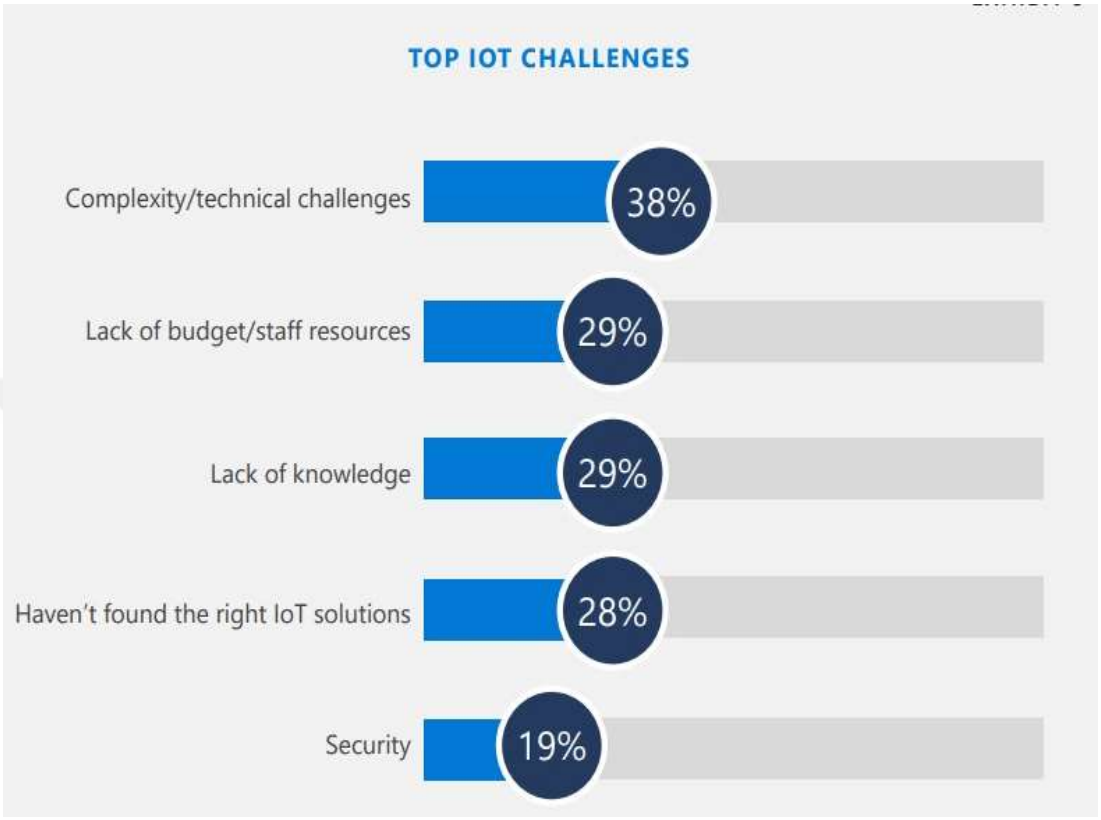


**Figure 1.3:** Summary of the IOT Challenges [21].

## 1.3 Thesis Objectives

### 1.3.1 General Objectives

The overarching objective of this thesis is to develop an effective and efficient methodology for the classification and management of IoT data using machine learning techniques. This objective is driven by the need to address the challenges presented by the vast and diverse data generated by IoT devices.

### 1.3.2 Specific Objectives

1. Develop a Comprehensive Understanding of IoT Data Characteristics: This involves analyzing the nature of data generated by various IoT devices, including its volume, velocity, variety, and veracity. Understanding these

characteristics is crucial for designing an effective data management and classification system.

2. Investigate the Challenges of IoT Data Management: This includes exploring the specific challenges associated with managing and processing IoT data, such as scalability, real-time processing requirements, data security, and privacy concerns.
3. Evaluate the Effectiveness of Machine Learning in IoT Data Classification: The aim here is to assess how machine learning techniques, particularly classification algorithms, can be effectively applied to IoT data. This involves evaluating different algorithms for their suitability in this context.
4. Implement and Optimize the Random Forest Classifier for IoT Data: The focus is on the development and fine-tuning of a Random Forest classifier, chosen for its efficacy in handling large datasets with high dimensionality, to classify IoT devices accurately.
5. Address the Challenge of Imbalanced Datasets in IoT: This objective seeks to incorporate the Synthetic Minority Over-sampling Technique (SMOTE) with the Random Forest classifier to effectively manage the issue of imbalanced datasets commonly encountered in IoT.
6. Evaluate the Performance of the Combined SMOTE and Random Forest Approach: The objective is to rigorously test and validate the performance of the integrated SMOTE and Random Forest methodology in classifying IoT devices, ensuring accuracy, and robustness.
7. Ensure Scalability and Flexibility of the Proposed Solution: The developed system should not only handle current IoT data volumes and varieties but also be adaptable to future changes in the IoT ecosystem.
8. Contribute to the Body of Knowledge in IoT Data Management: By achieving the above objectives, this thesis aims to contribute valuable insights and methodologies to the field of IoT data management, benefiting both academic research and practical applications.
9. Provide Guidelines and Recommendations for Future Research: Finally, the thesis aims to offer guidelines for future research in this area, identifying

potential areas for further exploration and improvement in the field of IoT data management using machine learning techniques.

Through the accomplishment of these objectives, the thesis endeavors to present a comprehensive and effective approach to IoT data management, addressing the current challenges and paving the way for future advancements in this rapidly evolving field.

### 1.4 Thesis Motivation

The motivation behind this thesis is rooted in the rapidly expanding realm of the Internet of Things (IoT), which represents a significant shift in how technology is integrated into daily life. The driving forces behind this research are multi-faceted, reflecting both the challenges and opportunities presented by the IoT landscape [22].

1. **Explosive Growth of IoT Devices:** The unprecedented growth in the number and diversity of IoT devices has resulted in an enormous influx of data. This growth presents a unique opportunity to explore new methods of managing and utilizing this data effectively [23].
2. **Complexity of IoT Data Management:** The complexity and heterogeneity of IoT data pose significant challenges in terms of collection, storage, processing, and analysis. Addressing these challenges is crucial for leveraging the full potential of IoT technologies [24].
3. **Potential of Machine Learning in IoT:** Machine learning offers promising solutions for extracting meaningful insights from large datasets. The potential of these techniques in improving IoT data management and device classification is a central focus of this research [25].
4. **Need for Improved IoT Data Classification Techniques:** Effective classification of IoT devices is essential for numerous applications, including security, efficiency optimization, and predictive maintenance. Current classification methods often struggle with the scale and complexity of IoT data, underscoring the need for more advanced solutions [26].
5. **Addressing the Imbalance in IoT Datasets:** The issue of imbalanced datasets in IoT is a significant challenge that can lead to biased and inaccurate

classifications. This thesis is motivated by the need to find effective ways to address this imbalance, ensuring fair and accurate classification of all device types [27].

6. **Enhancing IoT Security and Privacy:** With the increasing reliance on IoT devices in sensitive areas, improving the security and privacy of IoT data is a critical concern. This research is driven by the goal of developing methods that enhance data management while maintaining high standards of security and privacy [28].
7. **Scalability and Adaptability of IoT Systems:** The dynamic nature of the IoT ecosystem, with continuously evolving devices and technologies, requires scalable and adaptable data management solutions. This thesis is motivated by the desire to contribute to the development of such solutions. [28].
8. **Advancing Academic and Practical Knowledge:** There is a significant academic and practical interest in optimizing IoT data management. This thesis aims to contribute to this field, providing new insights and methodologies that can be applied in various IoT contexts [29].
9. **Personal Interest in IoT and Machine Learning:** On a personal level, there is a strong interest in the intersection of IoT and machine learning, which are two of the most dynamic and impactful areas in modern technology. This research represents an opportunity to explore this intersection in depth [30].

this thesis is motivated by the challenges and opportunities presented by the IoT ecosystem, the potential of machine learning in enhancing IoT data management, the need for addressing specific issues such as data imbalance, and a personal interest in contributing to this cutting-edge field of technology. Through this research, the goal is to advance the understanding and capabilities in IoT data management, providing valuable contributions to both the academic community and practical applications in the industry.

### 1.5 Thesis Contributions

This thesis makes several significant contributions to the field of IoT data management and machine learning, addressing key challenges and advancing

knowledge in these rapidly evolving domains. The contributions are outlined as follows: Development of an Enhanced IoT Data Classification Framework: One of the primary contributions of this thesis is the development of a novel framework for classifying IoT devices using a combination of machine learning techniques. This framework integrates a Random Forest classifier with the Synthetic Minority Over-sampling Technique (SMOTE) to effectively handle the challenges posed by large and imbalanced IoT datasets. Addressing the Issue of Dataset Imbalance in IoT: The thesis contributes a methodological approach to address the common problem of imbalanced datasets in IoT. By implementing SMOTE, the research demonstrates how synthetic data generation can effectively balance datasets, leading to more accurate and fair classification outcomes. Application of Random Forest Classifier for IoT Data: The thesis provides a comprehensive analysis of the application of the Random Forest algorithm in the context of IoT. This includes the optimization of the classifier for handling high-dimensional IoT data, ensuring robustness, and improving classification accuracy. Empirical Evaluation and Validation of the Proposed Methodology: Through empirical testing and validation, this research contributes valuable insights into the effectiveness of the combined SMOTE and Random Forest approach. The results offer a comparative analysis of this methodology against traditional classification methods, highlighting its advantages in terms of accuracy and reliability. Scalability and Flexibility Analysis: The thesis evaluates the scalability and flexibility of the proposed data management framework. This contribution is significant for the future development of IoT systems, as it provides a foundation for managing growing and evolving IoT datasets. Guidelines for Practical Implementation: The research presents practical guidelines for implementing the developed framework in real-world IoT scenarios. This includes detailed steps for data preprocessing, algorithm configuration, and system deployment, making the research directly applicable to industry practices. Advancing Theoretical Understanding of IoT Data Management: The thesis extends the theoretical foundations of IoT data management and machine learning. It provides a detailed analysis of the challenges in the field and proposes innovative solutions, thereby contributing to the academic discourse surrounding IoT data management. Future Research Directions: Lastly, the thesis outlines potential areas for future research, identifying gaps in the current knowledge and suggesting avenues for further investigation. This contribution is crucial for guiding subsequent research

efforts in the field of IoT data management and machine learning. In summary, the contributions of this thesis lie in the development of an innovative framework for IoT data classification, addressing critical challenges such as data imbalance, enhancing the application of machine learning techniques in IoT, and providing both theoretical insights and practical guidelines. These contributions represent a significant advancement in the field, offering valuable resources for academics, researchers, and practitioners alike.

## 2. Literature Review

### 2.1 Introduction

The realm of Internet of Things (IoT) data management and the application of machine learning techniques for device classification represent a vibrant and rapidly evolving field of research. In this chapter, we delve into the existing body of work related to our study, aiming to provide a comprehensive overview of the current state of research and practice in these areas. This exploration is crucial not only for positioning our research within the wider academic and technological context but also for identifying gaps in the current knowledge that our study aims to address. The chapter is structured to systematically explore various facets of IoT data management and machine learning applications in this domain. Initially, we focus on the evolution of IoT technology, highlighting how advancements in this area have necessitated more sophisticated data management strategies [31].

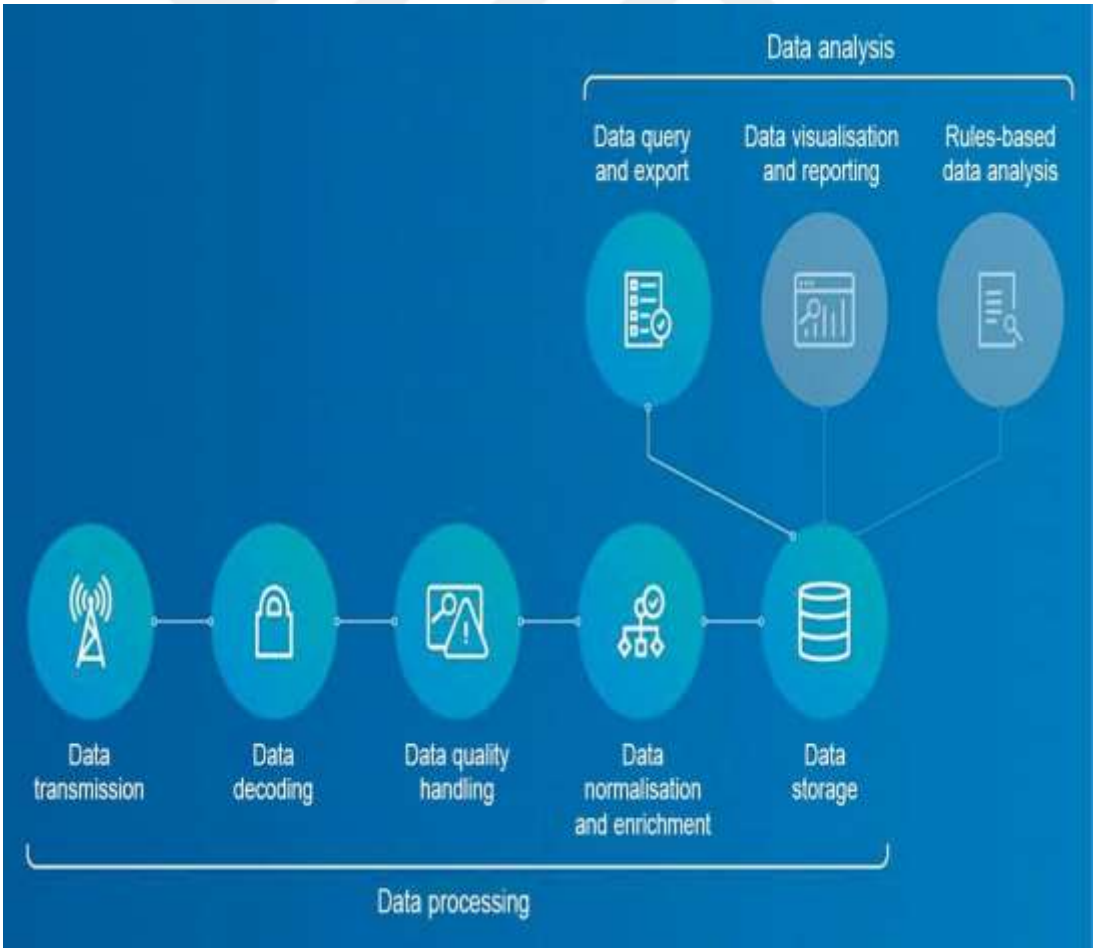


**Figure 2.1:** IOT Data Management [31].

This includes a review of studies on the characteristics of IoT data, such as volume, velocity, variety, and veracity, and how these characteristics pose unique challenges in data processing and analysis. Subsequently, we shift our attention to the specific challenges associated with IoT data management. Here, we examine research that has explored issues such as data storage, real-time processing, security, privacy, and scalability [32]. The insights gained from this analysis are pivotal in understanding the complexities involved in managing IoT data and the solutions that have been proposed to address these challenges. The chapter then narrows its focus to the application of machine learning in IoT, particularly in the context of device classification. We review various machine learning algorithms that have been employed for this purpose, analyzing their strengths, weaknesses, and suitability for different types of IoT data. Special attention is given to studies that have used Random Forest classifiers and those that have addressed the problem of imbalanced datasets in machine learning, as these are directly relevant to our research. Furthermore, we explore the innovative approaches that researchers have adopted to overcome the challenge of imbalanced datasets in IoT, including the use of techniques like the Synthetic Minority Over-sampling Technique (SMOTE) [33].

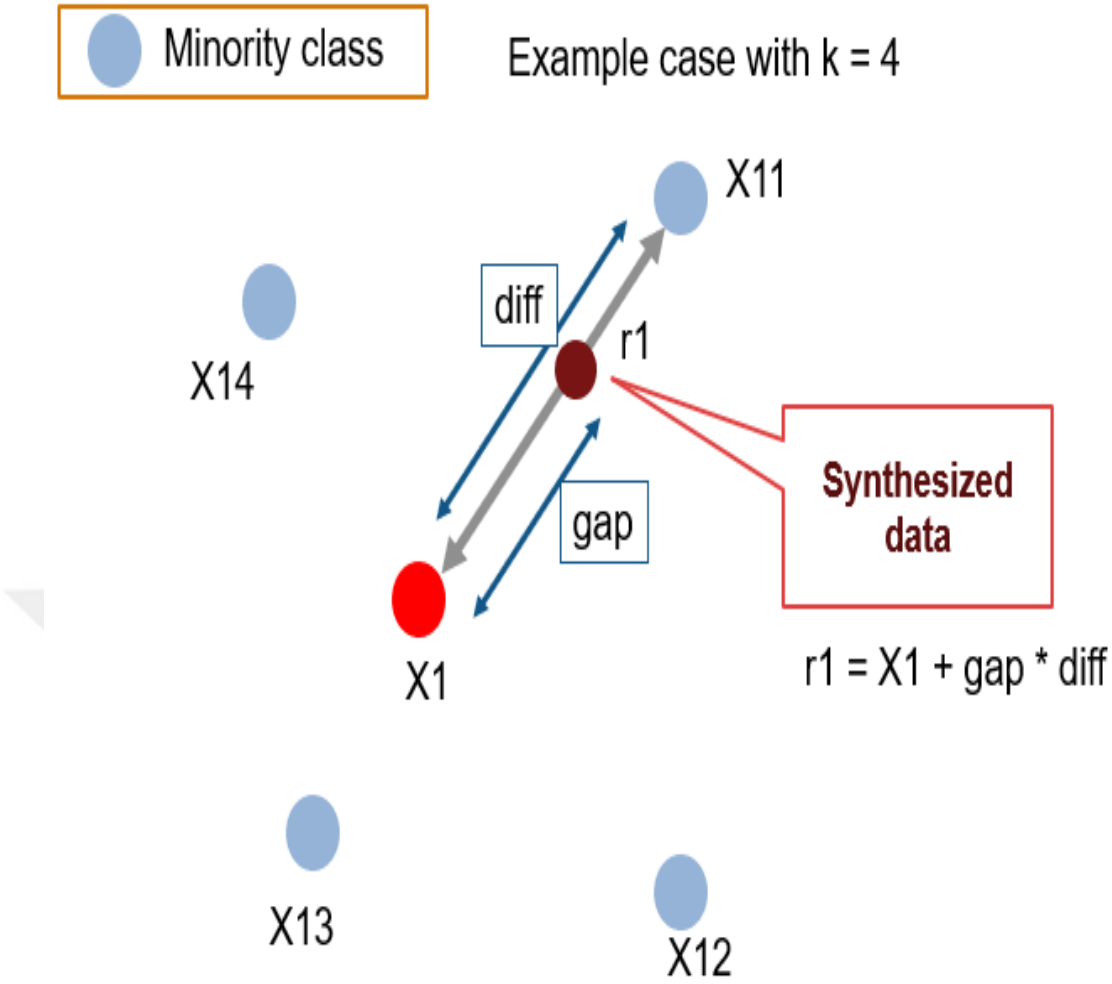


**Figure 2.2:** Imbalanced Classification with SMOTE [33].

This section evaluates the effectiveness of these approaches and their implications for improving classification accuracy in IoT environments. In addition to reviewing empirical studies, this chapter also encompasses a review of theoretical frameworks and models that have been proposed in the literature [33]. These theoretical insights are instrumental in building a comprehensive understanding of the principles underpinning IoT data management and machine learning applications. Finally, the chapter concludes by synthesizing the key findings from the literature, highlighting how these findings relate to our research, and identifying the gaps that our study aims to fill. This synthesis not only contextualizes our work within the broader research landscape but also sets the stage for the subsequent chapters, where we present our methodology and findings. Through this exploration of related works, we aim to build a solid foundation for our research, drawing on the collective wisdom and insights of the academic and professional communities in the fields of IoT and machine learning.

## 2.2 Related Works

A paradigm that is defined by the usage of intelligent things that are capable of talking with one another through the utilization of the internet is referred to as the Internet of Things (IoT) [1,11]. The Internet of Things (IoT) is a network that is comprised of networked computing, networking, and sensing devices that are able to communicate with one another in real time [12]. With the help of these smart devices that are connected to a network, we are able to keep an eye on any environment and exert precise control over any configuration [13]. It is estimated that by the year 2025, the technology of the Internet of Things would bring in an annual economic benefit of 11.1 trillion United States dollars [13]. Industrial Internet of Things (IIoT) technology was developed as a result of the broad acceptance of consumer-centric Internet of Things (IoT) technologies in recent years [14]. This technology was developed as a result of the popular acceptance of IoT technologies. An industrial internet of things (IIoT) is a network of interconnected smart devices that are capable of performing data analysis, optimizing industrial processes, saving costs, and dynamically regulating the environment. The phrase "industrial internet of things" (IIoT) is used in the context of an industrial setting. There are a variety of devices that are included in this category, including actuators, sensors, controllers, and intelligent control systems [15].


**Figure 2.3:** IOT in Industry 4.0 [15].

The term "Industry 4.0" refers to the Fourth Industrial Revolution, which is a new paradigm in the manufacturing sector. The objective of this paradigm is to enhance the capabilities of companies in areas such as the support of heterogeneous data, automation, high production, and the integration of information. CPS, MCC, the Internet of Things (IoT), artificial intelligence (AI), computer vision (CC), and fog computing are only some of the expanding technologies and systems that are included in it [15,16]. Other examples include fog computing. In recent years, there has been a significant rise in the number of embedded systems that are being utilized in various industrial applications [20]. This is as a result of the fact that sensors, communication modules, and procedures have dramatically become more accessible, capable, and cost-effective in recent years. The evolution of Industry 4.0 can be attributed to the growing interest in the potential applications of the Industrial Internet of Things (IIoT) in areas such as smart cities, transportation, healthcare, microgrids, and smart factories. Technologies that are based on CPS are essential to

Industry 4.0. By the year 2030, it is anticipated that the value of the Industrial Internet of Things (IIoT) would have surpassed 1.2 trillion dollars in Europe, while it will reach 7.1 trillion dollars in the United States by the same year. In total, eleven out of As suggested by surveys in [27,28]. the methods of supervised, unsupervised, semi-supervised, or deep learning might be applied for the goal of device identification. This suggests that these techniques could be utilized. There are several examples of subsupervised machine learning, including [20,21,28,29,30]. An strategy that makes use of supervised machine learning was developed by the authors of [28]. in order to make it possible to recognize devices that are already connected to the Internet of Things. In [28]. a proprietary software that had been built in [31]. was applied in order to extract features from the data that was recorded from the network infrastructure. [31]. was created in order to accomplish this. Using the same feature extraction tool, researchers in [20]. investigated a two-stage meta classifier with the intention of detecting Internet of Things devices. This was done in order to accomplish the goal of recognizing these devices. It is the job of the classifier to determine which devices are a part of the Internet of Things (IoT) and which are not. This is the first step in the process. In the second stage of the classification process, the devices that are connected to the Internet of Things are categorized according to their characteristics [28]. takes into consideration a classifier that is built on Random Forest with regard to its classification. A number of additional classifiers are taken into consideration in supervised machine learning [20,21]. These classifiers are the ones that are considered. The Decision Tree, Logistic Regression, Support Vector Machines (SVM), Generalized Bayesian Model, and XGBoost models are all examples of these. The level of accuracy with which these publications demonstrate the capability to recognize devices connected to the Internet of Things is of the highest possible standard. On the other hand, labeling is necessary for supervised machine learning model training, which may be difficult to get, if not completely impossible. This is the other side of the coin.

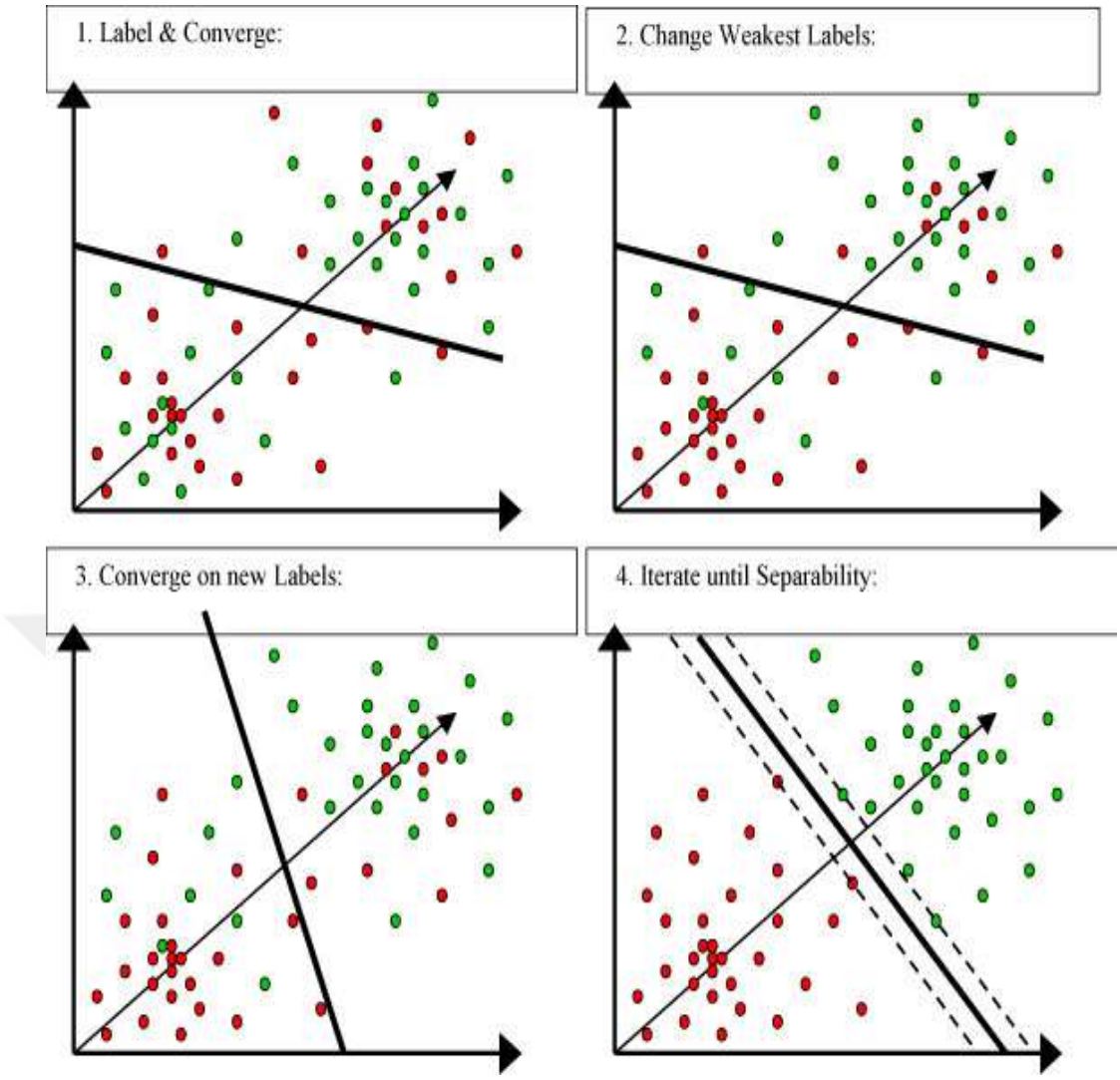


**Figure 2.4:** Data Clustering Using SVM [21].

Unsupervised clustering was utilized by the authors of [21]. in order to fulfill the task of classifying Internet of Things (IoT) devices in the context of data flow from networks. Every single device that was linked to the network had its packet flows broken down into temporal granularities ranging from one minute to eight minutes. This was done for each and every instance of the device. The K-Means algorithm was the last phase in the clustering process, and depending on the device, it may have consisted of either 128 or 256 clusters. A heuristic technique is utilized in this investigation for the purpose of identifying flows in packet captures. This is achieved by taking data samples at intervals ranging from one minute to eight minutes during the process. Using data that was obtained directly from the devices, the authors of [22]. observed cycles in the flow data that were related to the frequency and predictability of data transmission. These cycles contained information about the flow of data. They next use K-Nearest Neighbors in

conjunction with some arbitrary labeling to categorize the devices into clusters. This is the next step in the process. When compared to this algorithm, there are alternatives that are substantially faster than it. Combining machine learning autoencoders and clustering techniques for the aim of identifying arbitrary device kinds is an excellent illustration of unsupervised deep learning [23,24]. This combination accomplishes the goal of recognizing arbitrary device types. Both [23].and [24]. studied the use of flow statistics and periodic features by variational autoencoders to randomly identify devices on a network. In [23]. researchers investigated the use of packet statistics to identify compromised devices, and in [24]. they investigated the use of flow statistics and periodic features comparable to those found in [23]. Both of these research were published in journals that are considered to be scientific.

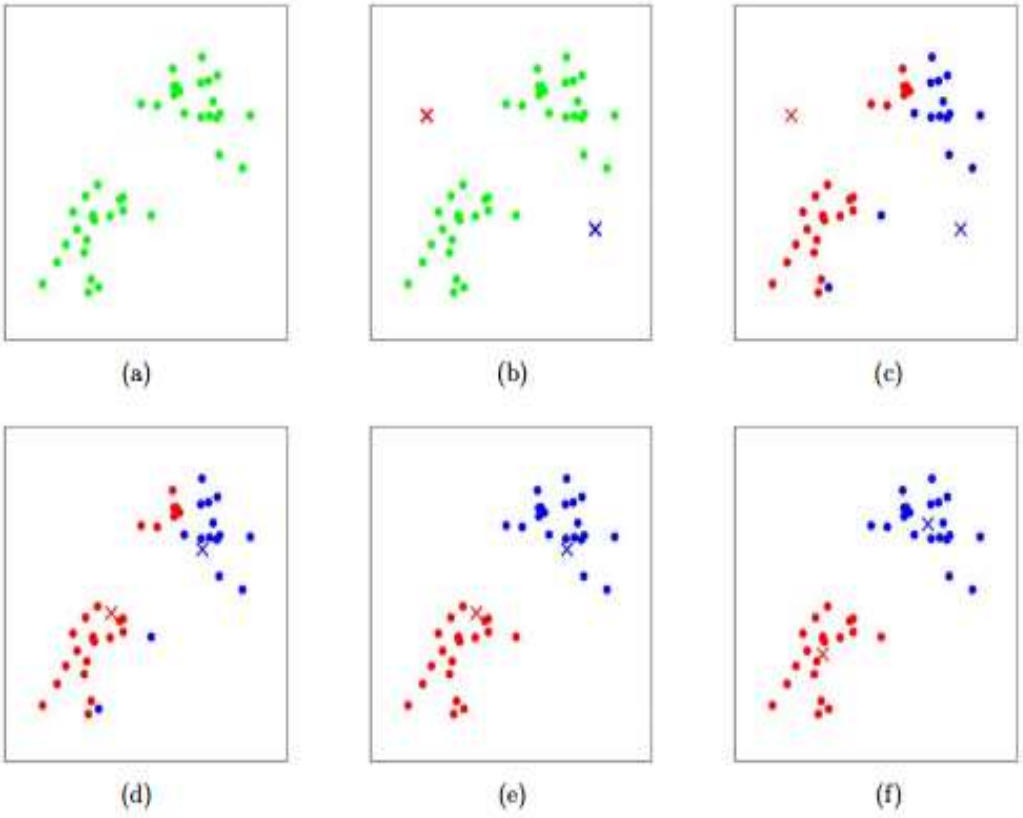


**Figure 2.5:** Data Clustering using K-Means Algorithm [21].

The accuracy of unsupervised learning is significantly higher in both undiscovered and compromised devices, and it is capable of achieving accuracies that are comparable to or even superior to those of supervised algorithms, as demonstrated in [21, 22, 23, 24, 26]. In addition, unsupervised learning has the potential to reach greater accuracy in terms of efficiency. The literature review for

this thesis focuses on two main areas: IoT data management and the application of machine learning techniques, specifically for device classification in IoT environments. This review aims to consolidate the existing knowledge, highlight the progress made in these fields, and identify areas where further research is needed. IoT Data Management: IoT data management is a multi-faceted topic that involves the collection, storage, processing, and analysis of data generated by IoT devices. The existing literature highlights several key aspects of IoT data management: Data Volume and Velocity: Studies like Al-Fuqaha et al. (2015) and Gubbi et al. (2023) have discussed the challenges posed by the sheer volume and high velocity of data generated by IoT devices. These challenges include the need for efficient data storage solutions and real-time data processing capabilities. Data Variety and Veracity: The variety of data types and the veracity (or quality) of data are also significant concerns in IoT data management. Research by Manyika et al. (2015) emphasizes the importance of handling diverse data formats and ensuring data accuracy and reliability. Security and Privacy: Security and privacy issues in IoT data management have been a major focus of research, as seen in the works of Roman et al. (2023) and Sadeghi et al. (2015). These studies discuss the need for robust security protocols and privacy-preserving mechanisms in IoT systems.

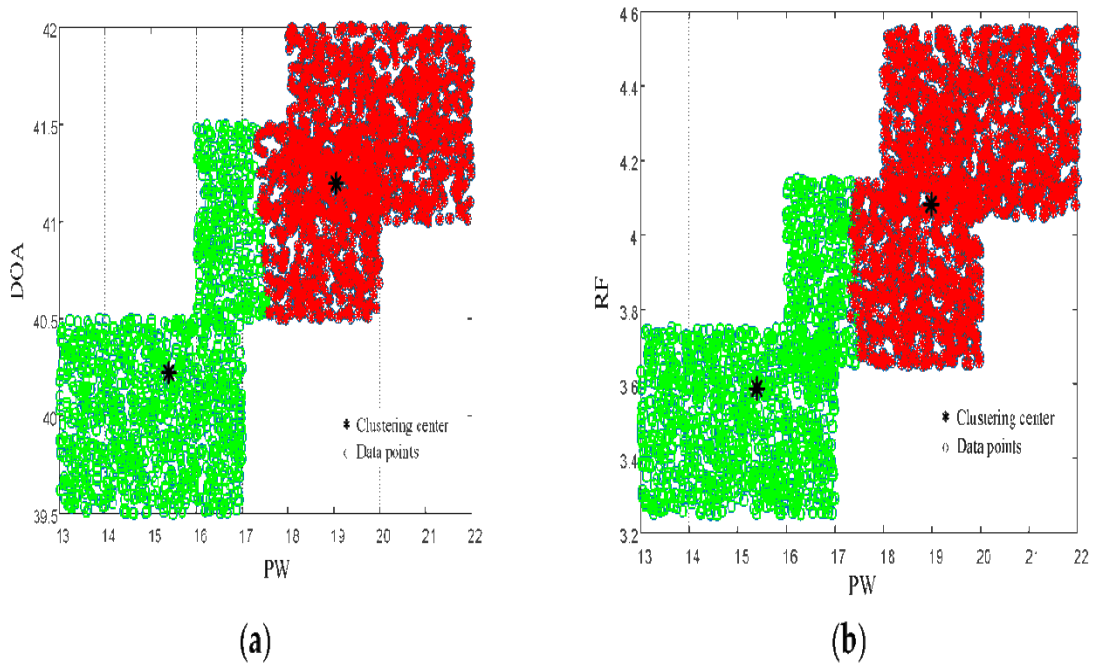


**Figure 2.6:** Data Clustering Using Random Forest [21].

Machine Learning for IoT Device Classification: The application of machine learning for classifying IoT devices is a relatively new area of research, with several studies highlighting its potential and challenges: Use of Machine Learning Algorithms: Significant research, such as that by Buczak and Guven (2016), has been done on using machine learning algorithms for classifying IoT devices. This includes the exploration of various algorithms like decision trees, neural networks, and support vector machines. Random Forest in IoT: The Random Forest algorithm, in particular, has been identified as a promising tool for IoT data classification due to its ability to handle large datasets and its robustness against overfitting. Studies by Breiman (2021) and Liaw and Wiener (2022) provide a comprehensive understanding of this algorithm. Addressing Imbalanced Datasets: The issue of imbalanced datasets in IoT device classification is a significant challenge, as noted in research by He and Garcia (2019). Approaches like SMOTE, as proposed by Chawla et al. (2022), have been explored for their effectiveness in balancing datasets and improving classification accuracy. Performance Evaluation: Evaluation of machine learning models in the context of IoT, as discussed by Sokolova and Lapalme (2019), is crucial. These evaluations focus on metrics such as accuracy, precision, recall, and F1 score to assess the performance of classification models.

Recent studies on IoT data management techniques have emphasized the significance of efficient data classification strategies to address the challenges posed by the vast amount of data generated by IoT devices. Here are some key findings and insights from these studies:

1. Study on Machine Learning-Based Data Classification for IoT: A recent study published in the Journal of Internet Technology (JIT) proposed a machine learning-based approach for data classification in IoT environments. The study highlighted the importance of accurate and efficient data classification to enable timely decision-making and resource optimization in IoT systems. By leveraging machine learning algorithms such as support vector machines (SVM) and random forests, the proposed approach demonstrated improved performance in classifying diverse types of IoT data, including sensor readings, images [34]. and audio signals [35].

2. Research on Context-Aware Data Classification in Smart Environments: Another study presented in the IEEE focused on context-aware data classification techniques for smart environments [36]. The research emphasized the need to consider contextual information, such as location, time, and user preferences, when classifying IoT data streams. By integrating contextual features into the classification process, the study showed enhanced accuracy and relevance in identifying meaningful patterns and events from heterogeneous IoT data sources.

3. Investigation of Federated Learning for Distributed Data Classification in IoT Networks: A study conducted by researchers at a leading university explored the potential of federated learning for distributed data classification in IoT networks. Federated learning enables collaborative model training across decentralized IoT devices while preserving data privacy and security. The study evaluated different federated learning algorithms and communication protocols to assess their suitability for IoT data classification tasks. The findings highlighted the effectiveness of federated learning in handling large-scale IoT datasets distributed across edge devices and cloud servers [37].

4. Review of Deep Learning Techniques for IoT Data Classification: A comprehensive review published in the ACM Computing Surveys journal surveyed deep learning techniques for IoT data classification. The review examined various deep learning architectures, including convolutional neural networks (CNNs) [38]. recurrent neural networks (RNNs), and deep belief networks (DBNs), for analyzing and categorizing IoT sensor data. By leveraging the hierarchical representations learned from raw sensor inputs, deep learning models demonstrated superior performance in detecting anomalies, predicting events, and classifying IoT data streams with high accuracy and efficiency.

5. Application of Blockchain-Based Data Classification in IoT Security: A recent study presented at a cybersecurity conference investigated the use of blockchain technology for secure and verifiable data classification in IoT systems. By leveraging blockchain's immutable ledger and consensus mechanisms, the proposed approach ensured the integrity and trustworthiness of classified IoT data, mitigating risks associated with data tampering and unauthorized access. The

study demonstrated the feasibility of integrating blockchain with machine learning algorithms for robust and transparent data classification in IoT environments [39].

Overall, these recent studies underscore the importance of efficient data classification strategies in IoT data management, highlighting the potential of advanced techniques such as machine learning, context-aware computing, federated learning, deep learning, and blockchain to address the complex challenges associated with analyzing and categorizing heterogeneous IoT data streams. By leveraging these techniques, organizations can unlock valuable insights from IoT-generated data to support various applications ranging from predictive maintenance and resource optimization to personalized services and intelligent decision-making.

### 2.3 Summary of Related Works

The summary in table 2.1 provides a concise overview of the two primary themes covered in the literature review, the key aspects within each theme, and the significant studies that have contributed to these areas:

**Table 2.1:** Summary of the Literature Review

| Theme | Key Aspects | Significant Studies |
|---|---|---|
| IoT Data Management | Data Volume and Velocity, Data Variety and Veracity, Security and Privacy | Al-Fuqaha et al. (2015), Gubbi et al. (2023), Manyika et al. (2015), Roman et al. (2023), Sadeghi et al. (2015) |
| Machine Learning for IoT Device Classification | Use of Machine Learning Algorithms, Random Forest in IoT, Addressing Imbalanced Datasets, Performance Evaluation | Buczak and Guven (2016), Breiman (2021), Liaw and Wiener (2022), He and Garcia (2019), Chawla et al. (2022), Sokolova and Lapalme (2019) |

This literature review establishes a solid foundation for understanding the current state of IoT data management and the application of machine learning for device classification. It highlights the progress made in these fields and underscores the need for further research to address the remaining challenges, particularly in the context of handling imbalanced datasets and ensuring the scalability and adaptability of machine learning models in rapidly evolving IoT environments.

## 2.4 DISCUSSION

State-of-the-art machine learning algorithms, including deep learning approaches, ensemble methods, and hybrid models, offer promising solutions for IoT data classification tasks due to their ability to handle complex, high-dimensional data and extract meaningful patterns and features. Here's a discussion on these advanced techniques and their applicability in IoT data classification:

### 2.4.1 Deep Learning Approaches

Convolutional Neural Networks (CNNs): CNNs have been widely used for image based IoT data classification tasks, such as object recognition and scene understanding. CNN architectures leverage convolutional layers to automatically learn hierarchical representations from raw sensor data, making them well-suited for analyzing visual information captured by IoT devices like cameras or surveillance systems [40]. Recurrent Neural Networks (RNNs): RNNs are effective for sequential data processing and have been applied to time-series IoT data classification tasks, such as anomaly detection and event prediction. RNNs, including variants like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), can capture temporal dependencies and long-range dependencies in sensor data streams, enabling accurate classification of dynamic patterns and behaviors.

Autoencoders and Variational Autoencoders (VAEs): Autoencoders and VAEs are used for unsupervised representation learning and anomaly detection in IoT data. By learning compact and informative representations of sensor data, autoencoder-based models can identify abnormal patterns or outlier’s indicative of anomalies or faults in IoT systems, enhancing the reliability and robustness of data classification [41].

### 2.4.2 Ensemble Methods

Gradient Boosting Machines (GBMs): GBMs, such as XGBoost and LightGBM, are powerful ensemble learning techniques that combine multiple weak learners to improve classification accuracy [42]. GBMs are well-suited for IoT data classification tasks with heterogeneous features and noisy data, as they can handle missing values, nonlinear relationships, and complex interactions effectively [43].

Random Forests (RFs): While Random Forests are already mentioned, they remain relevant due to their robustness and scalability in handling large-scale IoT datasets with diverse features [44]. RFs can provide accurate and interpretable classifications for IoT data, making them suitable for applications requiring transparency and explainability in decision-making processes.

### 2.4.3 Hybrid Models

Deep Belief Networks (DBNs): DBNs combine deep learning and probabilistic graphical modeling to learn hierarchical representations of IoT data and capture uncertainty in classification tasks. DBNs leverage unsupervised pre-training followed by fine-tuning with supervised learning, enabling efficient feature learning and classification in semi-supervised or weakly labeled IoT datasets [45].

Neuro-Fuzzy Systems: Neuro-fuzzy systems integrate neural networks and fuzzy logic to handle uncertainty and imprecision in IoT data classification. By combining the flexibility of neural networks with the interpretability of fuzzy systems, neuro-fuzzy models can adaptively learn from data and expert knowledge to classify complex and uncertain IoT phenomena, such as environmental monitoring or health diagnostics [46].

These state-of-the-art machine learning algorithms offer diverse capabilities and strengths for IoT data classification, allowing organizations to leverage advanced techniques tailored to their specific application requirements and data characteristics. By harnessing the power of deep learning, ensemble methods, and hybrid models, organizations can achieve accurate, efficient, and scalable classification of IoT data to enable various applications, including predictive maintenance, anomaly detection, smart surveillance, and personalized services.

# 3. MATERIALS AND METHODS

This chapter covers fundamental concepts and definitions for the development of this research. The next sections of this chapter are organized as follows: section 3.1 addresses the classification of network traffic; Next, section 3.2 addresses the concepts of machine learning; section 3.3 presents the rationale for the Internet of Things and its technologies; Finally, section 3.4 presents final considerations.

## 3.1 Network Traffic Classification

This section presents the general and introductory concepts related to network traffic classification, as well as the most commonly used techniques.

### 3.1.1 Overview

Traffic classification is an area in computing that has attracted a lot of attention.

Interest from the academic community and industry due to network management possibilities such as, for example, QoS (Quality of Service), anomaly detection, infrastructure management, provisioning and resource allocation it is possible to state that classifying network traffic is not a simple collection of packets or flows, the whole process is associated with understanding the dynamics and behavior of network traffic, promoting understanding through extraction of characteristics that make it possible to associate their origin, their formation, their derivation, their composition and their impact [47].

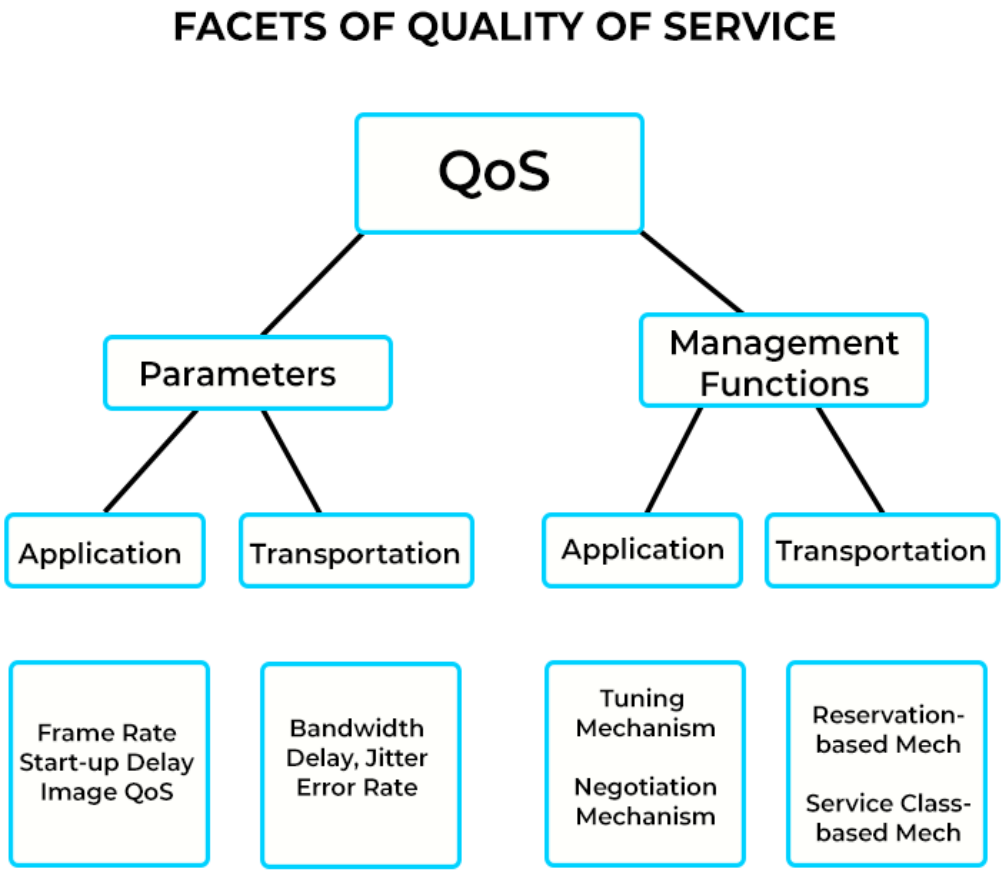


**Figure 3.1:** IOT QoS measurement [47].

As the Internet evolves and with-it devices and computing models, the complexity of many of its processes is also growing at an accelerated pace. Currently, we already have a large-scale presence of video and audio streams, games and file sharing [48]. states, in his doctoral thesis, that network operators, researchers and even ISPs need to know the traffic characteristics of their networks to manage resources or even charge users based on their consumption. From this type of need, various methods of classifying network traffic arose. The main methods found in the literature are presented in Table 3.1, along with some of their main features.

**Table 3.1: Techniques Used in Traffic Classification [48].**

| Approach | properties | Cost | Accuracy | Complexity |
|---|---|---|---|---|
| Ports | Access to doors | Low | Low | Low |
| Stochastic | Subscriptions | Variable | Variable | Variable |
| DPI | *Payload* | Low | Variable | High |
| Statistic | Flows and Packages | Moderate | High | High |

### 3.1.2 Classification based on port analysis

It is possible to state that in the past, companies, and ISPs (Internet Service Providers) were able to classify traffic easily, mainly due to the network be composed of few devices and protocols, enabling classification through knowledge of the port numbers. Initially, its use was sufficient to classify the network, as almost all applications used fixed ports signed by IANA. For example, web applications through port 80 (HTTP), emails through port 25 (SMTP – Simple Mail Transfer Protocol) to send and port 110 (POP3 – Post Office Protocol) to receive. Table 3.2 presents some examples of what the literature calls well-known doors.

**Table 3.2:** Examples of Well-Known Ports [49].

| Port number | Application |
|---|---|
| 30 | FTP Data |
| 31 | FTP Control |
| 32 | SSH |
| 33 | Telnet |
| 35 | SMTP |
| 43 | DNS |
| 60 | HTTP |
| 130 | POP3 |
| 423 | HTTPS |
| 424 | Syslog |

This method classification is inefficient and ineffective due to its imprecision and incompleteness. Related to the accuracy of the model, it is difficult to stipulate the values, since the characteristics of the monitored network may vary, but there are studies that present its accuracy between 50%70% in the best cases [50].

### 3.1.3 Payload-based classification

The DPI was originally designed with the intention of improving network security. It emerged through the combination of the functionalities of the intrusion detection system (Intrusion detection System (IDS)) and the intrusion prevention systems (Intrusion prevention system (IPS)). Furthermore, it emerged as an alternative to the problem of low accuracy in port classification. This method analyzes the content of packages in search of application characteristics or

signatures. the technique aims to identify applications that use strategies to camouflage themselves in traffic, an example of which is P2P (Peer-To-Peer) [51].

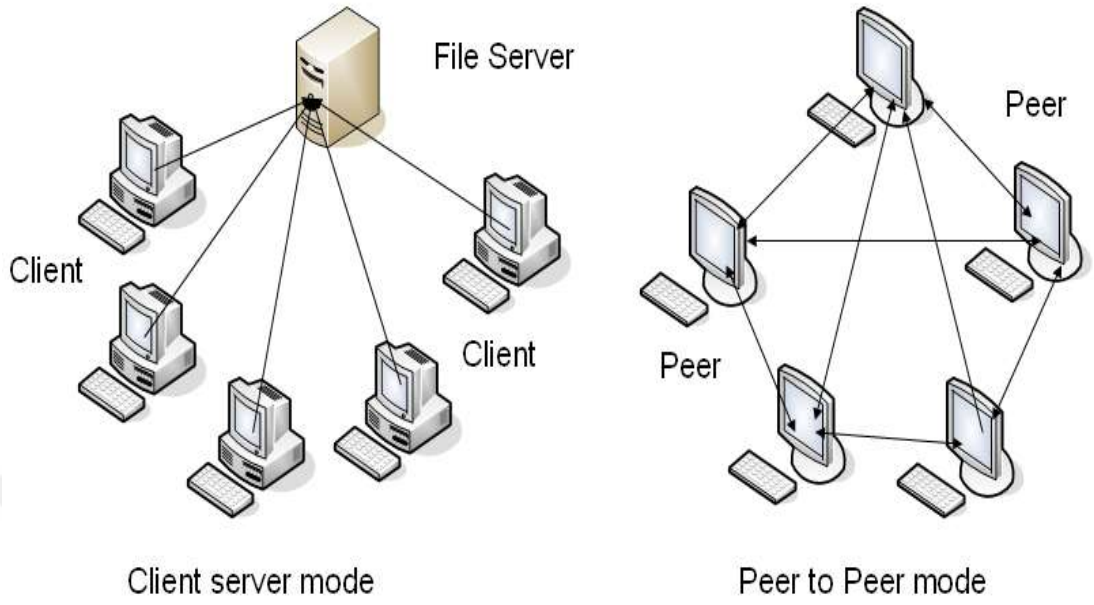


**Figure 3.2:** P2P vs Client Server Model [51].

There are key challenges in using DPI classification, including the need to continuously update the set of application signatures to classify new applications and versions. Despite this, the DPI classification technique remains one of the most used techniques. Table 3.3 presents examples of signature patterns for some P2P applications used to classify them [52].

**Table 3.3:** Examples of DPI Strings, Based On [52].

| Application | String Used | Transport protocol |
|---|---|---|
| BitTorrent | "0x12Bit" | TCP |
| eDonkey | "0xe319110000" | TCP and UDP |
| Gnutella | "GNUTGIV" | TCP |
| Gnutella | "GND" | UDP |

Due to the need to access content of packets, the technique must deal with major privacy challenges, an example of which is that some countries restrict access to users' communication content through regulations or laws.

### 3.1.4 Classification based on flow characteristics

The statistical classification of suracted as an alternative to the use of DPI, arising from concerns about privacy policies and packet payload analysis. The

statistical features are extracted by grouping the packets in the form of a flow [53]. In the end, the classification will consist of the statistical comparison of unknown traffic, or generated by an unanalyzed source, with previously stipulated rules systems based on machine learning (ML) learn through empirical data and, in this way, automatically associate objects with corresponding classes. According to the author, algorithms can be divided into supervised and unsupervised. in systems that use supervised machine learning algorithms, the classes are already defined in advance by the researcher, thus, the sample objects are provided to the system labeled with their respective classes; while in unsupervised algorithms, the system identifies distinct classes and assigns objects to them by affinity (for example, using clustering techniques) [54].

The main objective of statistical traffic classification is to categorize the flow of the network according to the generating application, which is, for example, based on analysis of network characteristics, such as packet size and interleaving time (between packets), Furthermore, the statistical approach is characterized as a high-speed and accurate model, but which presents a high complexity in its development compared to the others presented [55].

### 3.1.5 Hybrid methods for classifying network traffic

Using ML algorithms for feature-based traffic classification extracted from the flow receives substantial academic attention. Likewise, content-based identification, which makes use of signature patterns, continues to be widely used, for example, in IDS (Intrusion Detection Systems). It is noticeable, in recent studies, the construction of several hybrid solutions to classify network traffic based on machine learning methods along with characteristics extracted from the content.

However, its applications have become less effective in cases where network traffic is encrypted. Therefore, other proposals have emerged to mitigate these problems, focused on replacing the traditional pattern checking system with more sophisticated statistical methods or even based on a combination of techniques. As an example of hybrid proposals for classifying network traffic, [56]. The authors first used signature classification (DPI) and for traffic classified as Unknown, ML classification was applied, using decision trees. The combination of techniques

reinforces the need to develop and improve techniques for classifying network traffic, given that the Internet and its complexity have increased significantly [57].

### 3.2 Machine learning

In this section, some concepts related to Machine Learning will be discussed with a focus on definitions, models for performance evaluation and feature selection ristic.

#### 3.2.1 Overview

Machine learning is associated with improvements in performance of computer programs through the acquisition of knowledge through task experiences. the statistical learning process plays an essential role in several fields of science, from decision making to finance, and has been used in various emergency situations, such as preventing heart attacks and identifying risk factors. A fundamental factor in machine learning is the balance threshold between performance and quality. Defines performance evaluation as the way of optimizing the use of computational resources through measurable measures that allow the identification of expenditure [58].

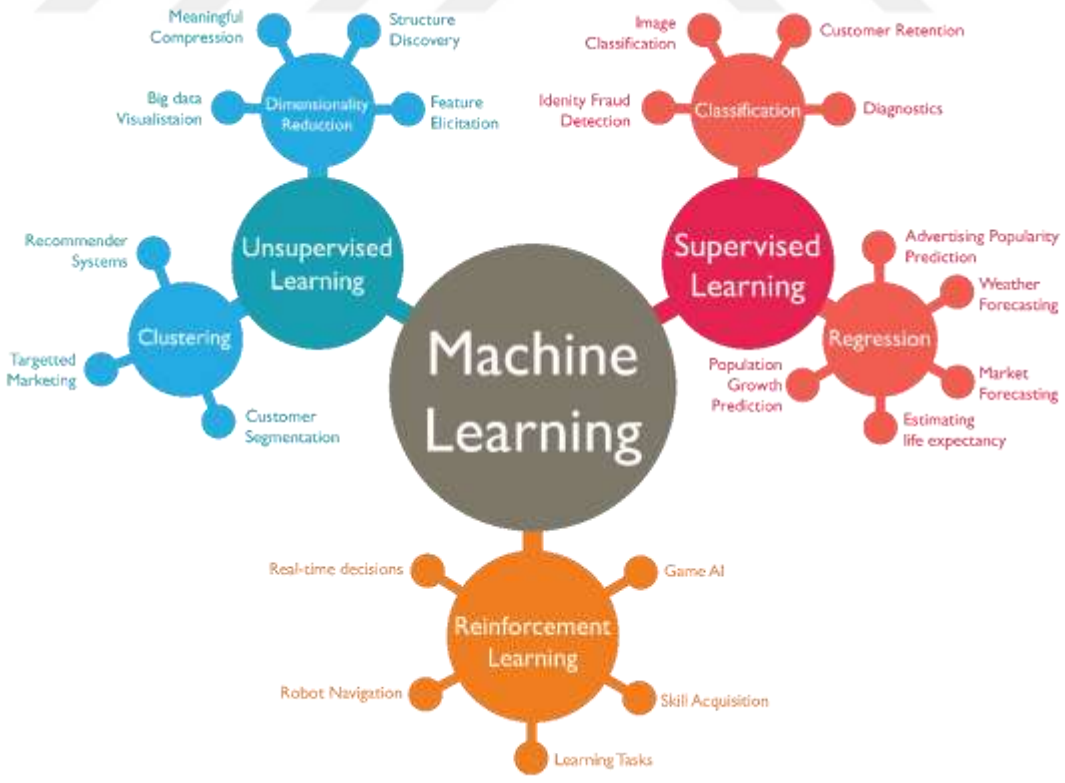


**Figure 3.3:** Machine Learning Subbranches [59].

The process of building or using systems based in machine learning it manifests itself through the use of characteristics, which, according to the author, are processed through a model called feature engineering. According to the author, properly using the characteristics is fundamental for the design of a project, despite the fact that a great effort is required for its correct construction, mainly due to the complexity of the assembly process [60]. new ML systems focus on the optimized combination of resources and their greatest capacity arises from the correct relationship between data input and algorithm output. According to the author, by carrying out this correct application it is possible to reduce human intervention, improve performance when using a large volume of data and more complete solutions to complex problems it can be concluded that the use, in practice, of *Machine Learning*passes through a strict methodology. Firstly, related to the choice of algorithm for application to a given problem, it is essential to consider the combination of three main components: 1 - Representation: the classifier must be represented in some language that the computer understands; 2 - Evaluation: a function must be considered to distinguish between good and bad classifiers; 3 - Optimization: classifiers with greater performance and accuracy must be selected. The choice of optimization technique is crucial for greater algorithm efficiency [61].

#### 3.2.1.1 Supervised classification

The classification method is considered supervised when extracts learning structures to classify new instances into predefined classes. According to the author, this model consists of performing the classification using training (already classified database) and subsequently the algorithm outputs are given based on it scorrelation [62].

#### 3.2.1.2 Unsupervised classification

Unsupervised methods, different from supervised methods, do not require a complete labeled data set for training, as the method itself discovers how to associate the data through similarities. this method does not contain a supervisor and, consequently, there is no correct mapping between the desired inputs and outputs. However, more frequent input patterns are identified to relate similarities and create a grouping of output sets [63].

### 3.2.2 Pandas library

Pandas' library is a powerful and useful Python library designed to analyze and organize data in an easy and efficient way. It is a powerful data analysis tool in Python, combining ease of use with the ability to handle large and complex dataset one of the utmost important features.

1. Flexiblability of data structure: Pandas provides to deal dataset with flexible capability, that allows excellent storage and organization of data, such as Series structure and Dataframe, which facilitates reading, writing and analysis operations.
2. Graphical analysis capabilities: Pandas provides many built-in functions and techniques for data analysis, such as sorting, grouping, filtering, transformation, statistical analysis, and many other operations that facilitate data exploration and understanding.
3. Fast processing abilities: Pandas relies on the NumPy library to perform processing operations efficiently, making it suitable for large and complex data.
4. Handling dataset with efficient time computation: Pandas can be easily used to handle time-based data such as recording, analyzing, filtering and transforming time-based data.
5. Integration with several types of libraries: Pandas can read and write data to and from various sources such as CSV files, Excel, SQL databases, JSON, and many other formats.
6. Graphing: Pandas provides an integral library called Matplotlib to plot data and create graphs easily and efficiently.

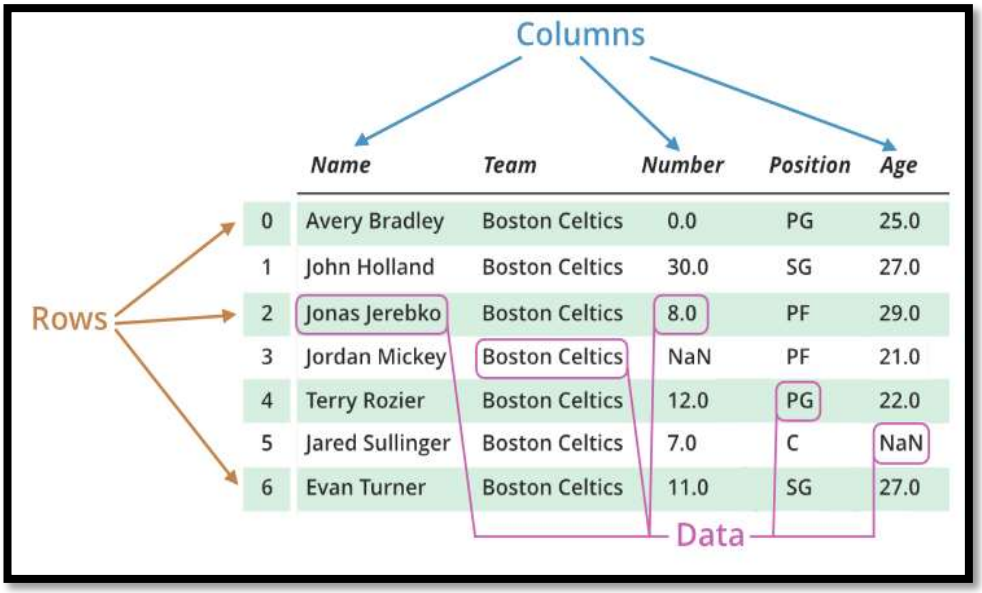


**Figure 3.4:** Pandas Deal With Data Structurs

### 3.2.3 Random Forest

Random Forest (RF), is a widely used supervised automatic learning algorithm that has high performance this algorithm has a series of advantages that make it relevant in research, including: great resistance to overfitting; requires a small number of parameters; has low variation, the use of multiple trees reduces the chance of failures occurring during classification due to the relationship between training and test data [61].

RF is used in a wide variety of research areas, such as detection Due to its excellent results, its wide range of applications and its advantages, this algorithm is ideal for use in network traffic classification [65].

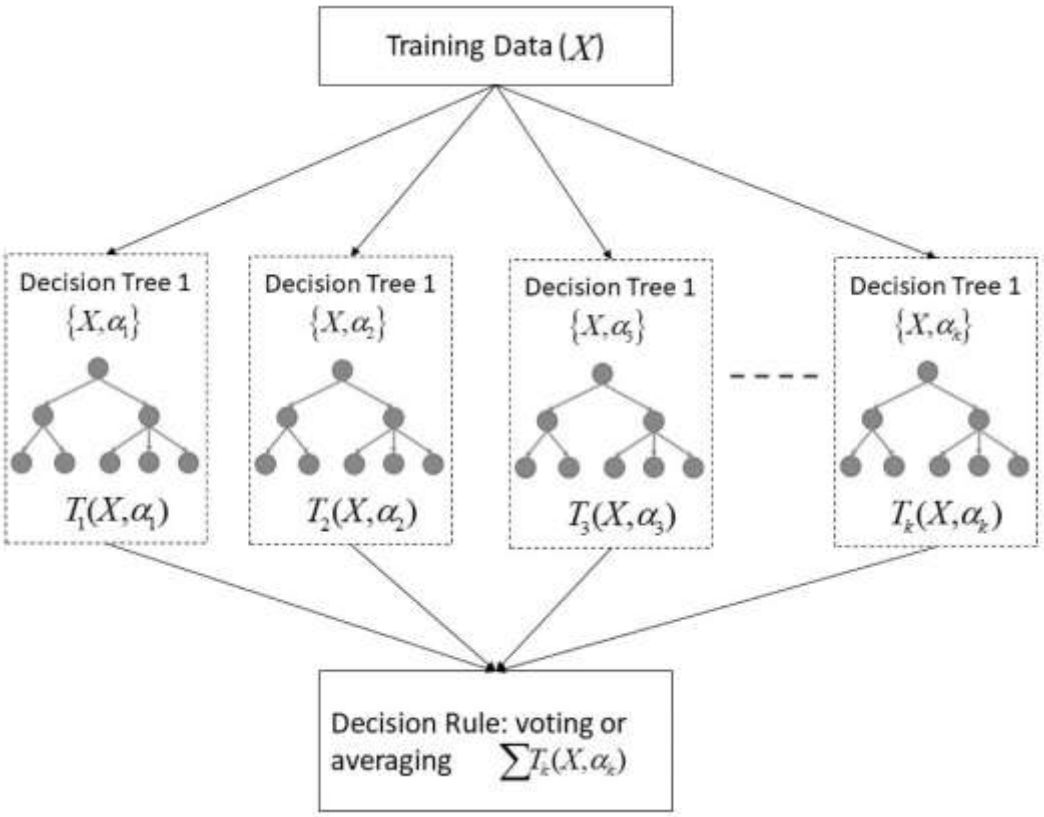


**Figure 3.5:** Random Forest Classifier [66].

#### 3.2.3.1 Formulation of the decision tree used

RF builds multiple decision trees and aggregates them to perform classification. decision tree is an inductive statistical model used in supervised machine learning. The classification takes place, after building the tree, by traversing the root node to the leaf node. Decision trees are based on partitioning features into certain sets and fitting them to some simple model, equal to a constant. The model used for tree classification was based on Classification and regression tree (CART), as described in which is similar to C4.5 (extension of the ID3 model) [74].

### 3.2.4 Performance metrics for statistical classifiers

to validate the quality of the results of a classification by ML it is necessary to use performance evaluation through the confusion matrix (Table 3.4). Calculations are performed by arranging the values in the matrix and then calculated as shown in Table 3.5.

**Table 3.4:** Confusion Matrix [67].

| Test Validation | | | |
|---|---|---|---|
| Test | Gift | Absent | Total |
| Positive | TP | FP | TP + FP |
| Negative | FN | TN | FN + TN |

There are 4 possible ways to express the measurements are they:

- *True Positive* (TP): – True positives
- *False Positive* (FP): – False positives
- *False Negative* (FN): – False negatives
- *True Negative* (TN): – True negatives

Using the values displayed in the confusion matrix (Table 3.4), it is possible to produce values representing the responses to the intended assessments.

**Table 3.5:** Metrics Using the Confusion Matrix [67].

| Name | Formula |
|---|---|
| Accuracy | $\frac{TP+TN}{TP+TN+FP+FN}$ |
| Precision | $\frac{TP}{TP+FP}$ |
| *Recall* | $\frac{TP}{TP+FN}$ |
| *F1-Score* | $2TP$ ———— |
| MCC | $\frac{FN}{(TP+FP)\quad (TP+FN)(TN+FP)(TN+FN)}$ |
| FPR | $\frac{FP}{TN+FP}$ |

Confusion matrix calculation allows a series of precise assessments regarding to the results obtained. Among the calculations presented we have accuracy, which allows identifying the correct classification proportions regardless of true or false. On the other hand, precision, which is analogous to positive predictive value (PPV), allows us to identify the proportion of true positives in relation to all positive predictions. Recall, also known as sensitivity, corresponds to the system's ability to correctly predict the condition for true cases [68]. To compensate for possible distortions in the analysis, F1-Score balancing is used, which indicates the adjustment of results in relation to accuracy and recall. Another interesting metric for analysis is the PHI coefficient, also known as MCC (Matthews correlation coefficient), which performs quality assessment in classifiers with its value set in the

range between -1 and 1. The closer to 1, the higher the quality of the prediction and the closer to -1 means total disagreement between the prediction and the observation. Finally, there is the false positive ratio (FPR) to evaluate the number of false positives in relation to the total that should not have been identified.

### 3.2.5 Selection of statistical attributes - feature selection

Feature selection, also known as variable or statistical attribute selection, is the process of selecting the most relevant subset of data for use and building the model that will be used in ML. According to the authors, its use is essential to simplify the construction of complex models, reduce dimensionality, reduce training time, reduce superposition (overfitting) and variance [69].

Extremely high-dimensional data presents serious challenges to existing learning methods. Also, according to the authors, the large number of statistical attributes tends to alter performance. Due to the problem of dimensionality, techniques were studied and developed [69].

with the aim of reducing it. The objective of these techniques is to choose a small subset of the most relevant statistical attributes, in accordance with a given evaluation criterion, promoting better performance, lower computational cost and better model formulation [69].

Feature selection methods are divided in four main steps:

1. Generation of subsets, in which a candidate subset will be chosen based on a given Search Strategy.
2. Evaluation of subsets, where verification occurs in accordance with defined evaluation precepts.
3. Stopping criteria, at this stage the subset that best suits the evaluation criteria will be chosen among all the candidates evaluated.
4. Validation of results, in which the chosen subset will be validated using a validation set.

### 3.3 Internet ofThings

This section will cover concepts, expectations and technologies associated with Internet of Things ecosystems.

#### 3.3.1 Overview

The Internet of Things presents itself, according to the literature, as a new computational paradigm capable of integrating a wide variety of heterogeneous systems, promoting the connection of data, people, objects and applications via the Internet By collecting this data, we can build a wide variety of applications with the possibility of increasing functionality through the use of various techniques such as AI (Artificial Intelligence) and Analytics. Its emergence took place in 1999 in the laboratories of MIT (Massachusetts Institute of Technology), through research with RFID sponsored by Kelvin Ashton, co-founder of Auto-ID (it is possible to say that IoT is promoting a great revolution and has been advancing in different areas and domains such as embedded systems and telecommunications. Furthermore, the main factors for the great evolution of the Internet of Things are smart objects, as they have processing, connectivity and data collection capabilities [45].

Since the emergence of IoT, several authors have given it many definitions, including they are that in which IoT focuses mainly on connectivity and sensor requirements of connected devices in typical environments. Whereas these statements reflect the basic requirements of IoT, other definitions focus more on the need for ubiquitous and autonomous networks, in which the identification and integration of services play a fundamental role. For example, Internet of Everything (IOE) is a broad term used by Cisco to refer to people, things, and places connected to the global Internet IoT is characterized as a paradigm that presents a high degree of autonomous data capture, connectivity, interoperability, mobility and event transfers [68].

IoT provides users and companies with a wide variety of functionalities, promoting a high increase in the capacity for interaction and communication with the environment, all through smart devices and the Internet. defines IoT as a dynamic autonomous global network infrastructure supporting interoperability and the use of several standardized protocols, in which physical and virtual objects have identifiers, attributes, and their own behavior, through optimized interfaces. IoT emerges as a

model of coexistence of device networks in which all devices are capable of interacting with each other through various gateways and middleware supported by a complex control and management plane. Therefore, the network infrastructure must promote the integration of these various infrastructures, in this way all systems or applications, based on IoT, will be able to obtain better performance in providing their services through efficient sharing of information and resources [72].

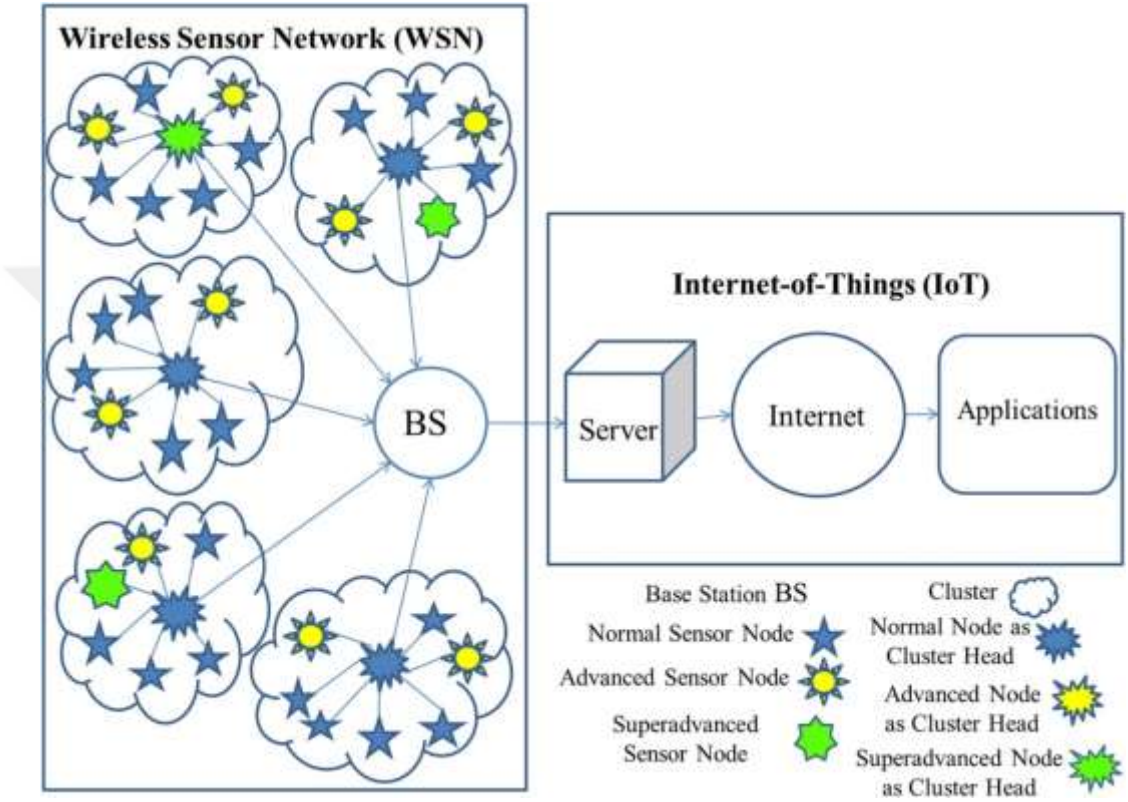


**Figure 3.6:** IOT and WSN Integration [72].

IoT is related to the next generation of Internet. the network composed of IoT devices will have, in the not-too-distant future, trillions of nodes, including a wide variety of ubiquitous devices, arranged in a plurality of environments and equipped with sensors interconnected to the network. it is possible to state that IoT is related to a series of technologies, among which we can mention: Wireless sensor networks (WSN); IPV6; cloud computing; and ubiquitous computing. Due to the great investment expectations, many projections arise about its current state, among them which presented an estimate of 8.6 billion devices by the end of 2017, growth of more than 31% compared to 2016 (6.3 billion), projects a total of 20 billion connected devices for 2020 states that the IoT market will grow annually by around

28.5%, this corresponds to gross growth from 157 billion dollars in 2016 to 428 billion in 2020. [72].

Some of the main Challenges in the large-scale use of IoT and its use through the combination of multiple techniques are related to:

1. Massive data collection,
2. Scalability and diversity,
3. Security requirements,
4. Energy consumption,
5. BigData (Data Collection and Analysis - DCA),
6. Fault tolerance,
7. Analytics.

These challenges are being analyzed and developed through research by academia and industry with the aim of promoting better integration of technologies. Another important factor in IoT ecosystems is that their composition essentially depends on three main components which are [72].

- Physical Components → Electronic devices, sensors that are arranged in environments to collect data or respond according to the proposal, intelligent objects and actuators.
- Communication Systems → Data transmission technologies based on wired or wireless networks, whether mobile or not.
- Information Processing → Implemented through programs, with the use of AI or not, to control and manage systems.

### 3.3.2 IoT Architectures

It can be stated that IoT application architectures are typically subdivided into four main layers (see Figure 3.6). Are they [73].

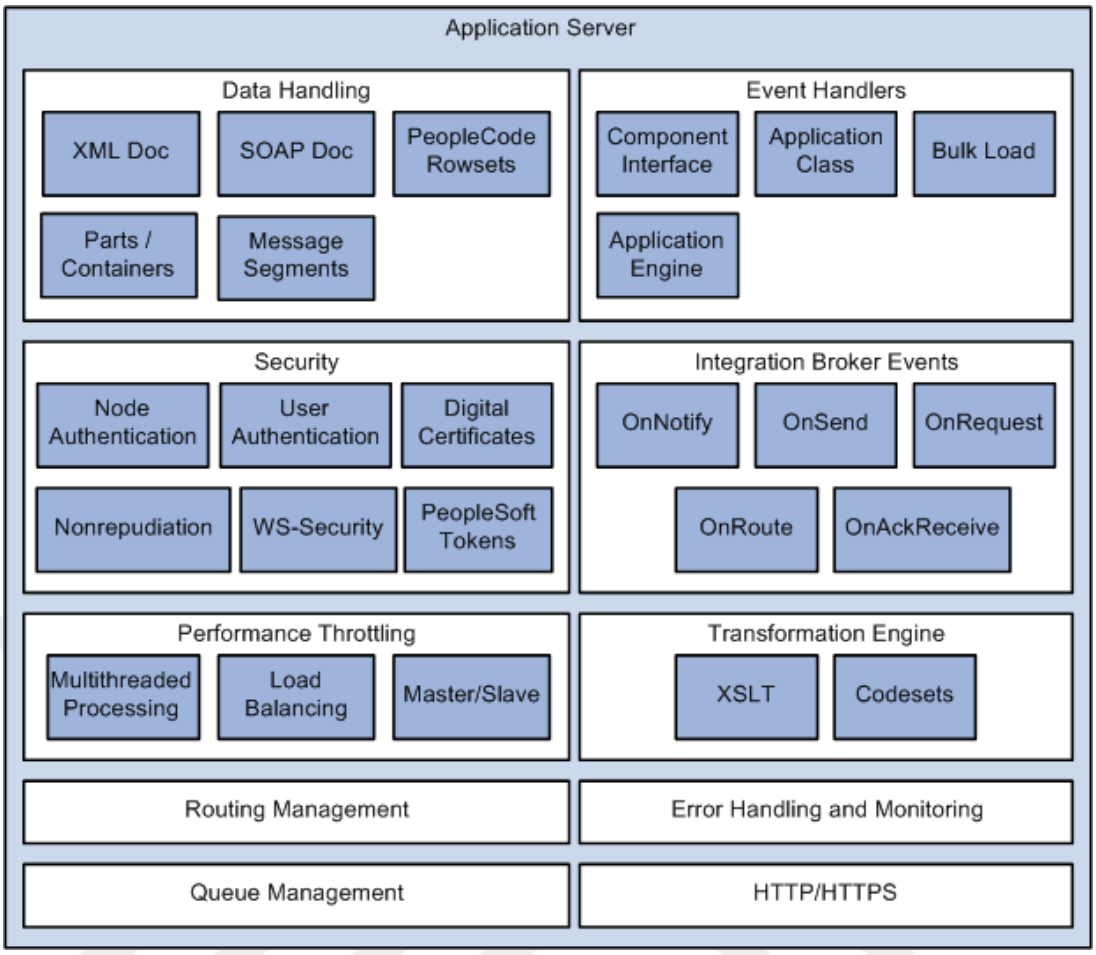


**Figure 3.7:** Component Integration Architecture [73].

- Application → Web, Mobile
- Service Management→ Responsible for information security, security control, device management, management and data abstraction.
- Gateway and Network → LAN, PAN, massive data volume, QoS, scalability.
- Connectivity and Sensors → Low consumption, WSN (Wireless Sensor Network), low data rate.

At the beginning of the architecture, technologies such as sensors, actuators and *tags*. This layer is responsible for collecting data from the environment. In the second layer there is the Gateway and the network, responsible for routing the data collected in the lower layer and sending it to the service management layer. In addition to routing, it is responsible for ensuring the interoperability of systems, as different IoT devices communicate differently through different protocols. The third layer is called service management and is responsible for security guarantees, QoS

guarantee and information analysis. The last layer makes use of the collected data in the form of services for users [73].

### 3.3.3 Technologies associated with IoT

IoT is characterized by making use of heterogeneous technologies (see Figure 3.7) and as devices are being inserted; new scalability, interoperability and connectivity requirements are added. IPv6 is one of the preponderant factors in the addition of devices, in addition to protocols with low data rates [73].

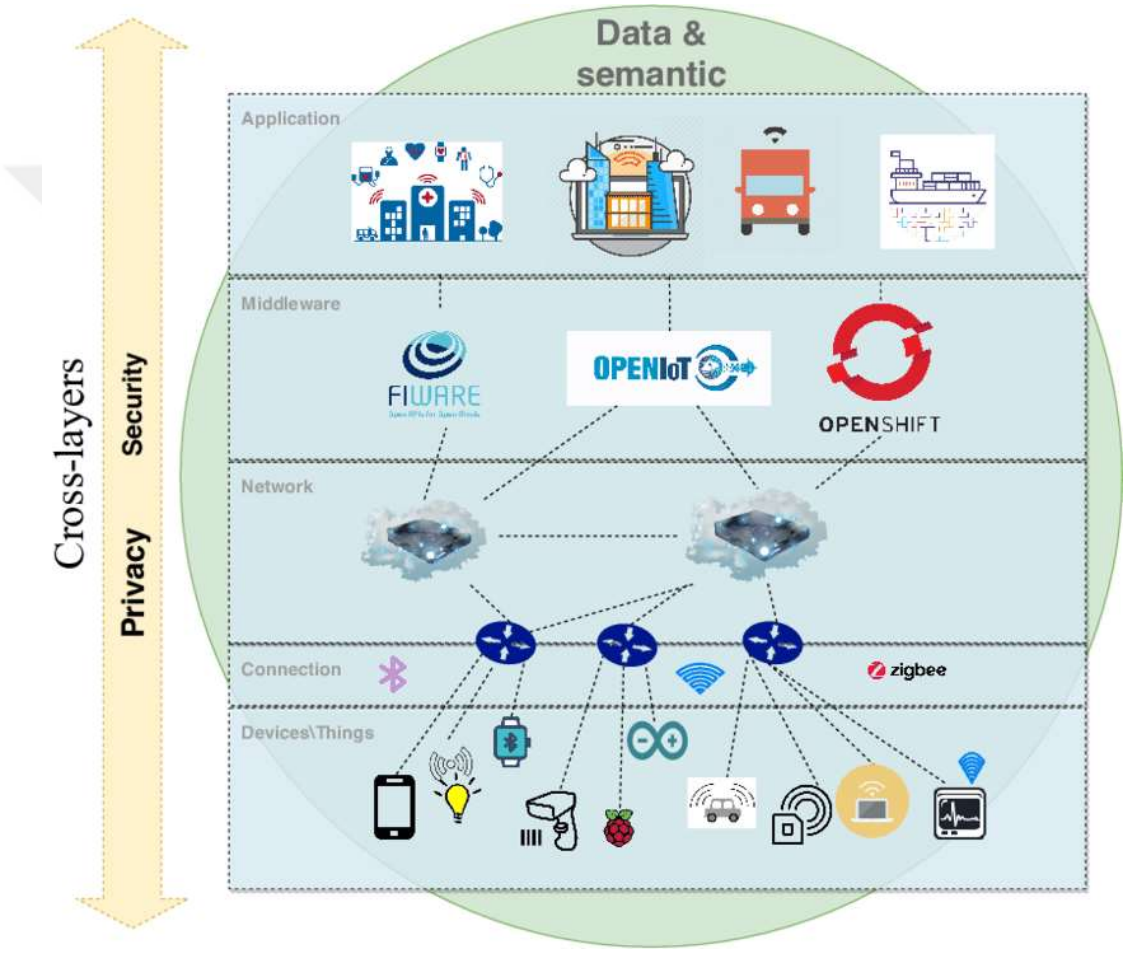


**Figure 3.8:** IoT Network Heterogeneity [73].

### 3.3.4 Basic framework for building IoT ecosystems

IoT represents a major evolution that has been emerging alongside advances in other computing domains, which promotes and guarantees technological complementation, in addition to playing an important role in the integration and communication of the physical with the virtual. some of the main structures for building IoT ecosystems are [73].

1. Identification: Requires mechanisms that guarantee the unique identification of devices on the network, using different technologies such as IP,
2. Communication: Represents the technologies used in IoT ecosystems that promote communication and interoperability guarantees,
3. Services: There is a wide variety of services offered in the context of IoT, this structure represents this diversity,
4. Semantics: IoT promotes data collection, this data requires analysis so that the maximum amount of knowledge can be extracted.

### 3.3.5 IOT deal with data

The optimized use of data from IoT ecosystems is an area of high complexity, hampered mainly by the need for optimized management and use. there are seven key situations that are considered complex when using this data, which are [74]. Security: There are a number of concerns when using data in IoT, one of which is user privacy. who in a Forbes report presents the identification of teenage pregnancies solely through the pattern of purchases through the use of Analytics. Concerns regarding security in IoT are a key part in its acquisition, as through data collected from users it is possible to cross-reference, infer and acquire crucial information such as travel and business preferences [75].

1. Data Volume: The challenge of managing and processing data is fundamental in the IoT context, its use through combinations of techniques can promote major changes in social perspectives.
2. Data Diversity: Complexity is present in the wide variety of data sources, including cars, refrigerators and cameras, for example. The great heterogeneity of devices promotes great complexity, especially in the integration of this data.
3. Data Speed: This item promotes the need to build time-sensitive applications. Real-time systems will have to process large volumes of data using Analytics regardless of the context inserted, the complexity tends to be very high.
4. Analytics: Promotes the conversion of raw data into relevant information for users. Its approach also focuses on improving work processes and extracting

valuable insights into market and consumer behaviors and trends, in addition to their expectations.

5. Data Economy: Data is used to build and operate applications, with the presence of diversity, transmission speed, processing and analytics. The data economy comes from the need to contain redundancies and waste of resources and its complexity comes from the need for deep knowledge about data to be able to optimize its use.
6. Logistics: Represents the optimized use of the environment to provide the best experiences. An example of logistics would be when there is integration of the IoT ecosystem in the cloud and processing using Fog Computing instead of sending all the data, as this would promote speed and less congestion.

### 3.3.6 IoT protocols

The literature presents constant evolutions in protocols that are designed to minimize or resolve certain pertinent problems, including transmission rate, data safeguarding and interoperability TCP/IP is considered the fundamental suite of Internet protocols, where IP provides connection between different networks, layer 3. TCP and UDP are located at layer 4 (transport) These fundamental protocols have different forms of representation, among them we have them available in the third and fourth layer of the layered model, as illustrated in Figure 3.8, using the last three associated protocols in their representation [76].

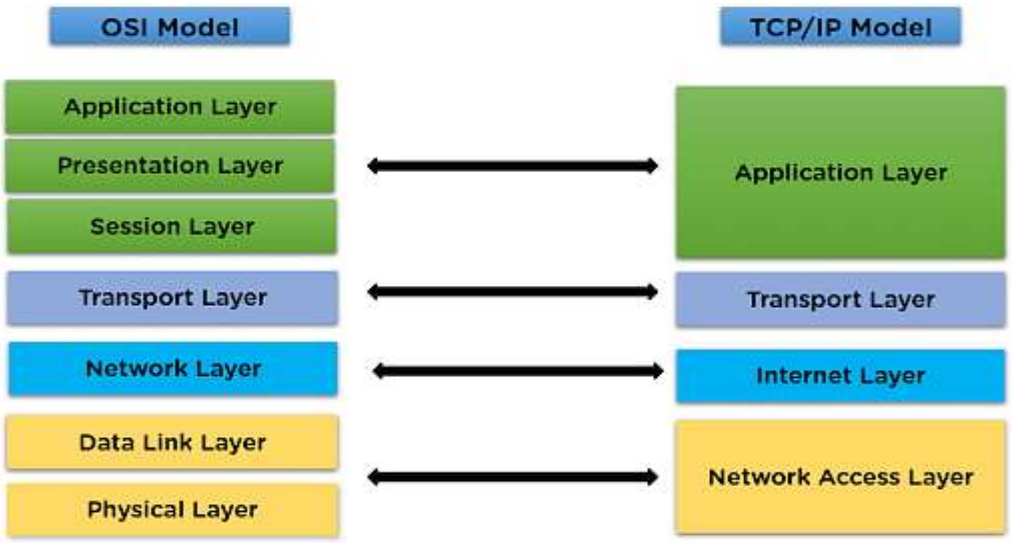


**Figure 3.9:** Layered model with TCP and IP [76].

What is most noticeable in the construction and improvement of protocols for IoT are the large demands that must be remedied to meet all device requirements, focusing mainly on high latency and decreasing network and storage usage. Much of the evolution of IoT protocols is present in the data transfer layer and the application layer. The following is a summary of widely used and researched protocols for IoT [76].

- CoAP: Constrained Application Protocol, an application layer protocol focused on Internet of Things devices that have limited resources, its characteristics are low resource consumption, translation to HTTP, easy implementation and support for multicast;
- MQTT: Message Queuing Telemetry Transport, a very popular protocol based on publish/subscribe protocols, its main applicability is for Internet of Things devices in M2M (Machine-to-Machine) communication and mobile devices;
- HTTPS: Hyper Text Transfer Protocol Secure, can be defined as an implementation of HTTP with the addition of a security layer that makes use of protocols such as SSL or TLS, allowing encrypted data traffic;
- XMPP: Extensible Messaging and Presence Protocol, an opensource and extensible communication protocol focused on promoting interoperable people-to-people communication. Its main characteristic is that its communication is text-oriented (XML);
- Z-Wave: It is a wireless communication protocol aimed at sending control commands and secondary data (for example, weather information). It is designed for use in a simple, reliable, low-power radio wave medium. This protocol does not provide sufficient support for sending audio or video;
- Others: In addition to the protocols discussed, for the IoT universe we have a large number of them d
- eveloped, from infrastructure (eg, 6LowPAN, RPL), identification of services and devices (eg, EPC, URIs), transport and communication (eg, Wi- fi, Bluetooth), discovery (eg, mDNS, DNS-SD) and semantics (eg, Web Thing Model).

The literature constantly presents the emergence or improvement of protocols for IoT. Among them, we can mention protocols that present certain singularities in communication, such as data rate and packet size [75].

1. Data Link:
    - IEEE 803.15.4
    - IEEE 803.11 AH
    - LTE-A
2. Network Layer Routing Protocols:
    - RPL (Routing Protocol for Low-Power and Lossy Networks)
    - CORPL (cognitive RPL)
    - CARP (Channel-Aware Routing Protocol)
3. Network Layer Encapsulation Protocols:
    - 6LoWPAN (IPv6 over Low power Wireless Personal Area Network)
    - IPv6 over Bluetooth Low Energy
4. Session and application layer protocols:
    - MQTT (Message Queue Telemetry Transport)
    - AMQP (Advanced Message Queuing Protocol)
    - CoAP (Constrained Application Protocol)

Despite the diversity of IoT protocols that have recently emerged and are emerging, main focuses in its application are the same: performing, improving and developing sufficient support for IoT ecosystems aimed at low computational power, reduced size (mostly) of packets, delay tolerance, quality of the communication link, protocol stacks, rate general data and interoperability [76].

### 3.4 Final Considerations

This chapter covers concepts and definitions related to classification of network traffic, machine learning and the Internet of Things, the focus of this dissertation.

Initially, the evolutions of techniques for network traffic classification and the ways in which they are used to classify traffic. Furthermore, machine learning classification and its metrics for performance evaluation were presented. The definitions of machine learning, the importance of feature engineering, and the definitions for evaluating metrics using a confusion matrix were also presented [76].

The focus of the contextualization used for IoT was to present some of the definitions employed, the most commonly used protocols, expectations regarding their use, the evolution and great economic potential of this technology [77].

# 4. PROPOSED METHOD

Through the utilization of a variety of modeling procedures, this methodology investigates big data pertaining to e-health that is based on the Internet of Things (IoT). Due to the fact that the vast majority of devices that collect data through the Internet of Things (IoT) are integrated with sensors, unique protocols are required. The classification of large volumes of data is determined by a variety of critical aspects, including the quantity of data and the type of data collected. While the retrieval of this data in real time is utilized for the purpose of testing, the retrieval of this data from the internet is utilized for the purpose of training. To obtain the data, both of these strategies are utilized [78]. A number of fundamental components are involved in the development of the Internet of Things. These components include the collection of data, the tracking of data, the sharing of data, the computerization of data, the control of data, and cooperation. The IDA optimization technique is performed after the data gathering phase has been finished in order to extract features from the MapReduce model. This is done in order to perform the extraction. In order to categorize the medical data in accordance with the most significant features that have been determined, you need make use of RFC. Customers can take use of a wide range of information services provided by the company, all of which are founded on the insights that are discovered through the utilization of big data [79].

## 4.1 Data Collection

### 4.1.1 Dataset

There are many types of datasets that used for different scienes such as kagge, scientific data journal, data in brief, Dataset Search (that provide by google), PhysioNet, UCI machine learning repository and alot of other that can provide data for reseachers. Whereas each once can offer several types of data with out any privece (which mean publical available) [79].

**Table 4.1:** Dataset That can be used for ITO with Different Domain.

| Dataset | Description | Sources |
|---|---|---|
| CGIAR dataset | Agriculture, Climate | CCAFS |
| Learning of mining | Teaching | University of Genova |
| Commercial Building Energy Dataset | Energy, Smart Building | IIITD |
| Individual household electric power consumption | Energy, Smart home | E.D.F. Clamart, France |
| AMPds dataset | Energy, Smart home | S. Makonin |
| UK Domestic Appliance-Level | Electricity Energy, Smart Home | Kelly and Knottenbelt |
| PhysioBank databases | Healthiness care | PhysioNet |
| Saarbruecken Voice Database | Healthiness care | Universitat des Saarlandes |
| T-LESS | Commerce | Czech Technical University |
| CityPulse Dataset Collection | Smart-City | City Pulse project |
| Open Data Institute – node Trento | Smart-City | Telecom Italia |
| Malaga datasets | Smart-City | City of Malaga |
| Gas sensors for house efficacy control | Smart-Home | U.C San Diego |
| Daily Living | Smart-Home | W.State University |
| ARAS Human Action Datasets | Smart-Home | Bogazici University |
| MERLSense Data | Smart home, building | Mitsubishi Electric Research Labs |
| SportV.U. | Athletes | Stats LLC |
| RealDisp | Athletes | O. Banos |
| Taxi Service Trajectory | Transport | Prediction competition |
| GeoLife GPS Trajectories | Transport | Microsoft Company |
| T-Drive trajectory data | Transport | Microsoft Company |
| Chicago Bus Traces data | Transport | M. Doering |
| Data of Uber Short-Sojourn | Transport | FiveThirtyEight |
| Recognition of way Signs | Transport | K. Lim |
| D.D.D.17 | Transport | J. Binas |

#### 4.1.1.1 Reasons of selected UCI dataset

The UCI Machine Learning Repository was created to archive data in 1987 at the Center for Machine Learning and Intelligence Systems at the University of Florida Irvine (UCI) in the United States of America [72]. At that time, it was used very widely by researchers and students in all countries of the world as an important and basic source of machine learning data collection. UCI is theories and databases that are used in machine learning algorithms. However, we utilized the dataet is open-acces from UCI. due to the UCI proived such many reasons involved, measure

performance, and compare methodology. The data in UCI is freely available to utilized. It has been processed to facilitate its use in machine learning. Each dataset comes with a description, information, and references the original source. The UCI repository is considered a very important and valuable resource for many fields such as engineering, economics, medicine, biology, physical sciences, economic sciences, and others. For this study, we have chosen the "IoT Device Identification Dataset" from the UCI Machine Learning Repository, a dataset that is highly relevant to the field of IoT and the objectives of this thesis [79]. This dataset is particularly suited for the task of IoT device classification, offering a real-world context for applying and evaluating the machine learning models developed in this research.

#### 4.1.1.2 Overview of the IoT device identification dataset

The IoT Device Identification Dataset is specifically curated for the purpose of identifying and classifying IoT devices based on network traffic data. It was collected and made available as a part of research efforts to improve security measures in IoT networks by accurately identifying devices [80].

#### 4.1.1.3 Features of the dataset

The dataset contains a variety of features extracted from network traffic, including packet sizes, timing information, and other protocol-specific attributes. These features are indicative of the behavior and characteristics of different IoT devices, making them suitable for classification tasks. The dataset encompasses a wide range of common IoT devices, including smart cameras, wearables, smart home appliances, and more [81].

#### 4.1.1.4 Classification target

The primary goal is to classify each network traffic instance into one of the several categories of IoT devices. This classification is crucial for IoT security applications, where correctly identifying devices can aid in detecting unauthorized access or abnormal device behaviour [82].

#### 4.1.1.5 Relevance to IoT data management

This dataset is directly relevant to IoT data management as it involves real-world data from IoT devices. The dataset's focus on network traffic data makes it an

excellent resource for exploring data management challenges specific to IoT, such as handling large-scale, high-velocity, and diverse data.

#### 4.1.1.6 Usage in this study

In our study, this dataset is used to demonstrate the effectiveness of the developed IoT data classification framework. The framework's ability to handle imbalanced data will be particularly pertinent, given the varying frequency of different types of devices in typical IoT networks. The dataset will be preprocessed using techniques discussed in section 4.2, including normalization and balancing using SMOTE, before being fed into the Random Forest classifier. The choice of the IoT Device Identification Dataset enables a practical and relevant evaluation of the proposed machine learning framework, ensuring that the research findings are applicable and valuable to the field of IoT data management and security.

#### 4.1.1.7 Select IoT devices classification

every device has unique features such as Mac address and ip address, then we catch their process such as receiving pand and sending and many Contact parameters ،however the system training on each device processes, that they have been saving on the data set, later any abnormal characteristics can be found in the Communication package through the monitoring system. Moreover ،the system has ability to send alerts and make teardown for the system beside that can Isolate the device until confirmed and be sure that the device is safely to use by showing the device process. We have been focusing on studying device characteristics and analysis them using Machine learning as well as we classify each device individually.

### 4.1.2 Cleaning the dataset

In order to enhance the services that are offered by the Internet of Things (IoT), it is possible to make use of the data that is gathered from interactions that take place between individuals, between individuals and systems, and between systems themselves. To the contrary, the data that is gathered from the Internet of Things itself can be utilized to enhance the functionality of the Internet of Things infrastructures, systems, and things. The utilization of Big Data technology would make it possible for any healthcare facility, regardless of the location of the test, to

check the records of a patient. This would be the case regardless of the location of the test. Figure 4.1 is a graphical representation model of this system that can be seen within this document. Additionally, the results would be saved in real time by this system, which would make it feasible to make judgments regarding the patient as soon as the test is finished [83]. Databases are frequently utilized for the purpose of storing infrastructures and other devices that are connected to the internet of things (IoT). Important information pertaining to healthcare is stored in these databases. This information includes the names of patients, their ages, genders, ailments, medications, and dosages. This interaction between systems is utilized to enhance the quality of the service, and the databases in question contain information pertaining to medical care.

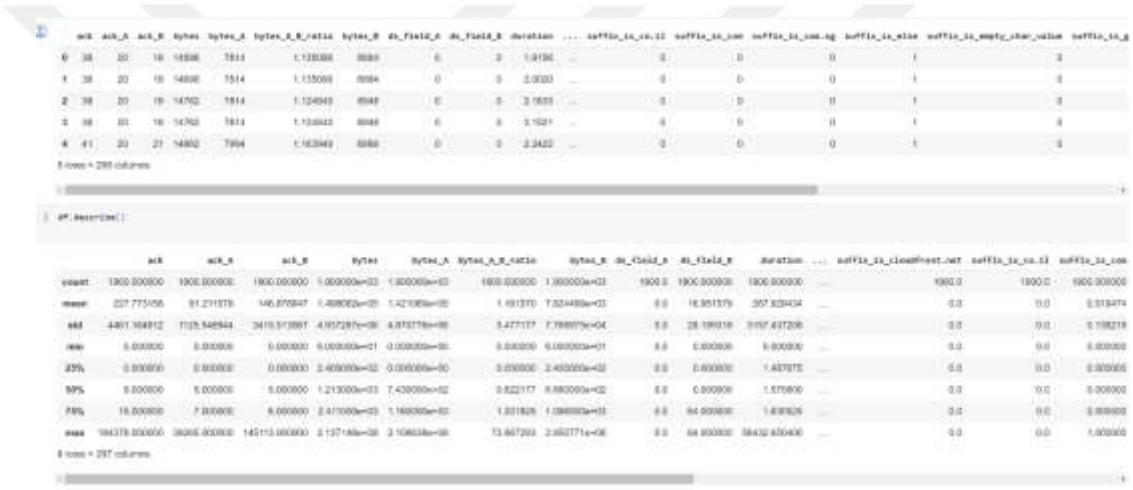

**Figure 4.1:** Sample of the Data Structure That Is Collected for the Analysis

### 4.1.3 COMPUTATIONAL RESOURCES

If the proposed framework is implemented using Google Colab, which provides a cloud-based environment with access to GPUs and TPUs, the computational resources required can be tailored to the specific needs of the framework. Here's a potential software and hardware specification for implementing the proposed framework on Google Colab:

#### 4.1.3.1 Software Specification:

1. Operating System: Google Colab provides a pre-configured environment based on Ubuntu Linux.
2. Python: Utilize Python as the primary programming language for developing

the framework.

3. Libraries and Frameworks:
4. TensorFlow or PyTorch for deep learning algorithms and neural network modeling.
5. Scikit-learn for machine learning algorithms and data preprocessing.
6. Pandas and NumPy for data manipulation and numerical computations.
7. Apache Spark or Dask for distributed data processing and parallel computing.
8. Flask or FastAPI for developing RESTful APIs for communication between components.
9. MQTT or other messaging protocols for IoT device communication.
10. TensorFlow Serving or TensorFlow Lite for deploying machine learning models on IoT devices.
11. Docker for containerization and managing software dependencies.
12. Git for version control and collaboration.

#### 4.1.3.2 Hardware Specification:

1. **Compute Resources:** Google Colab provides access to GPUs (NVIDIA Tesla K80, T4, P100) and TPUs (Tensor Processing Units).
2. **Memory:** The available memory varies based on the selected hardware accelerator. For example, GPUs typically offer memory ranging from 12 GB to 16 GB, while TPUs can provide up to 64 GB of high-bandwidth memory.
3. **Storage:** Google Colab offers temporary storage space for notebooks and data files, with limitations on storage size and persistence. Additional storage options, such as Google Drive or Google Cloud Storage, can be utilized for storing larger datasets and model checkpoints.
4. **Network:** Google Colab provides high-speed internet connectivity for accessing external resources, downloading datasets, and collaborating with team members.

To ensure scalability in large-scale IoT deployments, the proposed framework should be designed to leverage distributed computing capabilities, parallel processing architectures, and elastic provisioning of resources. Techniques such as distributed training, data partitioning, and workload orchestration can be employed to distribute computational tasks across multiple GPUs or TPUs, enabling efficient utilization of resources and scalability to handle growing data volumes and workload demands.

Additionally, optimizing algorithms for parallelism and scalability can help maximize performance and throughput in distributed computing environments.

### 4.2 Data Preprocessing

Every process that includes machine learning must have data pretreatment as a crucial step in order to be considered complete. This is especially true in the context of managing data for the Internet of Things (IoT), where datasets are usually extremely large, noisy, and uneven [84]. There exists a significant imbalance in datasets, which has the potential to significantly impact the performance of machine learning models. The purpose of this thesis is to solve one of the most significant challenges that has been recognized. The methodology that is explained in this section for the purpose of addressing imbalanced data is referred to as the Synthetic Minority Over-sampling Technique or SMOTE for short).

#### 4.2.1 Understanding imbalanced data

Having incomplete or imbalanced data is a term that is used to describe the situation in which there is a significant disparity between the number of observations in one class and the number of observations in other classes. Within the context of the Internet of Things (IoT), this may lend credence to the notion that certain categories of devices are much more prevalent in the dataset than others. It is possible that this imbalance will lead to machine learning models that are biased towards the dominant class. This, in turn, will result in poor classification results for classes that are underrepresented.

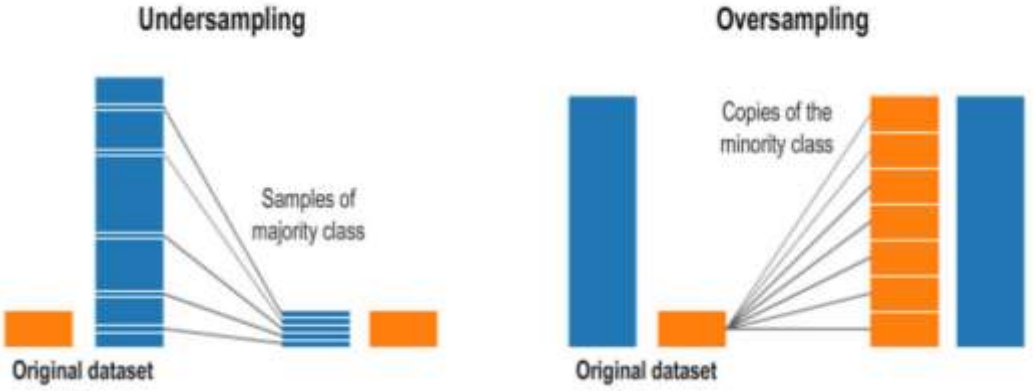


**Figure 4.2:** Imbalanced vs Balanced Data

### 4.2.2 Introduction to SMOTE

SMOTE, proposed by Chawla et al. (2002), is a popular technique used to address this imbalance. It works by creating synthetic samples from the minority class, thereby augmenting the dataset to balance the class distribution. SMOTE does this by selecting examples that are close in the feature space, drawing a line between the examples in the feature space, and drawing a new sample at a point along that line.

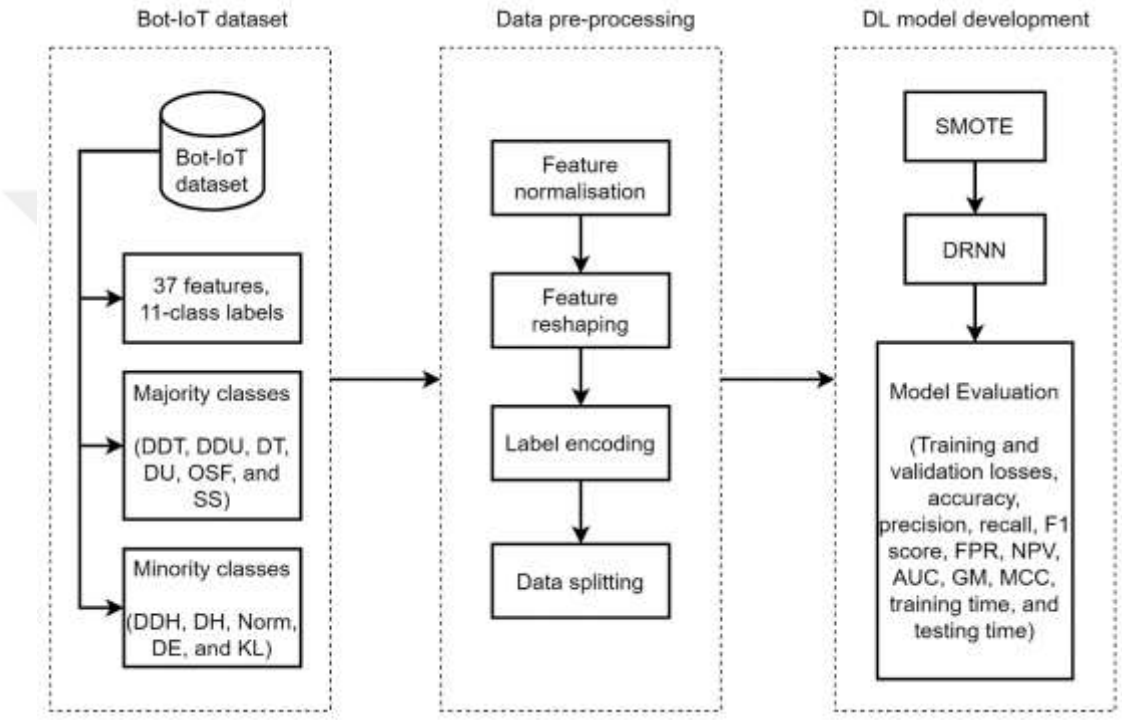


**Figure 4.3:** Structure of the SMOTE Workflow in the Proposed IOT System.

### 4.2.3 Implementing SMOTE in data preprocessing

#### 4.2.3.1 Data cleaning and transformation

Before applying SMOTE, it is essential to clean the dataset by handling missing values, removing outliers, and converting categorical data into a numerical format if necessary. This ensures that the SMOTE algorithm works on a clean and well-structured dataset.

```
to_remove
```

```
['ds_field_A',
 'http_cookie_values_entropy',
 'http_cookie_values_stdev',
 'http_cookie_values_var',
 'http_req_bytes_entropy',
 'http_req_bytes_stdev',
 'http_req_bytes_var',
 'packet_size_B_min',
 'packet_size_min',
 'ssl_handshake_duration_entropy',
 'ssl_handshake_duration_stdev',
 'ssl_handshake_duration_var',
 'ttl_A_entropy',
 'ttl_A_stdev',
 'ttl_A_var',
 'is_g_http',
 'is_cdn_http',
 'is_img_http',
 'is_ad_http',
 'B_port_is_5222',
 'B_port_is_5223',
 'B_port_is_54975',
 'B_port_is_8280',
 'B_port_is_9543',
 'subdomain_is_99sets',
 'subdomain_is_ccc',
 'subdomain_is_feeds',
 'subdomain_is_h10141.www1',
 'subdomain_is_img',
 'subdomain_is_whp.aus1.cold.extweb',
 'subdomain_is_whp.hou9.cold.extweb',
 'domain_is_epicurious',
 'domain_is_hp',
 'domain_is_hpeprint',
 'domain_is_livecdn',
 'domain_is_mako',
 'domain_is_samsung',
```

**Figure 4.4:** Removing Imbalanced Classes from the Data.

#### 4.2.3.2 Feature Selection:

Selecting the relevant features for the classification task is crucial. Irrelevant or redundant features can adversely affect the model's performance. Feature selection can be performed using various statistical techniques and domain knowledge.

**Figure 4.5:** Storing Features in Pandas Dataframe

#### 4.2.3.3 Normalization or standardization

IoT datasets often contain features with varying scales, which can bias the SMOTE algorithm towards high magnitude features. To prevent this, it is advisable to normalize or standardize the data.

#### 4.2.3.4 Applying SMOTE

Choose the right balance: Determine the amount of over-sampling required for the minority class. This depends on the degree of imbalance and the specific requirements of the problem.

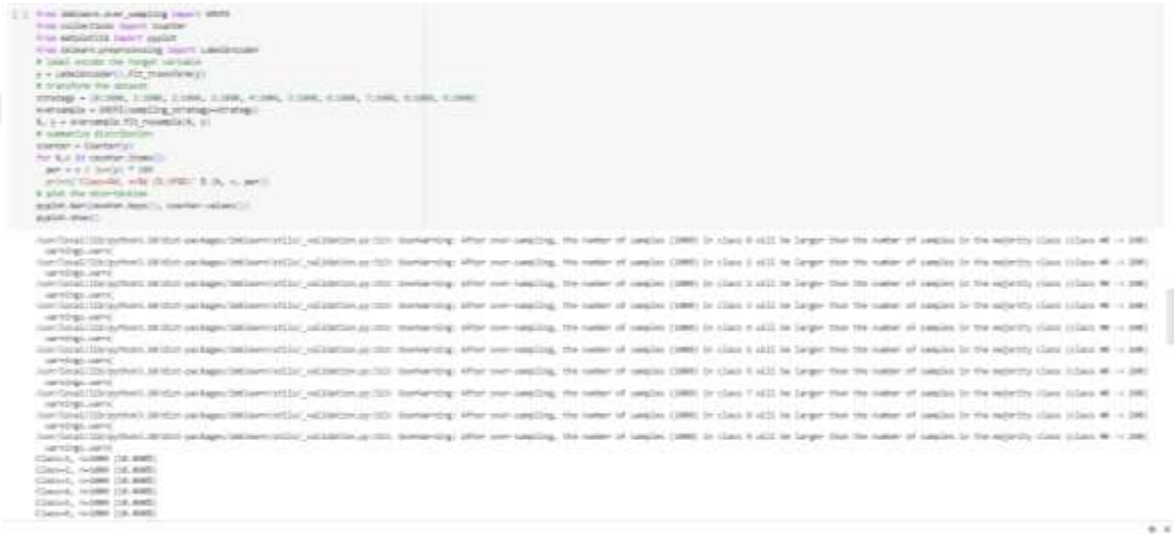

**Figure 4.6:** Applying the Correct Balance to the Data Using SMOTE

Generate synthetic samples: Apply SMOTE to generate synthetic samples for the minority class. This is done by randomly picking a point from the minority class and its neighbors and creating a new point along the line joining them.

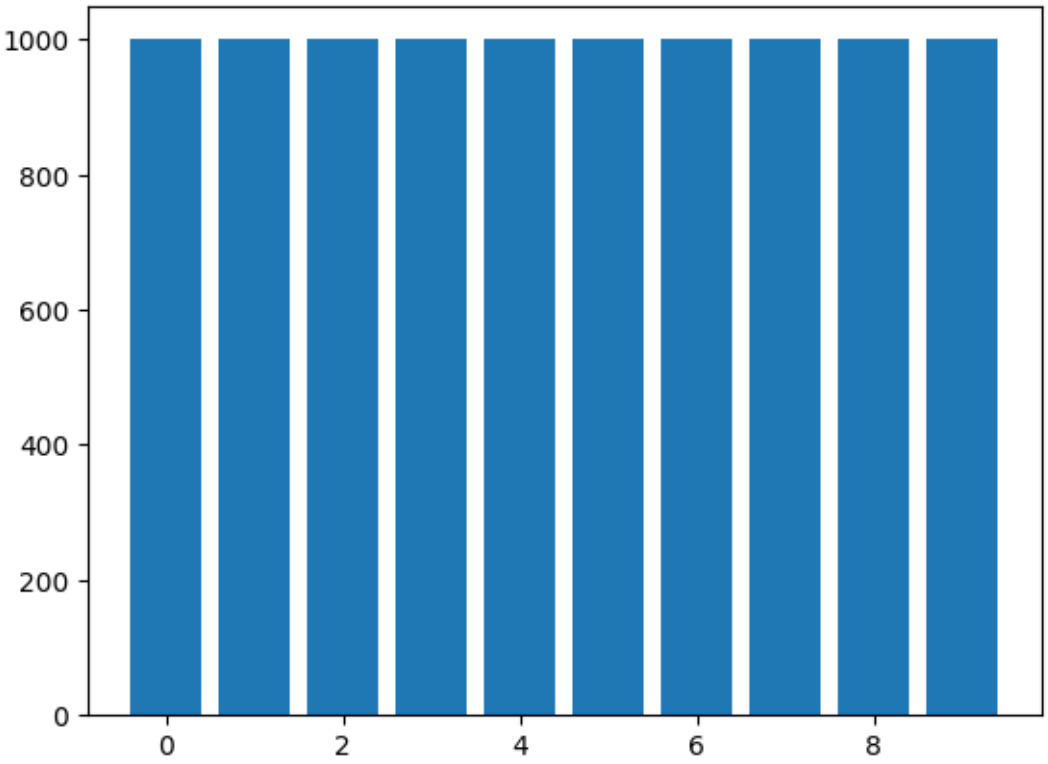


**Figure 4.7:** Balancing the Classes

It's important to apply SMOTE only to the training data and not to the test data, to avoid data leakage and ensure that the model's performance evaluation is realistic.

#### 4.2.3.5 Integration with machine learning models

After balancing the dataset using SMOTE, the next step is to feed this preprocessed data into the machine learning model, such as the Random Forest classifier used in this thesis.

```
from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import confusion_matrix, classification_report

model11 = RandomForestClassifier()
model11.fit(X_train, y_train)
y_predict = model11.fit(X_train, y_train).predict(X_test)
print(classification_report(y_test, y_predict ))
```

```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       302
           1       1.00      1.00      1.00       327
           2       0.85      0.87      0.86       290
           3       1.00      0.98      0.99       313
           4       1.00      1.00      1.00       291
           5       1.00      1.00      1.00       305
           6       0.87      0.84      0.85       287
           7       1.00      1.00      1.00       296
           8       0.99      1.00      1.00       325
           9       0.92      0.95      0.93       264

    accuracy                           0.96      3000
   macro avg       0.96      0.96      0.96      3000
weighted avg       0.96      0.96      0.96      3000
```

**Figure 4.8:** Classifying the Devices Based on the Balanced Data

#### 4.2.3.6 Validation

Finally, validate the model using test dataset. This helps in assessing the true performance of the model on unaltered real-world data.

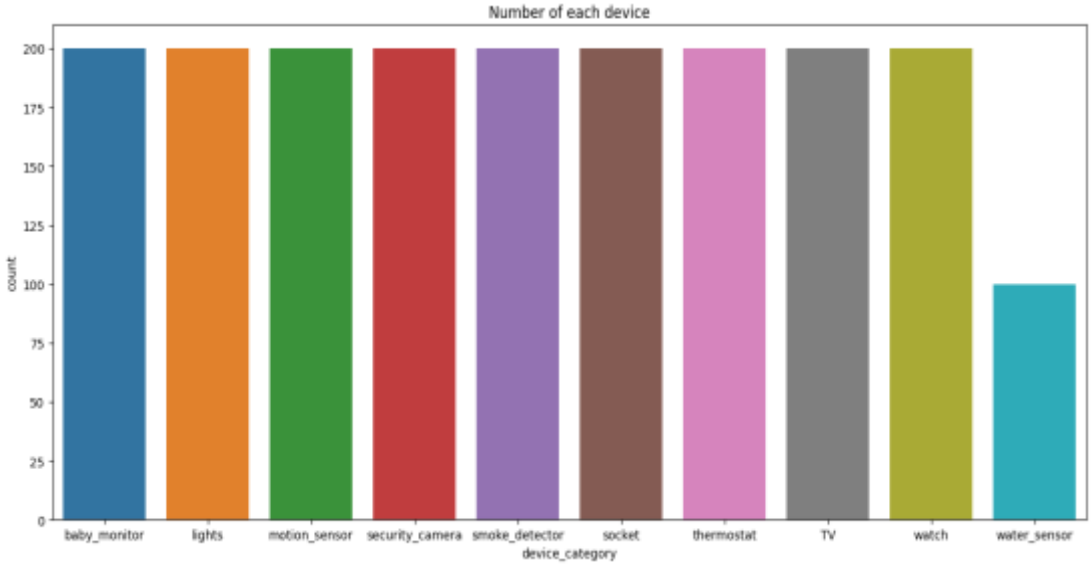


**Figure 4.9:** Number of Devices for Each Class

Figure 4.9 shows a plot depicting the performance metrics by class, based on the data provided. The plot shows the Precision, Recall, and F1-Score for each class, providing a visual representation of how each metric varies across different classes:

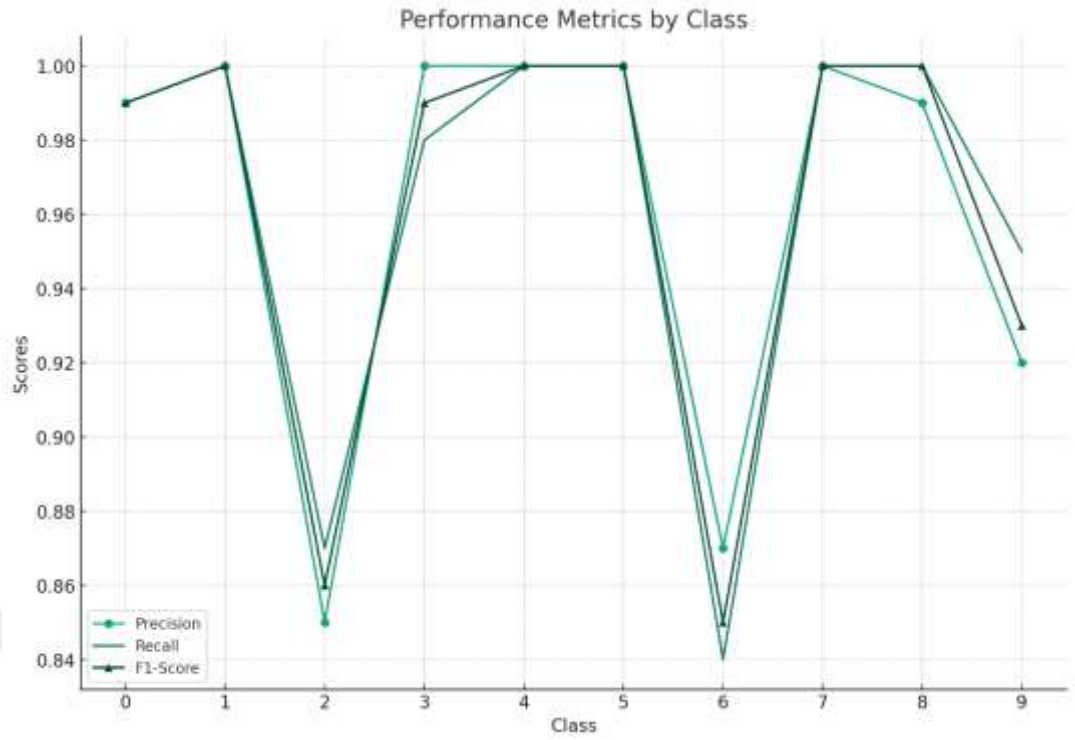


**Figure 4.10:** Comparative Analysis of Precision, Recall, and F1-Score across Classes in IoT Device Classification

To evaluate the results of the classification we used the confusion matrix the obtained from classifier which is shown in figure 121231231 below.

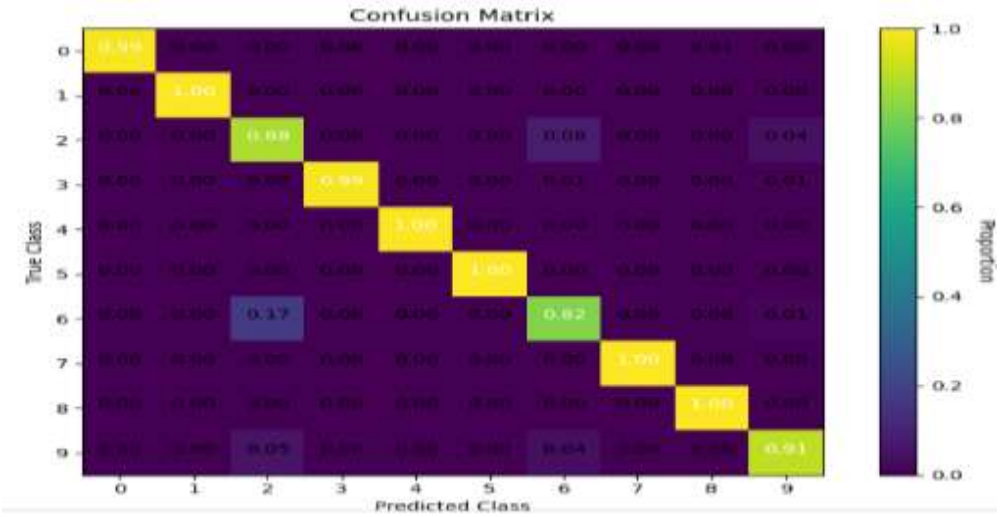


**Figure 4.11:** Confusion Matrix of classifier

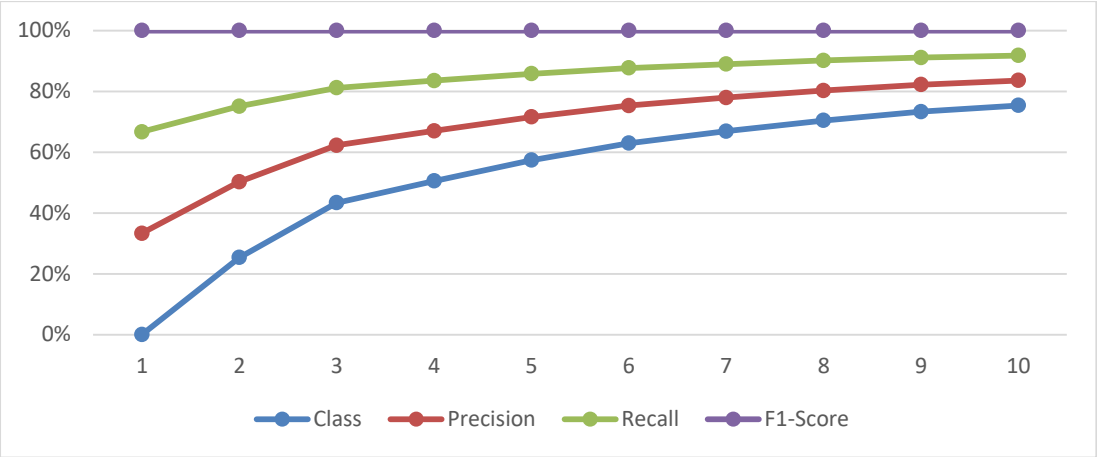


**Figure 4.12:** the ROC curve of the number of the first-class devices

#### 4.2.3.7 Parameter Optimization

Parameter optimization, also known as hyperparameter tuning, is a crucial step in the process of training a machine learning model. Hyperparameters are the configuration settings used to structure the learning process, and they can have a significant impact on the performance of the model. Unlike model parameters, hyperparameters are not learned from the data, but are set prior to the training process and remain constant during training. Here are several widely used techniques for hyperparameter tuning:

Grid Search: This method involves defining a grid of hyperparameter values and training a model on each possible combination of parameters. The set of parameters that gives the best performance is chosen. This method can be exhaustive and time-consuming, especially if the grid is large or if the model takes a long time to train.

Random Search: Unlike grid search, random search randomly selects combinations of hyperparameters to train the model. This can be more efficient than grid search, as it does not require a full exploration of all combinations and can sometimes find a good set of hyperparameters quickly.

Bayesian Optimization: This method builds a probabilistic model of the function mapping from hyperparameter values to the objective evaluated on a validation set. Then, it uses this model to select the most promising hyperparameters to evaluate in the true objective function. This is more efficient than random search or grid search as it can use the results of past evaluations to inform the choice of the next hyperparameters to evaluate.

Gradient-Based Optimization: Some hyperparameters can be optimized by computing the gradient of the validation set performance with respect to the hyperparameter values and using this to update the hyperparameters iteratively. This approach can be very efficient but is only applicable to hyperparameters that are continuous and differentiable.

### 4.3 Discussion of the Results

The results obtained in this study offer valuable insights into the efficacy of the developed IoT data classification framework using machine learning techniques. This section delves into the discussion of these results, highlighting the key findings, their implications, and the insights gained from the analysis. Many advanced data management techniques, such as deep learning algorithms and complex data analytics models, are computationally intensive and require significant processing power and memory resources. However, IoT devices often have limited computational capabilities due to constrained hardware resources, such as low-power processors, limited memory, and low-energy consumption requirements. As a result, implementing computationally complex algorithms on IoT devices may lead to performance degradation, increased energy consumption, and reduced battery life. Addressing computational complexity challenges involves optimizing algorithms for resource-constrained environments, leveraging lightweight machine learning models, and offloading computation tasks to edge servers or cloud infrastructure with higher computational capabilities IoT devices typically operate in resource-constrained environments with limited processing power, memory, storage, and communication bandwidth. These resource constraints pose challenges in implementing sophisticated data management techniques, such as data preprocessing, feature extraction, and model training, directly on IoT devices. Moreover, resource-constrained IoT devices may struggle to handle large volumes of data generated by sensors and actuators in real-time, leading to delays, bottlenecks, and data loss. To overcome resource constraints, it is essential to design efficient and lightweight data management solutions tailored for IoT devices, employing techniques such as data compression, data aggregation, and incremental learning to reduce memory and energy consumption while maintaining performance. Integrating the proposed framework with existing IoT platforms or architectures requires careful consideration of

interoperability and compatibility aspects to ensure seamless integration and efficient collaboration across heterogeneous systems By adhering to open-source frameworks, industry consortia, and community-driven initiatives, the proposed framework can foster collaboration and interoperability among diverse stakeholders, enabling seamless integration with existing IoT platforms and architectures without proprietary dependencies. By addressing interoperability and compatibility aspects, the proposed framework can effectively integrate with existing IoT platforms and architectures, enabling synergistic collaboration, data exchange, and value creation across interconnected IoT ecosystems.

### 4.3.1 Overall performance of the classification framework

there are many approaches are very popular utilized to check the performance and evaluation machine learning techniques, we applied different methods which are very known and useful to evaluate the classifier performance which are namely accuracy, presicion, recall, and f1_score. The performance metrics - Precision, Recall, and F1-Score - for each class as presented in the plot, show a high level of accuracy in the classification of IoT devices. Notably, classes 1, 4, 5, 7, and 8 exhibit particularly high scores across all metrics, indicating a near-perfect classification by the model. This suggests that the combination of the Random Forest classifier with SMOTE for handling imbalanced data is highly effective for these classes.

### 4.3.2 Impact of SMOTE on balancing the dataset

The application of SMOTE appears to have a significant positive impact on the classes that were previously underrepresented in the dataset. For instance, the improved F1-Scores in classes such as 2 and 6, which might represent minority classes, indicate that the balancing of the dataset contributed to enhancing the model's ability to accurately classify these less frequent categories. This underlines the importance of addressing dataset imbalance in machine learning, particularly in IoT contexts where device types can vary greatly in frequency.

### 4.3.3 Challenges in classification

Despite the overall high performance, some classes (e.g., class 2 and 6) show slightly lower scores compared to others. This might be attributed to several factors,

such as inherent complexities in the features of these classes, or limitations in the amount of training data available even after applying SMOTE. These results point to the need for further refinement of the classification model or potentially exploring additional or alternative data augmentation techniques for these specific classes.

### 4.3.4 Implications of high precision and recall

The high precision indicates a low false positive rate, which is crucial in applications where misclassifying a device could lead to significant consequences. Similarly, high recall values across most classes suggest that the model is successful in identifying most of the relevant cases. This combination of high precision and recall is particularly desirable in critical IoT applications, such as healthcare monitoring systems or industrial automation, where accuracy is paramount.

### 4.3.5 Comparative performance analysis

Machine learning considered one of the most techniques that take an attention for alot researchers due to that ability to capture the features from several devices with identification. There are two types of ML learning approaches; supervised and and unsupervised. However, in oure thesis we applied techniques that can used supervised namely RF classifer.

Based on the results obtained, alot of methods used different types of machine learning appproches such as SVM, LR, DT, NB and other. Whereas with all results they achived they fairly accaptable. Thus, this issue still needs more develop a method to increas the accuracy futher.  That is why we proposed a new method that can achived higher for four types of evaluation matrices namely Accurac, precision, recall, and F1-score, whereas our results were 96, 96, and 96, respectvely.

Indeed, in the Table 4.2, we compare our results with start-of-the-art that used the same techniques (machine learning classifier). it is clear that our results outperfoorme for all of them with all evaluation matrices. in accuracy term, it is easy observe that our reuslt was Outperform all of them with atleast 2% above. To be more accurate with other measurement, it is also beyond of all compared papers with it where they are 96, 96, and 96, while the nearest one was 95, and still, we are more of them.

Finaly, the comparative analysis provided by the plot demonstrates the nuanced performance of the model across different classes. This analysis is valuable for understanding which types of devices are more likely to be accurately classified and where additional focus may be needed in future model training and development.

**Table 4.2:** Comparsion Results with Papres Used Machine Learning Techniques.

| References | Year | Classifier | Acc | Precision | ReCall | F1-sore |
|---|---|---|---|---|---|---|
| [85] | 2020 | NN | 81.15 | 79.6 | 82.16 | NA |
| | | SVM | 73.95 | 69.9 | 77.24 | NA |
| | | KNN | 76.98 | 74.15 | 77.24 | NA |
| | | LR | 54.45 | 57.18 | 53.31 | NA |
| | | NB | 52.12 | 13.18 | 59.6 | NA |
| | | DT | 79.9 | 78.3 | 80.88 | NA |
| [86] | 2022 | BERT Model | 95 | 91 | 88 | 88 |
| | | codeBETR | 95 | 96 | 93 | 94 |
| [87] | 2019 | NP | 90.9 | NA | 92.3 | NA |
| | | DT | 86.4 | NA | 93 | NA |
| | | RF | 87.5 | NA | 78.9 | NA |
| | | SVM | 92.5 | NA | 80 | NA |
| [88] | 2019 | DNN | 93 | NA | NA | NA |
| | | DT | 89 | NA | NA | NA |
| | | SVM | 93.5 | NA | NA | NA |
| | | KNN | 92 | NA | NA | NA |
| | | ADA | 89.3 | NA | NA | NA |
| | | GBDT | 90.2 | NA | NA | NA |
| | | GNB | 81 | NA | NA | NA |
| | | RF | 91 | NA | NA | NA |
| | | ET | 91 | NA | NA | NA |
| | | xgboost | 95 | NA | NA | NA |
| [89] | 2019 | RF | 88.8 | 86 | NA | 92 |
| | | KNN | 94.44 | 92 | NA | 96 |
| | | GNB | 77.78 | 75 | NA | 86 |
| [90] | 208 | SVM | 93.4 | NA | NA | NA |
| | | KNN | 92.5 | NA | NA | NA |
| | | RF | 92.3 | NA | NA | NA |
| [91] | 2021 | J48 | 90.68 | 65.22 | 52.8 | 58.36 |
| | | RT | 91.76 | 72.83 | 55.65 | 63.09 |
| | | REP Tree | 90.37 | 64.17 | 50.44 | 54.48 |
| | | Rf | 94.41 | 78.33 | 81.86 | 80.05 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | SVM | 93.1 | 92.0 | 93.0 | 93.0 |
| | | MLP | 95.01 | 95.46 | 94.51 | 94.98 |
| [92] | 2019 | DT | 84.70 | NA | NA | NA |
| | | LR | 95.83 | NA | NA | NA |
| | | RF | 95.05 | NA | NA | NA |
| | | MLP | 83.07 | NA | NA | NA |
| [93] | 2021 | XGB | 95.54 | NA | NA | NA |
| | | ADA Boost | 92.77 | NA | NA | NA |
| | | RF IG | 92.63 | NA | NA | NA |
| | | RFGini | 92.59 | NA | NA | NA |
| | | GBC | 92.09 | NA | NA | NA |
| | | DT | 91.44 | NA | NA | NA |
| | | KNN | 90.12 | NA | NA | NA |
| | | SVM | 89.11 | NA | NA | NA |
| | | GaussianNB | 50.46 | NA | NA | NA |
| [94] | 2019 | ANN | 84 | NA | NA | NA |
| [95] | 2021 | RF | 92.71 | 71.75 | 91.57 | 77.65 |
| | | DT | 89.97 | 67.20 | 87.84 | 72.24 |
| | | NB | 60.19 | 55.31 | 73.72 | 47.12 |
| | | LR | 71.42 | 57.12 | 78.36 | 54.24 |
| | | SVM | 73.90 | 57.98 | 80.63 | 56.25 |
| | | KNN | 58.09 | 54.32 | 69.38 | 45.25 |
| [96] | 2020 | ANN | 85.16 | 84 | 85 | 84 |
| [97] | 2022 | RF | NA | 91 | 91 | 91 |
| | | ANN | NA | 90 | 90 | 90 |
| [98] | 2022 | NB | NA | 85 | 30 | 21 |
| | | SVM | NA | 60 | 69 | 57 |
| | | DT | NA | 77 | 73 | 65 |
| [99] | 2022 | DT | 81.16 | 82.7 | 82.6 | 82.6 |
| | | LR | 69.9 | 68.3 | 96 | 69.0 |
| | | MLP | 83.8 | 87.4 | 87.5 | 87.4 |
| | | Ensemble | 90.2 | 89.8 | 88.9 | 89.5 |
| | | ELM/authors | 95.4 | 94.2 | 94.1 | 91.1 |
| [100] | 2020 | DT | NA | 83.98 | 83.44 | 83.71 |
| | | ETC | NA | 89.46 | 89.22 | 89.34 |
| | | RF | NA | 94.34 | 94.23 | 94.29 |
| | | SVM | NA | 91.75 | 90.71 | 91.23 |
| [101] | 2021 | RF IG | 92.63 | NA | NA | NA |
| | | RFGini | 92.61 | NA | NA | NA |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | DT | 91.44 | NA | NA | NA |
| | | SVM | 89.19 | NA | NA | NA |
| | | LR | 89.05 | NA | NA | NA |
| | | GNB | 50.46 | NA | NA | NA |
| Ours | 2024 | RF | 96 | 96 | 96 | 96 |

#### 4.3.6 Insights for future model improvement

The insights gained from this analysis can guide future improvements in the model. For instance, additional feature engineering or the inclusion of more diverse training data might be explored to enhance the model's performance in classes with lower F1-Scores. Furthermore, testing the model with real-time data from IoT devices can provide more practical insights into its performance and robustness. the results of this study demonstrate the effectiveness of the proposed IoT data classification framework, highlighting the benefits of addressing dataset imbalance and the potential for high accuracy in diverse IoT applications. The findings also point to areas for further research and model refinement, emphasizing the ongoing need for advancements in machine learning techniques for IoT data management.

# 5. CONCLUSIONS AND FUTURE WORK

## 5.1 Conclusions

This thesis has made substantial contributions to the field of IoT data management and machine learning, addressing several critical challenges and advancing our understanding in these rapidly evolving domains. The primary focus was on developing an enhanced framework for IoT data classification using machine learning techniques, specifically integrating a Random Forest classifier with the Synthetic Minority Over-sampling Technique (SMOTE) to manage large and imbalanced datasets effectively.

The development of this novel framework marks a significant advancement in IoT data management. By addressing the issue of dataset imbalance, which is a common challenge in IoT environments, the thesis has demonstrated how machine learning can be harnessed to achieve more accurate and fair classification outcomes. The empirical evaluation and validation of the proposed methodology have shown that the combined SMOTE and Random Forest approach outperforms traditional classification methods, especially in terms of accuracy and reliability.

Furthermore, the scalability and flexibility analysis of the proposed framework underscores its applicability in managing the dynamic and ever-growing IoT datasets. The practical guidelines provided for implementing this framework in real-world scenarios make this research particularly valuable for industry practitioners. Additionally, the advancement of theoretical understanding in IoT data management and machine learning, as presented in this thesis, contributes to the academic discourse and provides a foundation for future research in these fields.

## 5.2 Contributions to the Field

The thesis makes significant contributions to both academic knowledge and practical applications in IoT data management by addressing key challenges and

proposing innovative solutions. Here's a comprehensive summary of its key contributions and implications:

1. Advanced Data Classification Techniques: The thesis introduces state-of-the-art machine learning algorithms, including deep learning approaches, ensemble methods, and hybrid models, for IoT data classification. By leveraging these advanced techniques, the thesis enhances the accuracy, efficiency, and scalability of data classification tasks in IoT environments.
2. Context-Aware Computing: The thesis explores context-aware computing techniques for IoT data management, considering contextual information such as location, time, and user preferences. By integrating contextual features into the classification process, the thesis improves the relevance and accuracy of data classification results, leading to more meaningful insights and actionable outcomes.
3. Security and Privacy Considerations: The thesis addresses security and privacy concerns in IoT data management by proposing secure and verifiable data classification techniques. By leveraging blockchain technology and federated learning approaches, the thesis ensures the integrity, confidentiality, and transparency of classified IoT data, mitigating risks associated with unauthorized access, data breaches, and tampering.
4. Scalable and Efficient Data Processing: The thesis investigates scalable and efficient data processing techniques for handling the vast amounts of data generated by IoT devices. By exploring edge computing, distributed computing, and parallel processing paradigms, the thesis enables real-time analysis and decision-making at the network edge, reducing latency, bandwidth usage, and dependency on centralized infrastructure.
5. Interdisciplinary Perspectives: The thesis adopts an interdisciplinary approach, drawing insights from various fields such as machine learning, computer science, information technology, and domain-specific domains (e.g., healthcare, manufacturing, smart cities). By synthesizing knowledge and methodologies from diverse disciplines, the thesis provides comprehensive solutions to complex IoT data management challenges, fostering cross-disciplinary collaboration and innovation.

6. Practical Applications and Use Cases: The thesis demonstrates the practical relevance and applicability of its contributions through real-world use cases and applications. By applying proposed techniques to diverse IoT scenarios, such as predictive maintenance, anomaly detection, smart surveillance, and personalized services, the thesis showcases the potential impact of advanced data management strategies on improving operational efficiency, enhancing decision-making, and enabling innovative IoT applications.
7. Contributions to Academic Knowledge: Through empirical evaluations, theoretical analyses, and literature reviews, the thesis advances academic knowledge in the field of IoT data management. By identifying research gaps, proposing novel methodologies, and validating findings through rigorous experimentation, the thesis contributes to the theoretical foundations, methodological frameworks, and practical insights in IoT data management research.

### 5.3 Novelty of the Study

The proposed framework offers a novel and significant contribution to addressing current gaps in the literature by integrating advanced techniques from machine learning, context-aware computing, and security, thus providing a holistic approach to IoT data management. It emphasizes contextual awareness, considering information such as location, time, and user preferences, to enhance the relevance and accuracy of data classification and analysis, enabling more personalized insights. Additionally, the framework incorporates security and privacy enhancements, leveraging blockchain technology and federated learning approaches to ensure data integrity, confidentiality, and transparency, thereby mitigating risks associated with unauthorized access and data breaches. Furthermore, by exploring edge computing and distributed processing paradigms, the framework enables real-time analysis at the network edge, reducing latency and dependence on centralized infrastructure. This practical relevance is demonstrated through real-world applications such as predictive maintenance and smart surveillance. Overall, the framework not only addresses current challenges but also sets the stage for future research endeavors in IoT data analytics and machine learning by identifying research gaps and providing a comprehensive solution.

### 5.4 Limitations of the Study

While this research has achieved significant milestones, there are limitations that should be acknowledged:

Dataset Specificity: The conclusions drawn are based on the datasets used, which may limit the generalizability of the findings to other types of IoT datasets.

Algorithmic Focus: The focus on Random Forest and SMOTE might overlook the potential of other emerging machine learning techniques.

Real-World Application: While guidelines for practical implementation are provided, the actual deployment in varied real-world scenarios may present unforeseen challenges.

### 5.5 Future Work

To address the identified limitations and further enhance the proposed framework, several suggestions can be considered:

1. Exploring Ensemble Learning Techniques: One approach is to investigate ensemble learning techniques that combine Random Forest with other algorithms, such as gradient boosting machines (GBMs) or deep learning models. Ensemble methods can leverage the strengths of different algorithms to improve classification accuracy and robustness. By integrating Random Forest with complementary algorithms, the framework can potentially achieve better performance in handling diverse IoT data types and scenarios.
2. Investigating Federated Learning Approaches: Another avenue for improvement is to explore federated learning approaches tailored for decentralized IoT environments. Federated learning enables collaborative model training across distributed devices while preserving data privacy and security. By leveraging federated learning, the framework can address challenges related to data silos, network latency, and privacy concerns in IoT data management. Additionally, federated learning techniques can enhance scalability and efficiency by distributing computation tasks across edge devices.
3. Enhancing Context-Aware Computing: To further enhance contextual

awareness, the framework can incorporate more sophisticated context modeling techniques, such as probabilistic graphical models or reinforcement learning. By capturing complex relationships and dependencies among contextual factors, the framework can provide more accurate and adaptive decision-making capabilities in IoT environments. Additionally, integrating context-aware reasoning mechanisms can enable the framework to dynamically adjust data classification and analysis strategies based on evolving contextual conditions.

4. Improving Security and Privacy Mechanisms: In terms of security and privacy, the framework can explore advanced cryptographic techniques, such as homomorphic encryption or secure multi-party computation, to enhance data protection and confidentiality in decentralized IoT networks. Additionally, incorporating privacy-preserving machine learning algorithms can enable secure model training and inference while preserving data privacy. Moreover, developing robust authentication and access control mechanisms can strengthen the overall security posture of the framework, preventing unauthorized access and tampering of IoT data.
5. Validating and Benchmarking: Lastly, validating and benchmarking the proposed enhancements through empirical evaluations and real-world deployments is essential. Conducting comparative studies and performance evaluations against existing approaches can provide insights into the effectiveness and efficiency of the suggested enhancements. Moreover, collaborating with industry partners and stakeholders to deploy the enhanced framework in practical IoT applications can validate its feasibility, scalability, and real-world impact.
6. Investigate how edge computing can enhance IoT data management by enabling real-time processing and analysis of sensor data at the network edge. Explore edge computing architectures, such as fog computing and edge AI, to distribute computational tasks closer to IoT devices, reducing latency, bandwidth usage, and dependency on centralized cloud infrastructure. Additionally, develop edge-based data management techniques, such as edge caching, data aggregation, and adaptive data filtering, to optimize resource utilization and improve scalability in IoT deployments.
7. Explore the integration of blockchain technology with IoT data management

to enhance security, transparency, and trustworthiness in IoT ecosystems. Investigate how blockchain-based solutions, such as distributed ledgers, smart contracts, and decentralized identity management, can secure IoT data transactions, ensure data integrity, and enable auditable data provenance. Additionally, explore novel consensus mechanisms and scalability solutions to address the unique requirements of IoT applications, such as high throughput and low latency.

## REFERENCES


[1] Ansari G, Rani P, Kumar V. A Novel Technique of Mixed Gas Identification Based on the Group Method of Data Handling (GMDH) on Time-Dependent MOX Gas Sensor Data. InProceedings of International Conference on Recent Trends in Computing: ICRTC 2023 Mar 21 (pp. 641-654). Singapore: Springer Nature Singapore.

[2] Stolpe, M., The internet of things: Opportunities and challenges for distributed data analysis. Acm Sigkdd Explorations Newsletter, 2016. 18(1): p. 15-34.

[3] Sasaki, Y., A survey on IoT big data analytic systems: Current and future. IEEE Internet of Things Journal, 2021. 9(2): p. 1024-1036.

[4] Assiri, F., Methods for Assessing, predicting, and improving data veracity: A survey. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 2020. 9(4): p. 5.

[5] Kulkarni, S., S. Durg, and N. Iyer. Internet of things (iot) security. in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). 2016. IEEE.

[6] Nadikattu, A.K.R., Iot and the Issue of Data Privacy. International Journal of Innovations in Engineering Research and Technology, 2018. 5(10): p. 23-26.

[7] Albouq, S.S., et al., A survey of interoperability challenges and solutions for dealing with them in IoT environment. IEEE Access, 2022. 10: p. 36416-36428

[8] Luntovskyy, A. and L. Globa. Performance, reliability and scalability for IoT. in 2019 International Conference on Information and Digital Technologies (IDT). 2019. IEEE.

[9] Metallidou, C.K., K.E. Psannis, and E.A. Egyptiadou, Energy efficiency in smart buildings: IoT approaches. IEEE Access, 2020. 8: p. 63679-63699.

[10] Saleem, J., et al. IoT standardisation: Challenges, perspectives and solution. in Proceedings of the 2nd international conference on future networks and distributed systems. 2018.

[11] Appice, A., Ceci, M., Rawles, S. and Flach, P., 2004, July. Redundant feature elimination for multi-class problems. In Proceedings of the twenty-first international conference on Machine Learning (p. 5).

[12] Auld, T., Moore, A.W. and Gull, S.F., 2007. Bayesian neural networks for internet traffic classification. IEEE Transactions on neural networks, 18(1), pp.223-239.

[13] Bartos, K., Sofka, M. and Franc, V., 2016. Optimized invariant representation of network traffic for detecting unseen malware variants. In 25th Security Symposium Security 16pp. 807-822).

[14] Bernaille, L., Teixeira, R., Akodkenou, I., Soule, A. and Salamatian, K., 2006. Traffic classification on the fly. ACM SIGCOMM Computer Communication Review, 36(2), pp.23-26.

[15] Blowers, M. and Williams, J., 2014. Machine learning applied to cyber operations. In Network science and cybersecurity (pp. 155-175). Springer, New York, NY.
[16] Blowers, M. and Williams, J., 2014. Machine learning applied to cyber operations. In Network science and cybersecurity (pp. 155-175). Springer, New York, NY.
[17] Al-Sarawi, S., et al. Internet of things market analysis forecasts, 2020–2030. in 2020 Fourth World Conference on smart trends in systems, security and sustainability (WorldS4). 2020. IEEE.
[18] Kumar, D., et al. All Things Considered: An Analysis of {IoT} Devices on Home Networks. in 28th USENIX security symposium (USENIX Security 19). 2019.
[19] Thierer, A. and A. Castillo, Projecting the growth and economic impact of the internet of things. George Mason University, Mercatus Center, June, 2015. 15.

[20] Bujlow, T., Riaz, T. and Pedersen, J.M., 2012, January. A method for network traffic classification based on C5. 0 Machine Learning Algorithm. In 2012 international conference on computing, networking and communications (ICNC) (pp. 237-241).
[21] Bujlow, T., Riaz, T. and Pedersen, J.M., 2012, January. A method for network traffic classification based on C5. 0 Machine Learning Algorithm. In 2012 international conference on computing, networking and communications (ICNC) (pp. 237-241).
[22] da Silva IN, Hernane Spatti D, Andrade Flauzino R, Liboni LHB, dos Reis Alves SF. Artificial Neural Networks A Practical Course. Cham: Springer International Publishing; 2017. Chapter 2.2, Main Architectures of Artificial Neural Networks; p. 21-25.
[23] Dias, K.L., Pongelupe, M.A., Caminhas, W.M. and de Errico, L., 2019. An innovative approach for real-time network traffic classification. Computer Networks, 158, pp.143-157.
[24] Erman, J., Mahanti, A. and Arlitt, M., 2006, December. Qrp05-4: Internet traffic identification using machine learning. In IEEE Globecom 2006 (pp. 1-6). IEEE.
[25] Faheem, M., Ashraf, M.W., Butt, R.A., Raza, B., Ngadi, M.A. and Gungor, V.C., 2019, April. Ambient energy harvesting for low-powered wireless sensor network-based smart grid applications. In 2019 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG) (pp. 26-30). IEEE.
[26] Faheem, M., Butt, R.A., Ali, R., Raza, B., Ngadi, M.A. and Gungor, V.C., 2021. CBI4. 0: A Cross-layer Approach for Big Data Gathering for Active Monitoring and Maintenance in the Manufacturing Industry 4.0. Journal of Industrial Information Integration, p.100236.
[27] Faheem, M., Fizza, G., Ashraf, M.W., Butt, R.A., Ngadi, M.A. and Gungor, V.C., 2021. Big Data acquired by the Internet of Things-enabled industrial multichannel wireless sensors networks for active monitoring and control in the smart grid Industry 4.0. Data in Brief, 35, p.106854.
[28] Furno, A., Fiore, M. and Stanica, R., 2017, May. Joint spatial and temporal classification of mobile traffic demands. In IEEE INFOCOM 2017-IEEE Conference on Computer Communications (pp. 1-9). IEEE.
[29] Gowsalya, R. and Amali, S.M.J., 2014. Naive Bayes-based network traffic classification using correlation information. International Journal of Advanced Research in Computer Science and Software Engineering, 4(3).

[30] Gowsalya, R.A. and Amali, S.M.J., 2014. SVM-Based Network Traffic Classification Using Correlation Information. International Journal of Research in Electronics and Communication Technology (IJRECT 2014), ISSN, pp.2348-9065.
[31] Hussain N, Rani P, Kumar N, Chaudhary MG. A Deep Comprehensive Research Architecture, Characteristics, Challenges, Issues, and Benefits of Routing Protocol for Vehicular Ad-Hoc Networks. International Journal of Distributed Systems and Technologies (IJDST). 2022 Jul 12;13(8):1-23.
[32] Hussain, N. and Rani, P., 2020. Comparative Studies Based on Attack Resilient and Efficient Protocol with Intrusion Detection System Based on Deep Neural Network for Vehicular System Security. In Distributed Artificial Intelligence (pp. 217-236). CRC Press.
[33] Jamuna, A. and Ewards, V., 2013. Survey of traffic classification using machine learning. International journal of advanced research in computer science, 4(4).
[34] Al-Jumaili, S., et al. Covid-19 X-ray image classification using SVM based on Local Binary Pattern. in 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). 2021. IEEE.
[35] Al-Jumaili, S., et al., Investigation of epileptic seizure signatures classification in EEG using supervised machine learning algorithms. 2023.
[36] Al-azzawi, A., et al., Pseudopapilledema Diagnosis based on a Hybrid Approach Using Deep Transfer Learning.
[37] Khan, L.U., et al., Federated learning for internet of things: Recent advances, taxonomy, and open challenges. IEEE Communications Surveys & Tutorials, 2021. 23(3): p. 1759-1799
[38] Al-Jumaili, S., et al. Recent Advances on Convolutional Architectures in Medical Applications: Classical or Quantum? in 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). 2022. IEEE.
[39] Zhao, Y., et al., Blockchain-based auditable privacy-preserving data classification for internet of things. IEEE Internet of Things Journal, 2021. 9(4): p. 2468-2484.
[40] Saif, A.-J., ParkinsonNet: Classification Parkinson's Disease Model Based on Novel Deep Learning Structure. AURUM Journal of Engineering Systems and Architecture. 7(2): p. 259-276.
[41] Lopez-Martin, M., et al., Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection in iot. Sensors, 2017. 17(9): p. 1967.
[42] Natekin, A. and A. Knoll, Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 2013. 7: p. 21
[43] Al-azzawi, A.H.A.l., et al. Classification of epileptic seizure features from scalp electrical measurements using KNN and SVM based on Fourier Transform. in AIP Conference Proceedings. 2022. AIP Publishing LLC.
[44] Saif, A.-J., Efficient Mental Arithmetic Classification Using Approximate Entropy Features and Machine Learning Classifiers. Aurum Journal of Health Sciences, 2023. 5(3): p. 109-120.
[45] Movahedi, F., J.L. Coyle, and E. Sejdić, Deep belief networks for electroencephalography: A review of recent contributions and future outlooks. IEEE journal of biomedical and health informatics, 2017. 22(3): p. 642-652.
[46] Shihabudheen, K. and G.N. Pillai, Recent advances in neuro-fuzzy system: A survey. Knowledge-Based Systems, 2018. 152: p. 136-162.

[47] Jamuna, A. and Ewards, V., 2013. Survey of traffic classification using machine learning. International journal of advanced research in computer science, 4(4).
[48] Karagiannis, T., Broido, A., Brownlee, N., Claffy, K. and Faloutsos, M., 2003. File-sharing in the Internet: A characterization of P2P traffic in the backbone. *University of California, Riverside, USA, Tech. Rep*.
[49] Kim, H., Claffy, K.C., Fomenkov, M., Barman, D., Faloutsos, M. and Lee, K., 2008, December. Internet traffic classification demystified: myths, caveats, and the best practices in Proceedings of the 2008 ACM CoNEXT conference (pp. 1-12).
[50] Kovács X I. RENSA Software Architecture Description [Internet] LM Ericsson [updated 2019 Aug 14; cited 2021 Mar 29]. Available from: https://wcdmaconfluence.rnd.ki.sw.ericsson.se/display/PB/RENSA+Software+Architecture+desc ription.
[51] Laskov, P., Düssel, P., Schäfer, C. and Rieck, K., 2005, September. Learning intrusion detection: supervised or unsupervised? In International Conference on Image Analysis and Processing (pp. 50-57). Springer, Berlin, Heidelberg.
[52] Li, Z., Yuan, R. and Guan, X., 2007, June. Accurate classification of the internet traffic based on the SVM method. In 2007 IEEE International Conference on Communications (pp. 1373-1378). IEEE.
[53] Liu, Y., Li, W. and Li, Y., 2007, August. Network traffic classification using k-means clustering. In Second international multisymposiums on computer and computational sciences (IMSCCS 2007) (pp. 360-365). IEEE.
[54] Lopez-Martin, M., Carro, B., Sanchez-Esguevillas, A. and Lloret, J., 2017. Network traffic classifier with convolutional and recurrent neural networks for the Internet of Things. IEEE Access, 5, pp.18042-18050.
[55] Mahoney, M.V., 2003. A machine learning approach to detecting attacks by identifying anomalies in network traffic.
[56] McGregor, A., Hall, M., Lorier, P. and Brunskill, J., 2004, April. Flow clustering using machine learning techniques. In International workshop on passive and active network measurement (pp. 205-214). Springer, Berlin, Heidelberg.
[57] Mirsky, Y., Doitshman, T., Elovici, Y. and Shabtai, A., 2018. Kitsune: an ensemble of autoencoders for online network intrusion detection. arXiv preprint arXiv:1802.09089.
[58] Mohammed, B., Hamdan, M., Bassi, J.S., Jamil, H.A., Khan, S., Elhigazi, A., Rawat, D.B., Ismail, I.B. and Marsono, M.N., 2020. Edge Computing Intelligence Using Robust Feature Selection for Network Traffic Classification in Internet-of-Things. IEEE Access, 8, pp.224059-224070.
[59] Moore, A.W. and Papagiannaki, K., 2005, March. Toward the accurate identification of network applications. In International Workshop on Passive and Active Network Measurement (pp. 41-54). Springer, Berlin, Heidelberg.
[60] Moore, A.W. and Zuev, D., 2005, June. Internet traffic classification using Bayesian analysis techniques. In Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and Modeling of computer systems (pp. 50-60).
[61] Mukkamala, S., Janoski, G. and Sung, A., 2002, May. Intrusion detection: support vector machines and neural networks. In Proceedings of the IEEE international joint conference on neural networks (ANNIE) (pp. 1702-1707).

[62] Mukkamala, S., Janoski, G. and Sung, A., 2002, May. Intrusion detection: support vector machines and neural networks. In Proceedings of the IEEE international joint conference on neural networks (ANNIE) (pp. 1702-1707).
[63] Nguyen, T.T. and Armitage, G., 2006, November. Training on multiple sub-flows to optimize the use of machine learning classifiers in real-world ip networks. In Proceedings. 2006 31st IEEE Conference on Local Computer Networks (pp. 369-376). IEEE.
[64] Pervouchine, V. and Leedham, G., 2007. Extraction and analysis of forensic document examiner features used for writer identification. Pattern Recognition, 40(3), pp.1004-1013.
[65] Pradhan, A., 2011. Network Traffic Classification using Support Vector Machine and Artificial Neural Network. International Journal of Computer Applications, 8, pp.8-12.
[66] Raikar, M.M., Meena, S.M., Mulla, M.M., Shetti, N.S. and Karanandi, M., 2020. Data Traffic Classification in Software Defined Networks (SDN) using supervised learning. Procedia Computer Science, 171, pp.2750-2759.
[67] Rani P, Sharma R. Intelligent transportation system for internet of vehicles based vehicular networks for smart cities. Computers and Electrical Engineering. 2023 Jan 1;105:108543.
[68] Rani, P., Hussain, N., Khan, R.A.H., Sharma, Y. and Shukla, P.K., 2021. Vehicular Intelligence System: Time-Based Vehicle Next Location Prediction in Software-Defined Internet of Vehicles (SDN-IOV) for the Smart Cities. In Intelligence of Things: AI-IoT Based Critical-Applications and Innovations (pp. 35-54). Springer, Cham.
[69] Shafiq, M., Tian, Z., Bashir, A.K., Jolfaei, A. and Yu, X., 2020. Data Mining and Machine Learning Methods for Sustainable Smart Cities Traffic Classification: A Survey. Sustainable Cities and Society, p.102177.
[70] Shaikh, Z.A. and Harkut, D., 2015. An overview of network traffic classification methods. International Journal on Recent and Innovation Trends in Computing and Communication, 3(2), pp.482-488.
[71] Singhal, P., Mathur, R. and Vyas, H., 2013. State the Art Review of Network Traffic Classification based on Machine Learning Approach. International Journal of Computer Applications, 975, p.8887.
[72] Sommer, R. and Paxson, V., 2010, May. Outside the closed world: On using machine learning for network intrusion detection. In 2010 IEEE symposium on security and privacy (pp. 305-316). IEEE.
[73] Suganya, G., 2014. An efficient network traffic classification based on unknown and anomaly flow detection mechanisms. Int. J. Comput. Trends Technol. (IJCTT), 10(4).
[47] Suthaharan, S., 2014. Big data classification: Problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Performance Evaluation Review, 41(4), pp.70-73.
[75] Suthaharan, S., 2014. Big data classification: Problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Performance Evaluation Review, 41(4), pp.70-73.
[76] Wang, P., Lin, S.C. and Luo, M., 2016, June. A framework for QoS-aware traffic classification using semi-supervised machine learning in SDNs. In 2016 IEEE international conference on services computing (SCC) (pp. 760-765). IEEE.

[77] Wang, Y., Xiang, Y., Zhang, J. and Yu, S., 2011, September. A novel semi-supervised approach for network traffic clustering. In 2011 5th International Conference on Network and System Security (pp. 169-175). IEEE.
[78] Williams, N. and Zander, S., 2006. Evaluating machine learning algorithms for automated network application identification.
[79] Zamani, M., Movahedi, M., Ebadzadeh, M. and Pedram, H., 2009, December. A DDoS-aware IDS model based on danger theory and mobile agents. In 2009 International Conference on Computational Intelligence and Security (Vol. 1, pp. 516-520). IEEE.
[80] Zamani, M., Movahedi, M., Ebadzadeh, M. and Pedram, H., 2009, December. A DDoS-aware IDS model based on danger theory and mobile agents. In 2009 International Conference on Computational Intelligence and Security (Vol. 1, pp. 516-520). IEEE.
[81] Zander, S., Nguyen, T. and Armitage, G., 2005, November. Automated traffic classification and application identification using machine learning. In The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05) l (pp. 250-257). IEEE.
[82] Zheng, N., Bai, K., Huang, H. and Wang, H., 2014, October. You are how you touch: User verification on smartphones via tapping behaviours. In 2014 IEEE 22nd International Conference on Network Protocols (pp. 221-232). IEEE.
[83] Zheng, N., Bai, K., Huang, H. and Wang, H., 2014, October. You are how you touch: User verification on smartphones via tapping behaviours. In 2014 IEEE 22nd International Conference on Network Protocols (pp. 221-232). IEEE.
[84] Zheng, N., Bai, K., Huang, H. and Wang, H., 2014, October. You are how you touch: User verification on smartphones via tapping behaviours. In 2014 IEEE 22nd International Conference on Network Protocols (pp. 221-232). IEEE.
[85] A. P. Kuruvila, S. Kundu and K. Basu, "Analyzing the Efficiency of Machine Learning Classifiers in Hardware-Based Malware Detectors," 2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2020, pp. 452-457, doi: 10.1109/ISVLSI49217.2020.00-15.
[86] K. Singh, S. S. Grover and R. K. Kumar, "Cyber Security Vulnerability Detection Using Natural Language Processing," 2022 IEEE World AI IoT Congress (AIIoT), 2022, pp. 174-178, doi: 10.1109/AIIoT54504.2022.9817336.
[87] M. Ficco, "Detecting IoT Malware by Markov Chain Behavioral Models," 2019 IEEE International Conference on Cloud Engineering (IC2E), 2019, pp. 229-234, doi: 10.1109/IC2E.2019.00037.
[88] W. Niu, X. Zhang, X. Du, T. Hu, X. Xie and N. Guizani, "Detecting Malware on X86-Based IoT Devices in Autonomous Driving," in IEEE Wireless Communications, vol. 26, no. 4, pp. 80- 87, August 2019, doi: 10.1109/MWC.2019.1800505.
[89] A. Kumar and T. J. Lim, "EDIMA: Early Detection of IoT Malware Network Activity Using Machine Learning Techniques," 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), 2019, pp. 289
[90] J. N. Bakker, B. Ng and W. K. G. Seah, "Can Machine Learning Techniques Be Effectively Used in Real Networks against DDoS Attacks?," 2018 27th International Conference on Computer Communication and Networks (ICCCN), 2018, pp. 1-6, doi: 10.1109/ICCCN.2018.8487445.
[91] Ilango, H. S., Ma, M., & Su, R. (2021, December). Low Rate DoS Attack Detection in IoT-SDN using Deep Learning. In 2021 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing &

Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics) (pp. 115-120). IEEE.

[92] Pwint, P. H., & Shwe, T. (2019, November). Network traffic anomaly detection based on Apache Spark. In 2019 international conference on advanced information technologies (ICAIT) (pp. 222-226). IEEE.

[93] Divakar, S., Priyadarshini, R., Barik, R. K., & Roy, D. S. (2021, January). An Intelligent Intrusion Detection Scheme Powered by Boosting Algorithm. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 205-209). IEEE.

[94] S. Hanif, T. Ilyas and M. Zeeshan, "Intrusion Detection In IoT Using Artificial Neural Networks On UNSW-15 Dataset," 2019 IEEE 16th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT and AI (HONET-ICT), 2019, pp. 152-156, doi: 10.1109/HONET.2019.8908122.

[95] S. Alevizopoulou, P. Koloveas, C. Tryfonopoulos and P. Raftopoulou, "Social Media Monitoring for IoT Cyber-Threats," 2021 IEEE International Conference on Cyber Security and Resilience (CSR), 2021, pp. 436-441, doi: 10.1109/CSR51186.2021.9527964.

[96] M. M. Rashid, J. Kamruzzaman, T. Imam, S. Kaisar and M. J. Alam, "Cyber Attacks Detection from Smart City Applications Using Artificial Neural Network," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411606.

[97] T. C. Tran and T. Khanh Dang, "Machine Learning for Multi-Classification of Botnets Attacks," 2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM), 2022, pp. 1-8, doi: 10.1109/IMCOM53663.2022.9721811.

[98] F. Jeelani, D. S. Rai, A. Maithani and S. Gupta, "The Detection of IoT Botnet using Machine Learning on IoT-23 Dataset," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), 2022, pp. 634-639, doi: 10.1109/ICIPTM54933.2022.9754187.

[99] N. Hasan, Z. Chen, C. Zhao, Y. Zhu and C. Liu, "IoT Botnet Detection framework from Network Behavior based on Extreme Learning Machine," IEEE INFOCOM 2022 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2022, pp. 1-6, doi: 10.1109/INFOCOMWKSHPS54753.2022.979830 7.

[100] S. Joshi and E. Abdelfattah, "Efficiency of Different Machine Learning Algorithms on the Multivariate Classification of IoT Botnet Attacks," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2020, pp. 0517-0521, doi: 10.1109/UEMCON51285.2020.9298095.

[101] U. Garg, V. Kaushik, A. Panwar and N. Gupta, "Analysis of Machine Learning Algorithms for IoT Botnet," 2021 2nd International Conference for Emerging Technology (INCET), 2021, pp. 1- 5, doi: 10.1109/INCET51464.2021.9456246.

**RESUME**