Crowd Counting with Wi-Fi Probe Requests: A Selective Information Elements-based Approach Supported by Generative Data Augmentation

Mohamed Chaaben

A Thesis

in

The Department

of

Concordia Institute of Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Quality Systems Engineering) at

Concordia University

Montréal, Québec, Canada

December 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mohamed Chaaben**

Entitled: **Crowd Counting with Wi-Fi Probe Requests: A Selective Information Elements-based Approach Supported by Generative Data Augmentation**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Quality Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Abdessamad Ben Hamza

_____ External Examiner
Dr. Jeremy Clark

_____ Supervisor
Dr. Nizar Bouguila

_____ Co-supervisor
Dr. Zachary Patterson

Approved by _____
Chun Wang, Chair
Department of Concordia Institute of Information Systems Engineering

_____ 2024

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Crowd Counting with Wi-Fi Probe Requests: A Selective Information Elements-based Approach Supported by Generative Data Augmentation

Mohamed Chaaben

Crowd monitoring is essential for smart city applications, particularly for optimizing public transit systems. To address this need, we propose a privacy-conscious crowd-counting pipeline using Wi-Fi probe requests. This pipeline is designed to adapt to the challenges posed by the randomization of Media Access Control (MAC) addresses, which serve as unique identifiers for devices on a network. Our approach leverages a random forest-based feature selection process to identify key Information Elements and frame attributes then applies DBSCAN clustering with adaptive parameter optimization for device counting. A diffusion model generates synthetic tabular data to mitigate the limited availability of labelled data, enhancing model robustness. Experimental results demonstrate improved accuracy in device counting, achieving a V-measure of 0.952, an average silhouette score of 0.789, and reliable clustering counts.

Acknowledgments

I would like to express my sincere gratitude to my supervisors, Professor Zachary Patterson and Professor Nizar Bouguila, for their insightful guidance and steady support throughout this journey.

My gratitude naturally extends to my family, whose education, trust, and support allowed me to flourish, both academically and personally.

My appreciation also goes to the team at BusPas Inc. for their collaboration, with special thanks to my professional mentor, Dr. Wissem Maazoun, for his valuable insights and expertise.

I am equally grateful to my lab mates, especially Asiye Baghbani and Siavash Farazmand, for their ongoing encouragement and practical advice along the way. This accomplishment is, in no small part, a reflection of their generous contributions.

To all who have been part of this journey, thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This introductory chapter sets the stage for the study by providing its context and highlighting the key gaps that motivated our work. It then defines the objectives of the research and elaborates on the contributions made. The chapter concludes with a concise outline of the thesis structure.

1.1 Problem Statement

Crowd monitoring has become increasingly important for local administrators seeking to enhance city services and create safer and more responsive urban environments. Nowadays, its applications reach far and wide, covering a range of different sectors. For instance, in intelligent buildings, monitoring systems are employed to enhance energy efficiency by adjusting lighting, ventilation, and heating according to real-time occupancy data, contributing to more sustainable management practices (Agarwal et al., 2010; Zou et al., 2018). In commercial settings like shopping malls and restaurants, customer queues and seating are monitored in order to allow for dynamic staffing adjustments, improving both service delivery and operational efficiency (Y. Wang et al., 2014). In safety-critical environments like stadiums, monitoring crowd density and movement, along with behavioural responses, is essential for anticipating risks of congestion or sudden surges (Dong et al., 2023). Such advances not only elevate safety but also enrich the shared experiences of those gathered.

To support these varied applications, computer vision-based solutions have proven effective (Junior et al., 2010), yet their real-world implementation often comes with high costs and privacy concerns. Indeed, significant expenses can arise from the need for extensive camera networks and the infrastructure to support them, especially when complex deep learning models are in use, requiring substantial power and processing resources. These technologies also facilitate the tracking of individual identities and locations, pushing against regulatory frameworks such as the European General Data Protection Regulation (European Parliament & Council of the European Union, 2016) and Quebec's Law 25 (Éditeur officiel du Québec, 2024). Additionally, practical challenges posed by variable lighting, viewpoint shifts, and occlusions further complicate deployment.

As smartphones have become ubiquitous, with 90% of adults in the United States owning one as of 2023 (Pew Research Center, 2024), Wi-Fi-based crowd monitoring has emerged as an alternative to vision-based solutions. For this purpose, it can rely on Wi-Fi device capabilities to broadcast probe requests.

Probe requests are management frames emitted by devices that perform active scanning to discover nearby Wi-Fi networks that are available for connection. This exchange is part of the larger communication process—illustrated in Figure 1.1—that helps devices and networks establish a handshake before connecting.

Assuming a one-to-one mapping between passengers and devices, catching probe requests has allowed for counting people and analyzing their permanence and return times. This approach proved effective until manufacturers implemented MAC address randomization for security and privacy reasons (Fenske et al., 2021). Rather than transmitting the real physical MAC address in probe requests, modern devices now send randomized virtual addresses at irregular intervals. Consequently, probe requests from a Wi-Fi device no longer retain a static MAC address but instead cycle pseudo-periodically through different random addresses. The adoption of MAC address randomization has been sporadic and varies across manufacturers and operating systems, making device monitoring more challenging still.

To work around MAC address randomization, probe requests are clustered based on the assumption that those from the same device exhibit enough intrinsic similarities to be grouped together. Various studies have investigated distinct features to assess this similarity. Some have focused on

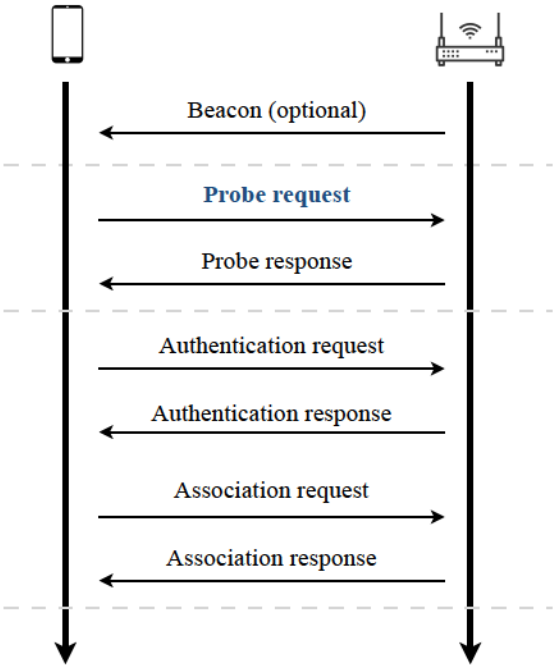


Figure 1.1: Illustration of the handshake process between Wi-Fi devices and an access point

patterns in the Received Signal Strength Indicator (RSSI), while others have examined timestamps or sequential numbering of probe request captures. However, RSSI and time-based approaches remain highly sensitive to signal path variations and can be easily distorted by obstacles, reflective surfaces, and interference. As for the sequence number, it is no longer reliable since it also is randomized.

In crowd monitoring, traditional methods typically depend on syntactic features that capture basic details, such as timestamps, RSSI values, or sequence numbers. The challenge, however, is to discern the semantic relationships within Wi-Fi probe requests. Semantic features, such as Information Elements (IEs), hold the potential to reveal behavioural context and user intent, offering valuable insights into crowd patterns—particularly when data fields have been randomized or obscured. IEs within probe requests have been proposed as an alternative feature, and several studies have leveraged the length or content of all available IEs to enhance device differentiation. Despite this, interpretability and generalizability issues remain problematic. The variability in IE usage—since they are optional and differ across devices—adds complexity to analysis, constraining consistency and weakening model robustness.

To improve crowd size estimation, scholars have commonly combined multiple features, such as MAC addresses, RSSI values, and IEs, rather than relying on only one feature. Despite this, challenges persist in verifying the exact number of devices due to MAC address randomization, which complicates efforts to confirm whether multiple probe requests originate from the same device. Some studies have attempted validation through manual counts or filtering based on RSSI, but these approaches lack precision. Consequently, many studies have relied on multiple Wi-Fi access points to enhance accuracy, though this approach adds complexity to the setup.

1.2 Contributions

In this study, we examine the use of a select subset of Information Elements (IEs) as features for crowd counting, focusing specifically on estimating passenger numbers at bus stops. This subset was identified through a random forest-based feature selection analysis, which demonstrated that certain IEs carry more significance than others, reducing the need to include all IE fields as previously done by Vanhoef et al. (2016). Our findings indicate that accurate counting is achievable using only four features: the probe request length and three specific IE fields. Additionally, we note the diminishing reliability of sequential numbering, likely due to randomization. We also introduce the use of "presence flags", a novel feature that, to our knowledge, has not been previously used.

Moreover, our study introduces a unique clustering metric by employing the Hamming distance rather than the commonly used Euclidean distance for probe request analysis, a choice that improves counting accuracy in our specific context.

The limited availability of probe request data presents an additional challenge to accurate crowd counting, prompting us to investigate generative models, specifically diffusion models and Generative Adversarial Networks (GANs), for data augmentation. By simulating realistic probe requests, this approach seeks to enhance model performance. Of particular interest is the fact that diffusion models and GANs remain underexplored for tabular data, which is the type of data used in our study. To our knowledge, this is the first application of deep generative models for data augmentation in the context of probe requests.

It is noteworthy that this thesis on Wi-Fi passenger counting at bus stops is an integral part of a project in collaboration with BusPas Inc. (2024), a company specializing in smart city solutions. For context, it has developed the "SCiNe", a smart device designed for installation at bus stops, equipped with dual cameras, a passive infrared sensor, a display screen, a Wi-Fi card, and other smart features. Our project is intended to be implemented within the SCiNe to enable counting functionality. Beyond the motivations identified in the literature, this research is closely aligned with the company's specific need for effective crowd-counting solutions. In fact, using Wi-Fi in particular offers unique advantages, as the SCiNe is solar-powered and operates with limited energy resources. Wi-Fi-based crowd-counting can be selectively activated during low battery conditions or when weather affects the reliability of computer vision. Moreover, because the device operates outdoors with a single Wi-Fi card, using Wi-Fi probe requests is more practical than alternative Wi-Fi options such as Channel State Information (CSI).

Finally, it is also worth highlighting that some of the literature on MAC address randomization has focused on MAC address "derandomization". Derandomization involves identifying a device's actual MAC address by defeating randomization techniques designed to conceal its true identity. However, in this work, our aim is neither to perform derandomization nor to identify individual devices, which ensures a privacy-preserving and responsible solution aligned with the legal and ethical standards.

1.3 Thesis Overview

The rest of the thesis is structured as follows:

- Chapter 2 offers a comprehensive review of Wi-Fi techniques for crowd monitoring, presenting a novel taxonomy that differs from the conventional framework seen in most Wi-Fi monitoring studies, which typically focus on passive versus active scanning. This chapter also examines relevant datasets and delves into deep learning methods for generating tabular data, which is the type of data used in this context.

- Chapter 3 outlines the methodology used to develop the pipeline, starting with an explanation of the feature selection process based on random forest techniques. This is followed by a

description of the density-based clustering approach and concludes with a detailed account of the diffusion-based method for data generation.

- Chapter 4 details the structure and composition of the dataset, examines feature distributions, and presents the results along with the evaluation metrics used to assess model performance.
- Chapter 5 concludes the thesis with a summary of the findings, the limitations of the solution, and a discussion of potential future research directions.

Chapter 2

Literature Review

This literature review is organized into four sections. The first section provides an overview of crowd-monitoring technologies that do not rely on Wi-Fi. The second section examines Wi-Fi-based crowd-monitoring methods, with a particular emphasis on crowd-counting key approaches and methodologies. In the third section, we review the datasets used in the context of Wi-Fi probe requests, noting the significant limitations in their availability. Finally, the fourth section discusses the challenges of generating synthetic tabular data to address these dataset limitations and enable more robust model development.

2.1 Non-Wi-Fi-based Crowd Monitoring

In the dynamic landscape of crowd-counting technologies, vision-based methods have traditionally been at the forefront, leveraging either single or multiple camera setups, whether standard or thermal. For instance, thermal imaging solutions, like those proposed by Gade and Moeslund (2014), have proven essential in scenarios where visibility is compromised, enhancing the ability to monitor and manage crowds under various conditions.

In recent years, deep learning has become fundamental to these methods, with ongoing advancements significantly enhancing accuracy and improving crowd-counting performance in complex environments. Notably, context-aware crowd-counting has emerged as a pioneering approach that dynamically adapts to environmental factors, allowing for more precise density estimations by

tackling challenges such as scale variation and occlusion (Liu et al., 2019). Another frontier in this domain is marked by vision-language models, with state-of-the-art tools like CrowdCLIP (Liang et al., 2023) leading the way. By incorporating text as an auxiliary modality through unsupervised learning, these models enable refined crowd pattern recognition without requiring extensive labelled data, which greatly alleviates the constraints of dataset dependency. More recently, the introduction of transformer-based models, exemplified by VMambaCC (Ma et al., 2024), has pushed the boundaries even further. With the ability to capture long-range dependencies, these models excel in handling densely packed scenes with intricate spatial layouts, making them indispensable in real-world applications where high precision is paramount.

While vision-based models perform well, they still depend on labelled data, which are costly and difficult to gather. They also face notable challenges, as pointed out by Singh et al. (2021). The primary one is related to rising privacy concerns, making it challenging to deploy freely in everyday environments. On top of that, cameras encounter technical hurdles like line-of-sight obstructions, adverse weather, low light, and high contrast. To tackle these limitations, researchers are exploring alternative people-counting technologies, particularly those that make use of radio frequency (RF) signals. RF-based counting is significantly more likely to perform effectively in low-light, smoky, or dusty environments, such as during fires or earthquakes, not to mention its ability to count people without compromising privacy.

Across the radio frequency spectrum, technologies such as ultra-wideband radars (J. W. Choi et al., 2012), wireless sensor networks (Yuan et al., 2013), and Zigbee communications (Lim et al., 2015) have been employed to monitor and count individuals. Recently, attention has turned to LiDAR sensors, which, unlike RF technologies, use laser light to distinguish individuals based on temperature variations within scanned areas. Pioneering work, including that of (Hasan et al., 2022), highlights the considerable promise of LiDAR for this purpose, while also noting the challenges—such as high hardware costs and the complexity of managing these advanced sensors. Another avenue involves aggregated mobile phone data, which can provide time series data on the number of people within specific geographical cells (Calabrese et al., 2014). However, such data are highly centralized, often lack privacy safeguards, and require special access to proprietary datasets.

2.2 Wi-Fi-based Crowd Monitoring

Wi-Fi-based crowd monitoring refers to the use of Wi-Fi signals to observe and analyze the presence, movement, and behaviour of crowds. With the ubiquity of Wi-Fi-enabled devices, such as smartphones, this method becomes increasingly appealing, especially given the ease of leveraging already existing Wi-Fi infrastructure. Additionally, its nature as a relatively energy-efficient and cost-effective solution makes it attractive for crowd monitoring.

Device-free crowd monitoring

In this approach, crowd dynamics are captured without requiring individuals to carry any device. It works by analyzing how people's movements disturb the Wi-Fi signal. The captured signal disturbances are then processed to infer crowd behaviour. For example, Xi et al. (2014) have developed a device-free crowd-counting method based on Channel State Information[1].

Similarly, Li et al. (2015) proposed a method for indoor crowd-counting using fluctuations in Wi-Fi signal strength (RSSI). As the number of people increases, the RSSI values drop and fluctuate. A five-layer neural network is then used to model the relationship between these RSSI variations and the presence of people.

Device-free monitoring, however, faces a major issue of scalability, limiting its use mainly to indoor settings. In fact, as the number of people increases, interpreting Wi-Fi signal variations becomes more challenging, especially in dense environments. Another factor that limits scalability is the interference from other nearby devices, which introduces noise and disrupts the system's accuracy (J. Wang et al., 2018). Additionally, device-free monitoring reliance on infrastructure—requiring the deployment of multiple Wi-Fi access points—makes it a costly approach. Those are the very limitations that corroborate forgoing this technology for crowd counting on the device of Buspas Inc.

[1]Channel State Information describes how a wireless signal propagates from the transmitter to the receiver, capturing variations in amplitude and phase across different antennas.

Device-based crowd monitoring

Beyond device-free approaches, the literature reflects a substantial body of research on device-based crowd monitoring. The ubiquity of mobile devices has increasingly drawn attention to this method, offering more possibilities for capturing crowd dynamics. Device-based monitoring can be categorized into active-intervention methods versus free-intervention methods.

Active-intervention monitoring

This approach requires direct and proactive human interaction during the monitoring process, such as users connecting to a specific access point or installing a mobile application.

J.-g. Park et al. (2010) explored how Wi-Fi signals can be used for indoor localization by creating a database of signal strengths captured from nearby Wi-Fi access points. To improve accuracy, users contribute by sharing their location data through a mobile app. This allows the system to link specific Wi-Fi signal patterns—also called fingerprints—to precise locations, making the system more reliable. In addition, Chon et al. (2013) extended this concept by designing a more comprehensive crowd-sensing approach, where mobile devices collect large-scale Wi-Fi data alongside other contextual information like pictures. Users actively participated by capturing images and other environmental data to further enhance the system's accuracy.

Active-intervention monitoring has a few real-world applications yet, as users are typically reluctant or disinterested in installing mobile applications or connecting to a specific network for monitoring, especially without incentives or benefits. Privacy and data collection concerns make voluntary participation even less likely. Free-intervention monitoring could, accordingly, be more practical. It leverages existing Wi-Fi systems to collect data passively, requiring no active involvement from users as they go about their daily routines.

Free-intervention monitoring

In this approach, crowd dynamics are captured and processed without the need for individual involvement. In fact, Wi-Fi-enabled devices automatically and sporadically send out Wi-Fi probe requests "in the air". By capturing these signals, we can analyze crowd patterns without requiring

manual intervention. Assuming a one-to-one correlation between passengers and devices, catching probe requests has allowed for counting people and analyzing their permanence and return times.

In the work of Yaik et al. (2016), an estimation of crowd size is presented via counting Wi-Fi probe requests emitted by smartphones. The authors captured these probes and identified unique MAC addresses to deduce the number of people. Tested at a public event, the method has shown a high alignment with manual counting, proving its reliability in tracking crowd size.

In much the same way, Pattanusorn et al. (2016) estimated the number of passengers on shuttle buses at Thammasat University, Thailand. They used a Raspberry Pi microprocessor equipped with a Wi-Fi adapter to sniff probe requests, from which they extracted the RSSI and duration of transmission to filter out non-passengers. If a device's MAC address continued to appear while the bus was moving, they assumed the device belonged to a passenger.

The probe request-based approach was broadly successful until manufacturers introduced MAC address randomization for security and privacy reasons (Fenske et al., 2021). This shift was particularly driven by growing concerns about privacy and data protection. In the European Union, for example, MAC address randomization became unavoidable with the enforcement of the General Data Protection Regulation (GDPR), which mandates stringent protections for user-related data (European Parliament & Council of the European Union, 2016).

Rather than transmitting the actual physical MAC address in probe requests, modern devices now use randomized virtual addresses at irregular intervals. Consequently, probe requests from a single device no longer retain a static and universally unique MAC address but instead change pseudo-periodically to different random addresses. The adoption of MAC address randomization has been sporadic, varying across manufacturers and operating systems, and its implementation is undocumented, making device monitoring even more challenging. Hence, some researchers have sought to study Wi-Fi device behaviour in relation to MAC address randomization. For instance, Fenske et al. (2021) conducted an extensive analysis of various solutions to assess the extent of randomization, the conditions under which it is applied, and whether tracking vulnerabilities are effectively mitigated.

As traditional probe request-based tracking methods have become outdated, researchers have shifted their focus toward defeating MAC address randomization, often referred to as derandomization. Early efforts, such as the work by Matte et al. (2016), explored the use of timing patterns. In their approach, inter-frame arrival times of probe requests were analyzed to group frames originating from the same device, even when different virtual MAC addresses were being used. Their method uses several timing-based distance metrics to group these frames.

Relying on signal travel time as a distinguishing feature can be unreliable in real-world environments, where random delays often occur due to factors like signal reflections and interference from multiple paths. These effects introduce variations that make timing-based measurements less precise and harder to interpret accurately (Uras et al., 2020). Alternatively, some studies have turned to crowd monitoring by relying on the RSSI values embedded in probe requests. For example, Fuada et al. (2020) presents an RSSI-based system for indoor monitoring, estimating the distance between devices and Wi-Fi nodes to construe their locations, making it possible to track people. The study by Hong et al. (2018) focused on tracking crowd trajectories in a multi-level museum. The authors were gathering over 1.7 million probe requests to deduce visitor trajectories. A Hidden Markov model was used to model visitor movements, with MAC addresses (both randomized and non-randomized) and RSSI fingerprints as features.

While the signal strength is more commonly used in indoor environments, it also finds applications in outdoor settings. Guillen-Perez and Cano (2019) tried to count and locate pedestrians at traffic intersections. The focus is classifying pedestrians as either moving or stationary (waiting) and pinpointing the location of stationary pedestrians at intersections by analyzing the RSSI from the captured probe requests.

Relying on RSSI has been shown to provide limited accuracy, with the scientific literature providing convincing evidence that it shows low reliability, at least when used as the primary or sole feature. This is mainly due to its instability and susceptibility to path loss, fading, and interference (Heurtefeux & Valois, 2012). An illustrative example of RSSI's limitations is found in the work by Groba (2019), which aimed to count participants in public demonstrations using a distance filter based on RSSI. The solution could capture only a small fraction of the actual attendance, underscoring the failure of RSSI-based methods in real-world applications.

The literature then shifted towards exploring other features, such as sequence number, which initially appeared promising. The sequence number assigned to each Wi-Fi packet is an incremental value initially intended to ensure data transmission order. Cai et al. (2021) used it as a key feature to distinguish devices with randomized MAC addresses, with the assumption that the same device sends probe requests with consecutive or closely related sequence numbers. They grouped probe requests with consecutive sequence numbers, with the number of groups corresponding to the count of unique devices. However, the sequence number is no longer a very reliable feature, as it has been almost systematically randomized since 2018 for privacy purposes (Fenske et al., 2021).

In recent years, the literature has seen the emergence of new methods aimed at overcoming the challenges posed by MAC address randomization in Wi-Fi-based monitoring. One such approach involves leveraging Information Elements (IEs) embedded in Wi-Fi probe requests, which carry data about device capabilities. For instance, Vanhoef et al. (2016) introduced the concept of generating unique device signatures using IEs to track Wi-Fi devices. Their method, although focused on Wi-Fi security rather than crowd monitoring, successfully tracked up to 50% of devices within a 20-minute window. Interestingly, their approach also utilized the sequence number in probe requests, which then had not yet been randomized.

To improve the accuracy of crowd size estimation, scholars have also combined multiple features rather than relying on just one. For example, Vega-Barbas et al. (2021) developed a method that estimates crowd size via a footprinting mechanism that circumvents MAC address randomization. They generated unique identifiers by leveraging a combination of MAC addresses, RSSI, and some IEs. To test the method, they collected data from two scenarios: an outdoor event and an indoor concert. Likewise, Tan and Chan (2021) conducted a study in a shopping mall in Hong Kong. Their approach uses a flow network to model probe requests as nodes and optimize associations by finding minimum-cost flows between frames. For validation, real MAC addresses were used in some instances as ground truth, while high RSSI values were employed to infer the proximity of devices with randomized MAC addresses. However, their validation was limited in scope and carried out in a highly controlled setting, raising questions about the method's scalability. In another study, Pérez-Hernández et al. (2024) clustered probe requests based on IEs and RSSI to create unique device fingerprints. Their system, tested on a university campus with six access points, successfully

detected 181 devices with an accuracy rate of 92.3%.

People-counting using Wi-Fi probe requests under MAC address randomization comes with a notable challenge: model validation. The difficulty lies in confirming whether two probe requests come from the same device. To address this, researchers often rely on manual counting or apply coarse-grained assumptions like RSSI filtering, though these methods tend to be imprecise. As a result, many studies use multiple Wi-Fi access points to improve accuracy, which helps but also adds considerable complexity to the setup.

A more refined approach to address the challenge of model validation involves using anechoic chambers. An anechoic chamber is an isolated space, shielded from external interference and engineered to prevent echoes of electromagnetic waves. When a device is placed inside, we can be certain that all detected signals originate solely from that device in a direct line of sight to the sniffer. Therefore, if two probe requests are received, we can confidently conclude that they originate from the same device–the only device in the chamber. To explore the potential of anechoic chambers, Uras et al. (2020) developed an approach to fingerprint devices by combining the IDs and lengths of IEs–though not the full IE content. A set of 23 smartphones was individually placed in a semi-anechoic chamber, which generated 15,151 probes employing randomized MAC addresses. However, this dataset was not yet publicly available.

Fortunately, the release of a labelled dataset in 2022 has fuelled further research and provided a stronger foundation for model validation (Pintor & Atzori, 2022b). The dataset was created in a semi-anechoic chamber, and each smartphone's probe requests were categorized individually, addressing the scarcity of labelled data. In their study, Pintor and Atzori (2022a) used this dataset to analyze Wi-Fi probe requests and compared the potential of various IE fields to improve clustering outcomes. Uras et al. (2022) also used this dataset to de-randomize MAC addresses for crowd-counting purposes. They grouped probe requests originating from the same device by combining frame arrival times with IE lengths and RSSI. Their approach employs only the IE identifier and length, excluding their content. Additionally, they propose a method for detecting pseudo-random MAC addresses, which only change when the device's Wi-Fi is toggled on or off.

The same dataset was as well explored in the work by Simončič et al. (2023), where a method for detecting crowd presence and movement is proposed. This method essentially clusters probe

requests based on RSSI values, frames arrival time, and various IEs. The authors further validated their approach using probe requests collected in different environments: a semi-controlled rural setting, another dataset from the Jožef Stefan Institute, and an uncontrolled urban environment in Catania, Italy.

In the quest for the most potential features, our study extends beyond a simple comparison of IEs by examining the relevance of all promising options, including RSSI, sequence numbers, and probe request length. We aim to show that only a specific subset of IEs is genuinely necessary. We also employ a new feature, which we call "presence flags", with the goal of achieving greater accuracy.

2.3 Data for Wi-Fi Probe Requests

Before the introduction of MAC address randomization, device tracking and analysis were relatively straightforward, with little need for ground-truth datasets. This ease had arisen from the static nature of MAC addresses, which were consistently linked to individual devices. Two public datasets were used in the literature and had deemed sufficient for research requirements.

- **Sapienza Dataset**

  The Sapienza dataset (Barbera et al., 2013) is a publicly available resource widely used for studies involving probe requests. Collected during large-scale measurement campaigns, it has supported numerous related studies and includes data from five distinct environments: a university campus, a shopping mall, a train station, the Vatican City area, and political rallies.

- **Hasselt Dataset**

  The Hasselt dataset (Robyns et al., 2015) contains about 123,000 probe requests captured by eight monitoring stations at the Glimps 2015 music festival in Ghent, Belgium. Only a few studies have utilized this dataset, likely because the Sapienza dataset was widely available and deemed adequate for research at the time, and also due to specific limitations within the Hasselt dataset itself. In particular, the Hasselt dataset records only one probe request per unique MAC address, assigns a sequence number of zero to each request, and obscures the SSIDs. These constraints severely limit its potential for in-depth analysis.

As device addresses are no longer static, datasets like Sapienza and Hasselt have become outdated. This shift has led researchers to collect their own data, though only a few have made them publicly available. Nevertheless, these datasets are often limited in scope and scale because the data collection process is labour-intensive, challenging, and costly. As a result, they fail to meet the growing demand for more comprehensive data, which is particularly important for deep learning applications. The following two datasets were created and used after the widespread adoption of MAC address randomization

- **Nile dataset**

  This dataset (Abdulrahem, 2021) was collected at night in a shopping centre and is the only publicly available one without anonymization, making it ideal for analysis. However, the brief 40-minute capture window significantly limits the scope of the analysis.

- **IPIN dataset**

  The IPIN dataset (Bravenec et al., 2022) was released as supplementary material for a case study carried out in Lloret de Mar, Spain, during the 2021 Indoor Positioning and Indoor Navigation conference. It was collected over only four days, and the capture device could not store radio information.

Beyond the above-mentioned datasets, most others remain proprietary and inaccessible to the research community (Cunche et al., 2014; Matte et al., 2016; Vanhoef et al., 2016), creating a significant barrier to developing generalizable solutions. The challenge extends even further when it comes to model validation, as with MAC address randomization, accurately determining whether two probe requests originate from the same device becomes difficult without proper labelling. In uncontrolled environments, where interference from other Wi-Fi signals is common, reliably linking signals to specific devices becomes highly problematic.

Fortunately, the Pintor dataset, a labelled one created by Pintor and Atzori (2022b), can help address this matter. Indeed, the creators tested each device individually in an anechoic chamber, ensuring precise signal identification. The dataset comprises 22 smartphones, with only five devices lacking MAC address randomization implementation. It encompasses a total of 315 PCAP (Packet CAPture) files of probe request captures. Using a Raspberry Pi 3, each individual device was sniffed

either within an empty anechoic chamber or a noisy environment, concurrently using three non-overlapping channels (1, 6, and 11).

2.4 Synthetic Data for Crowd Monitoring

2.4.1 Challenges of tabular data generation

Considering the importance of a labelled dataset for model validation and the inherent challenges of collecting representative data, generating synthetic data may provide a practical and cost-effective solution to expand probe request datasets. The growing field of synthetic data generation could provide a promising approach by training machine learning models to accurately learn and replicate the statistical patterns embedded in original datasets, thus generating new representative samples. This process would not only enable more rigorous validation but also function as an effective data augmentation technique, enhancing model robustness and improving the performance of deep-learning applications during training.

That said, much of deep learning research remains largely centred around image data. Although images are high-dimensional and complex, they typically contain uniform and structured features, with pixel values arranged in grids that follow consistent patterns and distributions. In contrast, tabular data, like the kind we work with, has often been overlooked despite its broad use and importance across many fields (Borisov et al., 2022). Tabular data differs from images in several key ways:

- **Mix of discrete and continuous features**:

  Real-world tabular data commonly includes a blend of categorical and numerical columns, which require specialized preprocessing before use in neural networks, typically designed to handle normalized floating-point values. These preprocessing steps, however, can introduce biases, disturb feature relationships, and create data sparsity, all of which may hinder model performance and reduce the quality of synthetic data generated (Xu et al., 2019). Additionally, it's essential that preprocessing is reversible to accurately reconstruct the original data from synthetic samples.

- **Non-Gaussian distribution**:

  Unlike image data, where pixel values generally follow a Gaussian-like distribution and are straightforward to normalize, continuous features in tabular datasets often deviate from this. They can exhibit complex patterns such as multimodal distributions with multiple peaks, long-tailed distributions with many outliers, or skewed distributions with an asymmetric shape. Since neural networks are optimized for Gaussian inputs (Borisov et al., 2022), this creates a challenge, requiring custom preprocessing and post-processing methods to ensure accurate representation of non-Gaussian data.

- **Imbalanced categories and sparse values**:

  In many real-world scenarios, categorical data is often imbalanced, with certain classes appearing much more frequently than others. Generative models frequently struggle to capture these under-represented classes due to mode collapse. This challenge is further exacerbated by missing values or sparse features in tabular datasets, where most entries may be zeros. Additionally, common methods for encoding discrete features, like one-hot encoding, increase the sparsity of the transformed data, adding yet another obstacle for generative models (Xu et al., 2019).

2.4.2 Generative models for data generation

Generative models are a class of machine learning models that aim to capture the underlying data distribution of a training set, allowing for the generation of new data points from that distribution. The main types of deep generative models are variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models (DMs). Figure 2.1 illustrates the architectural differences between them.

Variational autoencoders

VAEs generate data by learning a latent representation and applying a regularization term that keeps this representation close to a defined distribution. In other words, the input data x is encoded into a latent space z. After training, we can draw samples from the latent distribution $p(z)$ and

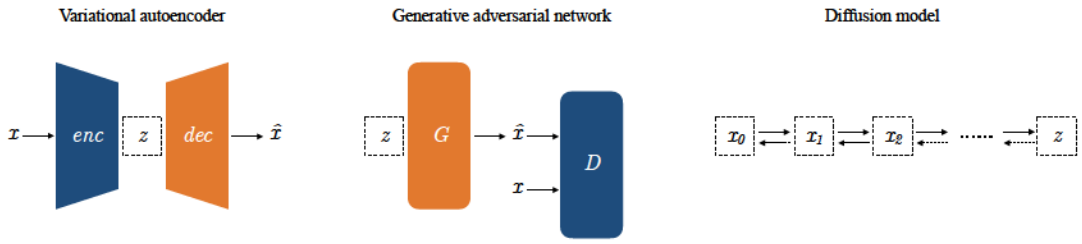


Figure 2.1: Illustrations of the three types of generative models relevant to tabular data synthesis. Coloured blocks indicate the Encoder (*enc*), Decoder (*dec*), Generator (G), and Discriminator (D) networks. A real dataset x is used to train the models, generating a synthetic dataset $\hat{x}$. The latent space from which the models sample is denoted by z.

decode these into new data $\hat{x}$. VAEs consist of two neural networks: an encoder network q_ϕ parameterized by weights ϕ and a decoder network p_θ parameterized by weights θ. The VAE's objective, the Evidence Lower Bound (ELBO), is optimized by maximizing the following expression:

$$\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x) \,||\, p(z)) \tag{1}$$

where D_{KL} denotes the Kullback-Leibler divergence. The first term promotes similarity between the reconstructed data $\hat{x}$ and the original data x, while the second term enforces a prior distribution on z, typically a standard Gaussian. This formulation ensures stable training and approximate likelihood estimation. However, it inherently limits the model's ability to capture complex data distributions, often producing less realistic samples (Bond-Taylor et al., 2021).

While some research has explored VAEs for generating tabular data (Abay et al., 2019; Acs et al., 2017; Xu et al., 2019), the notable success of GANs in image generation has shifted much of the focus to GAN-based models (Jordon et al., 2022). Even when used, VAEs are frequently integrated within GAN frameworks (Torfi & Fox, 2020; Torfi et al., 2022).

Generative adversarial Networks

Unlike VAEs, GANs rely on adversarial training between a generator and a discriminator, allowing the model to implicitly learn the data distribution. However, this adversarial setup often causes instability, as the training process can become imbalanced, leading to challenges such as

mode collapse and non-convergence (Bond-Taylor et al., 2021).

In the context of tabular data, E. Choi et al. (2017) introduced medGAN, an approach that combines an autoencoder with a GAN framework to generate synthetic health records, focusing specifically on discrete data types. N. Park et al. (2018) expanded on this concept with table-GAN, designed to handle mixed-type tabular data. Built upon DCGAN (Radford, 2015), table-GAN incorporates an auxiliary classifier to enhance performance. It is important to note that in these approaches, categorical variables were often represented as integers, which could lead to misinterpretation by suggesting ordinal relationships in non-ordinal data. In the work of Xu et al. (2019), Conditional Tabular GAN was introduced to address challenges in generating mixed-type tabular data. It effectively handles non-Gaussian, multimodal distributions and uses fully connected layers in its generator and discriminator to capture complex feature correlations. It also employs Wasserstein loss with gradient penalties for improved training stability.

Diffusion models

Diffusion Models (DMs) have since emerged as a more stable alternative for generative modelling. They gradually add noise to the data and then learn to reverse this process. By dividing the sampling process into smaller sequential steps, DMs enable smoother training. They have gained significant attention in image synthesis due to their high sample quality and reliability (Croitoru et al., 2023; Yang et al., 2023), yet their application to tabular data synthesis remains largely underexplored (Yang et al., 2023). In the study by Kotelnikov et al. (2023), the TabDDPM model is applied to generate privacy-preserving synthetic data and is evaluated across various datasets. However, this approach implements only a direct denoising model. In our work, we fine-tune a different diffusion model, TableDiffusion (Truda, 2023), to predict noise rather than just denoise, enabling us to reconstruct the data more effectively.

Chapter 3

Methodology

This section begins with the methodology for Random Forest feature selection, followed by an exploration of density-based clustering using DBSCAN (Density-Based Spatial Clustering of Applications with Noise) for data grouping, and concludes with details on the diffusion model applied for data generation.

3.1 Random Forest Feature Importance

We adopt a selective approach to assess the relative importance of each feature, considering both its relationship to other features and the target label. By applying feature selection, we focus on a subset of the most informative features, which helps reduce noise from irrelevant data and can improve the robustness (Tirelli & Pessani, 2011) and explainability of the device counting model.

Using feature selection techniques also reduces the dimensionality of our data, enhancing model usability and improving processing speed—both essential for deployment on an edge device. While dimensionality reduction methods like Principal Component Analysis (PCA) also lower dimensionality, PCA has certain drawbacks that make it less suitable to our context.

First, PCA sacrifices interpretability. By transforming original features into principal components, which are linear combinations of the original features, PCA produces abstract features that are harder to interpret. This lack of interpretability is a significant limitation, particularly in our context, where understanding distinct and non-standardized features is important.

Second, PCA assumes linear relationships among features, which may not accurately reflect real-world data patterns. If the relationships among features are non-linear, PCA is unlikely to capture this structure effectively, leading to potential loss of information.

Third, PCA maximizes variance in the dataset, assuming that components with the highest variance contain the most informative content. However, in supervised learning, high variance doesn't necessarily correspond to high relevance for the target variable. This could result in discarding low-variance features that are, in fact, valuable predictors of our target outcomes.

Finally, PCA may face challenges when applied to sparse datasets (Zou & Xue, 2018), such as ours, where a significant proportion of features contain zero values. This sparsity can undermine the effectiveness of PCA by distorting the variance-capturing process and potentially leading to less accurate or meaningful results.

Since some features in 802.11 frames are optional and their presence may vary, accounting for commonly included features is essential to ensure data consistency and enhance the model's generalizability. Random Forest-based feature selection can efficiently identify important subsets of features while considering those with high prevalence (Nicodemus, 2011). As an ensemble of decision trees, Random Forest (RF) estimates feature importance based on how well a feature increases the purity of leaf nodes in decision trees during training (Breiman, 2001).

Gini importance evaluates the significance of the features by measuring the reduction in node impurity, using the Gini impurity as the criterion. Leaf purity, often measured by Gini impurity, reflects the homogeneity of data points in a node after a split, where more important features lead to purer leaves and better data separation. For a node S containing records from k classes, the Gini impurity is computed as follows in (2):

$$G(S) = 1 - \sum_{i=1}^{k} P_i^2 \tag{2}$$

where P_i is the proportion of samples in class i at node S. The higher the Gini impurity, the less pure the node is, indicating more mixed classes and less information gained from the split. At each split in a decision tree, the Gini impurity is calculated for both the parent node S and its two child nodes. The importance of a feature is then determined by the weighted reduction in impurity for all

the splits that involve the feature across the forest. The Gini importance of feature j is given by (3):

$$I_G(j) = \sum_{t \in T} \sum_{s \in S_j} \Delta G(t, s) \tag{3}$$

where T represents all the trees in the forest, and S_j represents the nodes in a tree where splits on feature j occur. In a tree t, $\Delta G(t, s)$ is the reduction in impurity at node $s \in S_j$ after splitting based on feature j. It results from splitting the samples into two children nodes S_l and S_r with respective samples $p_l = \frac{n_l}{N}$ and $p_r = \frac{n_r}{N}$, as shown in (4):

$$\Delta G(S) = G(S) - p_l \cdot G(S_l) - p_r \cdot G(S_r) \tag{4}$$

3.2 Density-based Clustering

Probe requests originating from the same device should exhibit enough similarity to be grouped into the same cluster, eventually resulting in the number of clusters corresponding to the number of devices. Density-based clustering algorithms are particularly suited to this purpose, as they do not require predefining the number of clusters, unlike partitioning methods such as k-means and k-medoids. Furthermore, density-based methods are more flexible in identifying clusters of arbitrary shapes, while partitioning and hierarchical clustering methods tend to perform best for spherical clusters, making them less adaptable in complex scenarios.

The traditional use of clustering methods often relies on syntactic features, like proximity in a feature space, which alone may not be sufficient in the context of Wi-Fi probe requests. Due to MAC address randomization, significant noise is introduced, complicating reliance on surface-level syntactic features. Therefore, it becomes critical to base clustering on semantic features—those that capture device behaviour and intrinsic properties. By leveraging Information Elements (IEs), our approach seeks to uncover latent semantic structures within the probe requests, ensuring that clusters are not merely the result of syntactic similarity but are driven by meaningful behavioural patterns.

DBSCAN (Ester et al., 1996) is a widely used density-based clustering algorithm known for its ability to handle noise. It has two primary parameters:

- Epsilon (ϵ): the maximum distance between two probe requests in order to be considered in the same neighbourhood. It essentially determines the radius of the cluster around each probe.
- Minimum Points (*min-points*): the minimum number of probe requests required to form a dense region. If there are at least *min-points* probe requests within an ϵ radius of a particular probe, the latter is considered a core point, and a cluster is formed around it.

DBSCAN initiates by selecting an arbitrary point and identifying all neighbouring points within a distance ϵ. If the number of neighbouring points is greater than *min-points*, a cluster is established. The point and its neighbours are subsequently added to the cluster, and the point is marked as visited. If the number of neighbouring points is less than the specified *min-points*, the point is classified as noise. The algorithm continues iteratively until all points have been visited.

The selection of ϵ and *min-points* is often informed by empirical knowledge, with iterative adjustments performed through trial and error to achieve optimal clustering outcomes (Song et al., 2018). For data of dimensionality d, a commonly used heuristic suggests setting *min-points* to $d+1$.

The vanilla version of DBSCAN detects clusters using a single global density threshold, which limits its effectiveness for datasets with diverse intrinsic density levels. This drawback is especially pertinent to our situation, as probe requests are not uniformly standardized and can differ considerably between manufacturers and even across operating system versions.

Adapting to different density levels can be achieved by adjusting ϵ using a k-distance graph, where we compute the average distance to each point's k-nearest neighbours. Averaging these distances smooths the k-distance graph, reducing noise and helping to identify appropriate density thresholds. The k-distance d_k for a probe request p in the dataset D is the distance to its k-th nearest neighbours in terms of a chosen distance metric d. It is expressed in (5) as:

$$d_k(p) = \min_{R \subseteq D, |R|=k} \operatorname*{mean}_{q \in R} d(p, q) \tag{5}$$

By sorting these average distances, we can observe patterns related to the dataset's density levels. Points associated with noise tend to exhibit larger k-distances, whereas sharper transitions suggest density changes. These transitions – often called knees– correspond to potential ϵ values. Mathematically, if the k-dist plot is viewed as a function f of the sorted points, a knee point is where

the second derivative tends to zero. In the case of multiple density levels, the k-distance plot would typically feature smooth curves connected by regions of sharp variation.

We experimented with both Euclidean and Hamming distances as metrics for determining the k-nearest neighbours. Ultimately, we selected the Hamming distance for the final approach, as it produced superior clustering outcomes. Given n features, the Hamming distance between two probes p and q is defined in (6) as:

$$d_{\text{Hamming}}(p, q) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(p_i \neq q_i)} \tag{6}$$

3.3 Diffusion-based Data Generation

Diffusion models represent a recent class of implicit generative models that generate data by gradually injecting noise into data over T discrete time steps and then reversing this process to reconstruct the original data distribution. The diffusion model in this work is based on the formulation of Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020), where the noise injection and data recovery processes are modelled through two Markov chains. In the forward process, Gaussian noise z_t is introduced to the data x_t at each time step t, as shown in (7).

$$x_{t+1} = x_t + z_t, \text{ with } z_t \sim \mathcal{N}(0, \beta_t I) \tag{7}$$

As β_t increases over time, the noise variance grows, ensuring a smooth transition. Instead of the linear noise schedule from DDPM by Ho et al. (2020), we adopt a trigonometric noise schedule defined in (8) as:

$$\beta_t = \sin^2\left(\frac{\pi\, t}{2T}\right) \tag{8}$$

The trigonometric space provides a smoother noise schedule compared to the linear one, leading to more gradual data degradation, especially near the end of the process (Nichol & Dhariwal, 2021). In contrast, the linear schedule injects too much noise early on, potentially destroying information prematurely. As a result, fewer steps would be needed in the reverse process.

In the reverse process, starting from the fully noised state x_T, the aim is to recover the original data x_0. This is achieved by training a neural network M_θ to minimize the mean squared error between the predicted and actual noise, thereby estimating the noise added during the forward process. Specifically, the model learns to predict the noise $\hat{z}_t$ at each step, aiming to reconstruct $\hat{x}_0$, as expressed in (9):

$$\hat{x}_0 = x_T - \sum_{t=1}^{T} \hat{z}_t, \quad \text{where } \hat{z}_t \sim M_\theta(x_t) \tag{9}$$

For generating synthetic tabular data, we first map the raw mixed-type data into a latent space where it is represented homogeneously. The diffusion process is then applied within this latent space, which facilitates the generation of high-quality synthetic tabular data.

Unlike traditional generative models, diffusion models allow for the gradual reconstruction of data, ensuring that the semantic integrity of the original probe requests is maintained throughout the generation process. This characteristic makes diffusion models particularly well-suited for tasks requiring a deep understanding of latent semantic relationships between features, as is the case in Wi-Fi probe request clustering. By preserving these semantic relationships, diffusion models provide a robust foundation for clustering probe requests in scenarios where maintaining device behaviour characteristics is crucial.

Chapter 4

Data and Results

In this chapter, we present the data used for testing and validating our proposed Wi-Fi probe request-based crowd-counting model and discuss the results of our experiments. The analysis begins with an overview of the dataset composition, including key characteristics of the Pintor dataset. The unique features and limitations of each dataset are highlighted, demonstrating the challenges in modelling probe requests under MAC address randomization.

4.1 Dataset Overview

In this study, we use the Pintor dataset (Pintor & Atzori, 2022b) to validate our findings. This labelled dataset associates each probe request to the label of its generating device. It consists of capture files containing the transmitted messages over Wi-Fi channels. Multiple captures were taken for each device under various settings. These settings are divided into two main categories: active-screen modes (A, PA, and WA), in which the device's screen remained on during capture while playing a video, and inactive-screen modes (S, PS, and WS), where the screen was on standby. In power-saving modes (PA and PS), the device's power-saving setting was enabled, whereas in all other captures, it was disabled. Additionally, WA and WS modes indicate captures where the device's Wi-Fi interface was off; in all other modes, the Wi-Fi interface remained on but was not connected to any access point. A summary of the data exploration for the Pintor dataset is provided in Table 4.1, presenting an overview of its settings.

		Number of devices
Operating system	Android	17
	iOS	5
Environment setup	Anechoic chamber	8
	Noisy environment	14
MAC randomization	YES	17
	NO	5

Table 4.1: Device distribution in the Pintor dataset across different experiment settings

The importance of this dataset lies in its labelled nature. While our approach is unsupervised, specifically focused on calibrating a clustering algorithm, we could typically only use unsupervised metrics for model validation. However, we plan to leverage the available labels for an additional validation layer, in a manner similar to how supervised models are tested (without using the labels for learning or calibration). We contend that relying exclusively on unsupervised metrics, like the silhouette score, for evaluating clustering performance is insufficient. These metrics primarily assess the internal structure of the clusters, such as cohesion (how closely related the points within a cluster are) and separation (how distinct the clusters are from each other). While these aspects are important and can provide valuable insights, they may not fully capture the practical or intended purpose of the clusters, particularly in real-world applications where the clusters need to correspond to meaningful and domain-specific categories. Moreover, unsupervised metrics are susceptible to biases that can distort the assessment of clustering quality. This is particularly evident when clusters overlap or when the data exhibit varying densities across different regions. Some metrics prioritize compactness, which might cause them to miss subtler, yet semantically significant, distinctions between data points within a cluster.

The Pintor dataset predominantly includes devices with MAC address randomization enabled, which can be identified by examining the MAC address. A MAC address consists of two 6-byte sections. The first part, known as the Organizationally Unique Identifier (OUI), identifies the manufacturer or organization responsible for producing the device. The second part uniquely identifies each device within that manufacturer's product line. A 1 value in the seventh most significant bit

(B1) implies a random MAC address, whereas a value of 0 indicates a real physical MAC address. Figure 4.1 illustrates the structure of a MAC address.

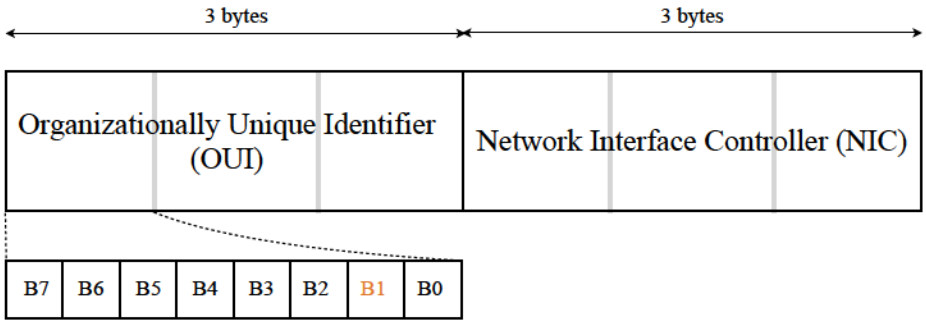


Figure 4.1: The structure of MAC address with the functional bits

Our primary focus in this project lies in the analysis of the IEs within the dataset. As previously noted, not all the IE fields are mandatory. Table 4.2 provides an overview of each IE's frequency within the dataset, detailing the percentage presence of each. To prepare the IEs for analysis, we extract their content from the dataset and convert it into numeric values. If an IE is missing or empty, we assign it a value of zero. If the IE is already numeric, we leave it unchanged. For arrays, we reduce them to a single numeric value by summing their elements. If the IE is a string, we calculate a numeric equivalent by adding up the ASCII values of each character. Prior to applying the machine learning algorithms, we applied min-max normalization. Notably, a single Probe Request can include multiple 221 and 127 IEs with distinct contents. To account for this, we combine the values of all instances of these IEs by summing them. Next, we add presence flags IE X* each containing a binary value to indicate whether the corresponding IE X is present and used in the probe request. These flags are intended to provide additional information about the probe request, helping to characterize it more clearly.

We initiated the development of a custom dataset, termed the "Concordia Dataset", with the aim of creating a labelled dataset specifically designed to address the challenges in device counting validation. This dataset project was launched to evaluate the process of dataset creation, which proved to be labour-intensive, time-consuming, and costly. The Concordia dataset is distinct from the Pintor dataset in that it is specifically tailored for device counting. Unlike the Pintor dataset

IE ID	Name	Description	Presence
0	Service Set Identify (SSID)	Name of the wireless network	100%
1	Supported Rates	Supported Data rates for communication	100%
50	Extended Supported Rates	Identical to supported rates, but used when the supported rates exceed 8 rates	99.9%
3	DS Parameter Set	Channel number	97.8%
45	HT Capabilities	Details on the High Throughput capabilities of the network (802.11n)	96.8%
221	Vendor Specific	Proprietary information unique to each manufacturer	90.9%
127	Extended Capabilities	Features beyond core standard capabilities, e.g., Wi-Fi Direct	83.5%
255	Element ID Extension	Allowing additional IEs to be used	14.3%
191	VHT Capabilities	Details on the Very High Throughput capabilities of the network (802.11ac)	11.1%
107	Interworking	Types of access networks, associated costs, and the type of venue.	3.3%

Table 4.2: Description of the information elements present in the Pintor dataset

approach, which involves isolating a single device in an anechoic chamber, our methodology intentionally introduces multiple devices into the chamber, each with an assigned label representing the device count (e.g., a configuration of five phones corresponds to the label five). Table 4.3 gives an overview of the Pintor dataset raw data.

MAC address	sequence number	length	RSSI	IE 0	IE 1	IE 3	IE 50	IE 45	IE 127	IE 107	IE 221	IE 191	IE 255	Label
ec:9b:f3:75:8e:40	4096	123	-35	0	551	0	414	270	12	0	0	0	7171	H
22:8f:aa:fb:bf:51	5248	131	-33	0	402	0	809	270	20	0	0	0	8001	P
22:8f:aa:fb:bf:51	5264	131	-31	0	402	0	809	270	20	0	0	0	8001	P
da:a1:19:9c:fd:cf	56144	127	-76	0	39	2	414	265	3	0	20730	0	0	B
da:a1:19:9c:fd:cf	56432	125	-48	0	39	3	414	265	3	0	20730	0	0	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
da:a1:19:b5:fa:fb	128	112	-66	0	39	3	414	269	8	0	20747	0	0	J
da:a1:19:b5:fa:fb	384	112	-50	0	39	6	414	269	8	0	20747	0	0	J

Table 4.3: Excerpt of raw data rows used in the study

We wanted to create our own dataset, the Concordia dataset, to understand the process behind

building such particular ones in anechoic chambers. Our goal was to assess how labour-intensive it is and uncover the specific steps and unique challenges involved. Drawing from the analysis of the Pintor dataset, we set out to develop ours with the following configuration. Through this experimental design, the dataset attempts to capture real-world complexities, such as signal interference, device proximity effects, and multi-device interactions, making it highly unique and particularly valuable for validating device counting algorithms. However, due to resource limitations and the intensive nature of the data collection process, we were able to compile only a small amount of data.

To ensure that the frames we capture originate exclusively from our devices, the ideal setup environment would be an isolated room, such as a Faraday cage, completely shielded from external Wi-Fi signals. Given the unavailability of such a room, we opted for a hemi-anechoic chamber as an alternative available at Concordia University, where we place a variety of smartphones within a one-meter line-of-sight from the scanner. Each experiment takes around 10 to 15 minutes. The scanner is a MacBook Pro M1 equipped with a Wi-Fi chipset supporting monitor mode, essential for capturing probe requests. We use Wireshark software to capture probe requests and store them in PCAPng files (Packet CAPture new generation). Table 4.4 gives an overview of the devices used so far in the Concordia dataset.

Device Name	**Operating System Version**	**Vendor**
iPhone 12	iOS 16.6	Apple
Galaxy A51	Android 13	Samsung
Iphone 5C	iOS 10.3.3	Apple
Galaxy S6	Android 7	Samsung
Galaxy Core	Android 4.4.2	Samsung
Nexus 5X	Android 8.1	LG Electronics
iPhone 4	iOS 7.1.2	Apple
Galaxy S9	Android 9	Samsung
iPhone 4S	iOS 9.3.4	Apple
iPhone 14	iOS 16.6	Apple

Table 4.4: Device names and operating system versions from the Concordia dataset

4.2 Experiments and Results

Figure 4.2 illustrates the Gini Importance of the available features in probe requests. It is clear that the probe requests length, IE 127, IE 221 contain a substantial amount of information compared to other features. We also include IE 45 in our analysis to capture any additional, albeit smaller, informational contributions.

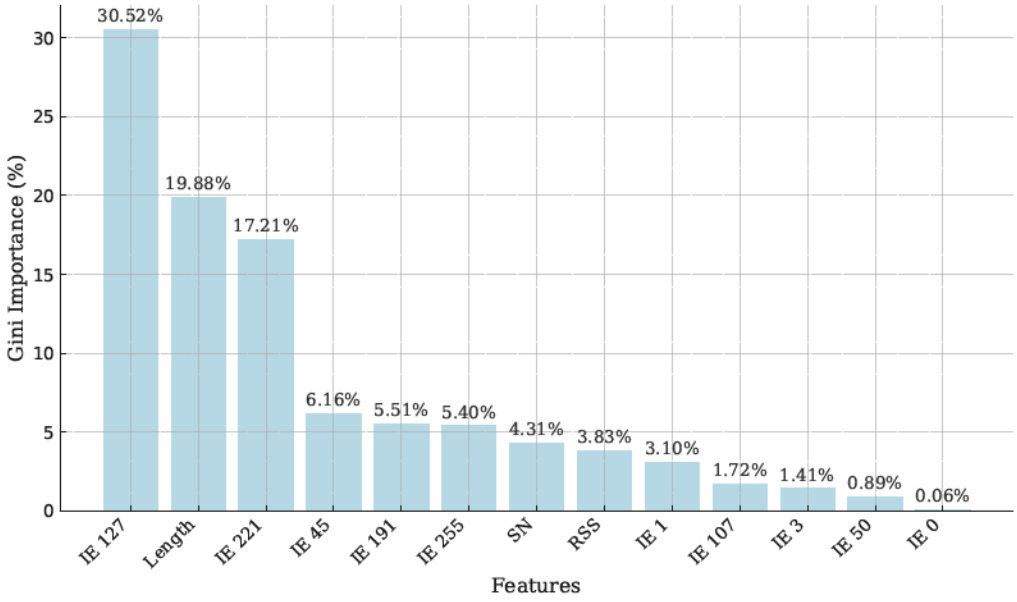


Figure 4.2: Gini importance scores for all features in the Pintor dataset

The Lorenz curve, depicted in Figure 4.3 further supports these observations by illustrating the cumulative percentage of features relative to their cumulative importance. If each feature contributes equally, a given proportion of features would account for the same proportion of cumulative importance, represented by the red line of equality. However, the observed Lorenz curve deviates below this line, forming a bowed shape, indicating that a small subset of features contributes disproportionately to the overall importance.

Based on the preceding analysis, we will proceed using the probe request length along with Information Elements (IEs) 221, 45, and 127. Our objective is to apply the DBSCAN clustering algorithm to group probe requests. Clustering is used under the assumption that probe requests originating from the same device exhibit sufficient intrinsic similarities. Thus, the number of clusters will correspond to the estimated number of distinct devices.

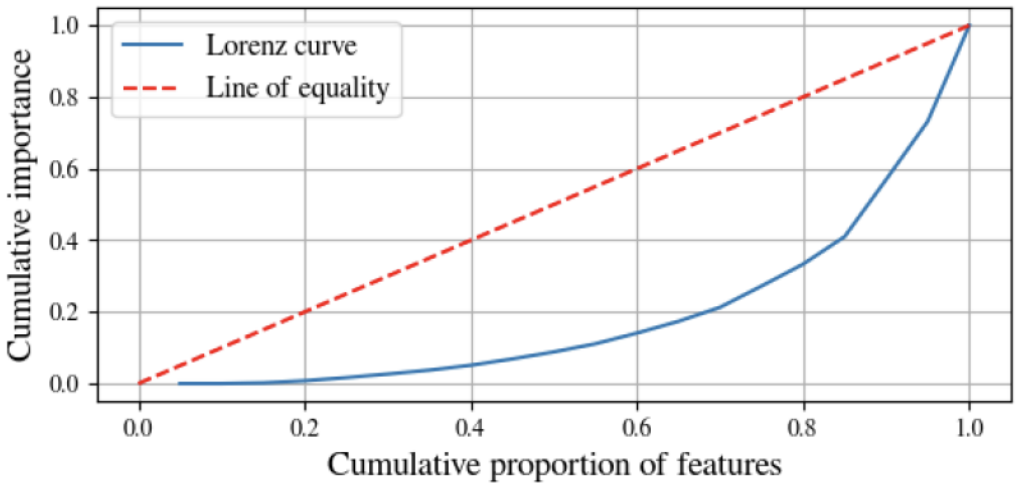


Figure 4.3: The Lorenz curve illustrating the cumulative importance as the proportion of features increases

The effectiveness of the clustering approach will be evaluated using 3 metrics:

- **Absolute Error**: it measures the disparity between the number of clusters generated and the count of devices contributing to our dataset.

- **The average silhouette score**: To evaluate clustering quality, we calculate the silhouette score, which helps determine how well-separated clusters are by quantifying the distance between each probe request and its surrounding clusters. The silhouette score $s(p)$ for a probe request p is defined as:

$$\mathrm{s}(p) = \frac{b(p) - a(p)}{\max[b(p), a(p)]} \tag{10}$$

  where $a(p)$ is the average distance between the probe request p and all other probe requests in the same cluster and $b(p)$ is the average distance between the probe request p and the probe requests in the nearest neighbouring cluster. A silhouette score of 1 means the probe request is well-clustered, being close to its own cluster and far from others. A score of 0 indicates the probe request is near the boundary between clusters, with no clear fit to either. A score of -1 suggests the probe request is likely misclassified, as it is closer to another cluster than its own.

- **V-measure**: it is an entropy-based metric used to quantify how successful a clustering is.

V-measure is computed as the harmonic mean of homogeneity and completeness scores.

A clustering satisfies homogeneity if probe requests within each cluster originate from the same device. A homogeneity score is calculated as follows:

$$h = 1 - \frac{H(C|K)}{H(C)} \tag{11}$$

where $H(C|K)$ is the conditional entropy of the class distribution given the proposed clustering. In the perfectly homogeneous case, this value is 0.

A clustering satisfies completeness if all probe requests coming from a given device are clustered in the same cluster. A completeness score is calculated as follows:

$$c = 1 - \frac{H(K|C)}{H(K)} \tag{12}$$

The V-measure is then expressed in (13) as:

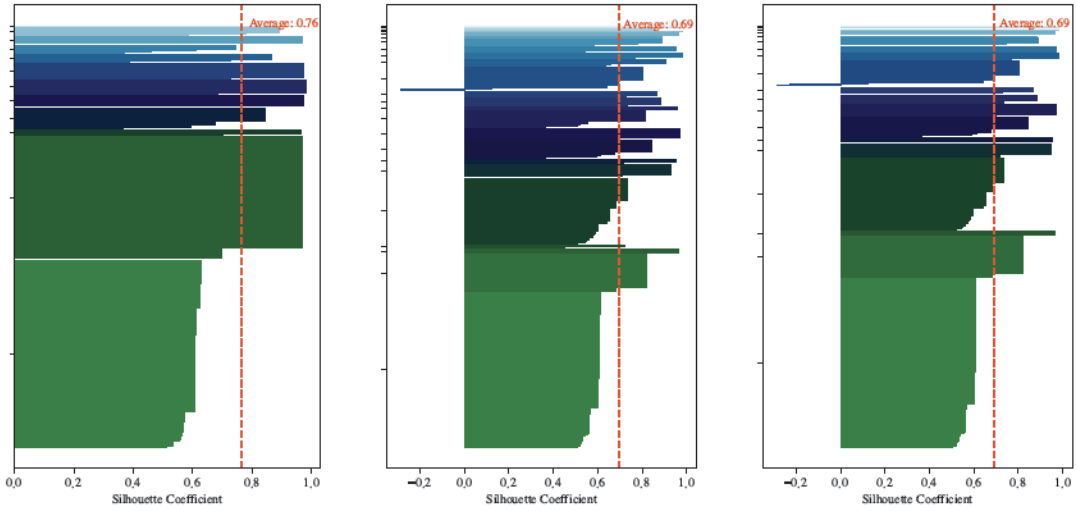
$$\text{V-measure} = \frac{2 \cdot h \cdot c}{h + c} \tag{13}$$

We run DBSCAN with the parameters defined for different (overlapping) subsets of the dataset.

- Subset 1, including data from 12 devices, with two devices not implementing MAC address randomization.
- Subset 2, including data corresponding to 20 devices, with three devices not implementing MAC address randomization.
- Subset 3, including data from 17 different devices, all implementing MAC address randomization.
- Subset 4, including data from 10 devices, all implementing MAC address randomization.
- Subset 5, including data from all 22 devices.

In Figure 4.4, the silhouette coefficients relative to each subset illustrate intra-cluster cohesion

and inter-cluster separation for the different subsets. Subset 1, with the highest average silhouette score, has values consistently close to 1, indicating well-separated clusters. This suggests that the selected features effectively differentiate data points, contributing to the subset's strong performance. The narrow distribution further indicates minimal overlap between clusters, underscoring the robustness of this clustering. In contrast, Subset 3, which has the lowest average silhouette score, shows lower values, including some negative scores, reflecting poorer clustering quality.



(a) Subset 1 (b) Subset 2 (c) Subset 3

(d) Subset 4 (e) Subset 5

Figure 4.4: Comparison of silhouette scores for the clustering subsets

It is important to emphasize that silhouette scores alone do not reliably guarantee effective

clustering outcomes. To ensure validity, they should be complemented by additional metrics, such as the supervised V-measure. For example, a high silhouette score does not necessarily indicate well-defined inter-cluster distinctions. Therefore, a comprehensive evaluation requires combining multiple metrics, including the silhouette score, V-measure, and absolute error, to provide a more robust assessment.

	Hamming distance			**Euclidean distance**			**Euclidean distance without presence flags**		
	V-measure	Silhouette score	Absolute error	V-measure	Silhouette score	Absolute error	V-measure	Silhouette score	Absolute error
Subset 1	0.957	0.764	0	0.885	0.917	8*	0.804	0.870	3*
Subset 2	0.919	0.692	4	0.867	0.852	6	0.671	0.776	8
Subset 3	0.909	0.691	2	0.873	0.869	1	0.641	0.785	6
Subset 4	0.915	0.701	3	0.878	0.873	1	0.659	0.790	11
Subset 5	0.940	0.711	3	0.865	0.851	7	0.564	0.618	9

* including one cluster labelled as noise

Table 4.5: Comparison of clustering metrics across the different subsets

The results in Table 4.5 reveal two key insights: the Hamming distance is more suitable than the Euclidean distance, and the presence flags contribute to a more robust and binary difference counting. Indeed, the use of Hamming distance creates a distinct separation between data points, resulting in a more gradual and stable k-distance curve characterized by multiple potential knees. This enables the knee point to be identified without interference from abrupt drop-offs. In contrast, Euclidean distance often produces a sharp, singular drop in the curve, where the transition to near-zero distances can cause instability, potentially leading to the selection of unrealistically small ϵ values.

Table 4.5 also highlights the significance of incorporating a supervised metric in conjunction with the silhouette score. Indeed, Subset 1 with Euclidean distance attains a high Silhouette score but results in inaccurate counting. This is most probably due to the cluster's compactness and the spatial distribution of probe requests. This observation also highlights how Hamming distance can mitigate this compactness, improving counting accuracy.

To assess the performance of the diffusion model, we compare the mean and standard deviation differences between the real and synthetic data. As we can depict from Figure 4.5, the points (very often overlapping) are positioned very close to the $y = x$ line in both the mean and standard deviation plots, indicating that the synthetic data replicates the statistical properties of the real data.

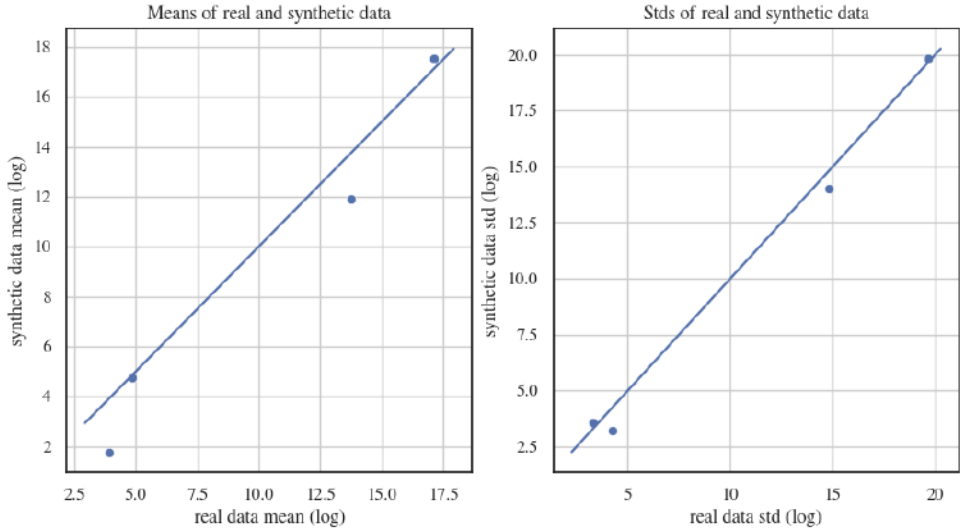


Figure 4.5: Comparison of the absolute log mean and standard deviations between original and diffusion-based synthetic data

To evaluate the performance differences between the used GAN model (Xu et al., 2019) and the diffusion model, we conduct a Principal Component Analysis (PCA) and present the first two principal components of the real data, GAN-generated synthetic data, and diffusion-generated synthetic data. It is clear in Figure 4.6 that the diffusion model's synthetic data closely mirrors the real data in terms of component distribution, scale, and spatial positioning. In contrast, the PCA plot of the GAN-generated data reveals numerous points that are poorly aligned, indicating a less accurate representation of the real data structure.

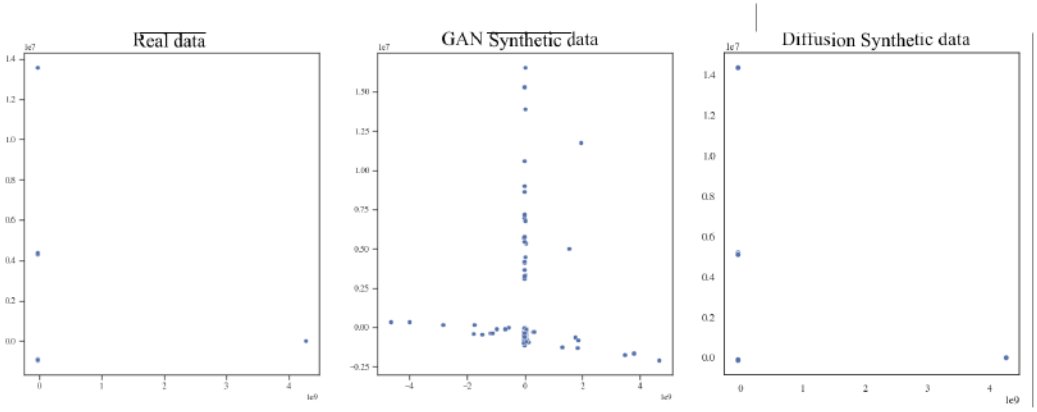


Figure 4.6: PCA plot of the first two principal components showing the distribution of original, GAN-generated, and diffusion-generated data

We further evaluate the differences between the GAN and the diffusion model by plotting the cumulative sum of real and synthetic data. The cumulative sum plot is important as it helps to visually assess how well the synthetic data captures the overall trends and distribution of the real data. A close match indicates good data generation, while significant discrepancies suggest potential issues with the model's performance. In both Figure 4.7 and Figure 4.8, we observe that the cumulative sum of the length feature is notably smoother than that of the other features, which lack a continuous distribution.

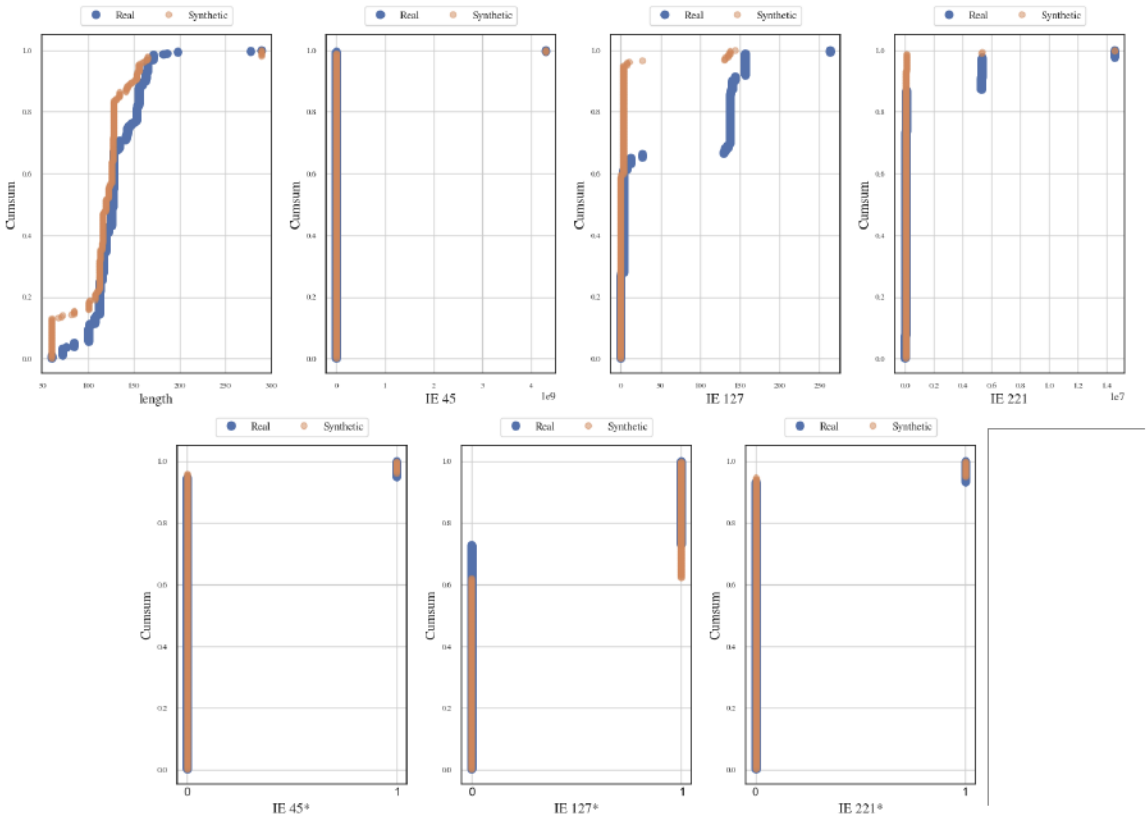


Figure 4.7: Cumulative sum comparison of real versus diffusion-generated synthetic data across all features

When synthetic data diverges from real data, this may highlight limitations in the generation process, potentially due to limited data diversity or sparsity in specific areas of the real dataset. Notably, the diffusion model captures non-Gaussian features more accurately, while the GAN performs better on Gaussian-like distributions, such as the length feature. However, the GAN struggles with

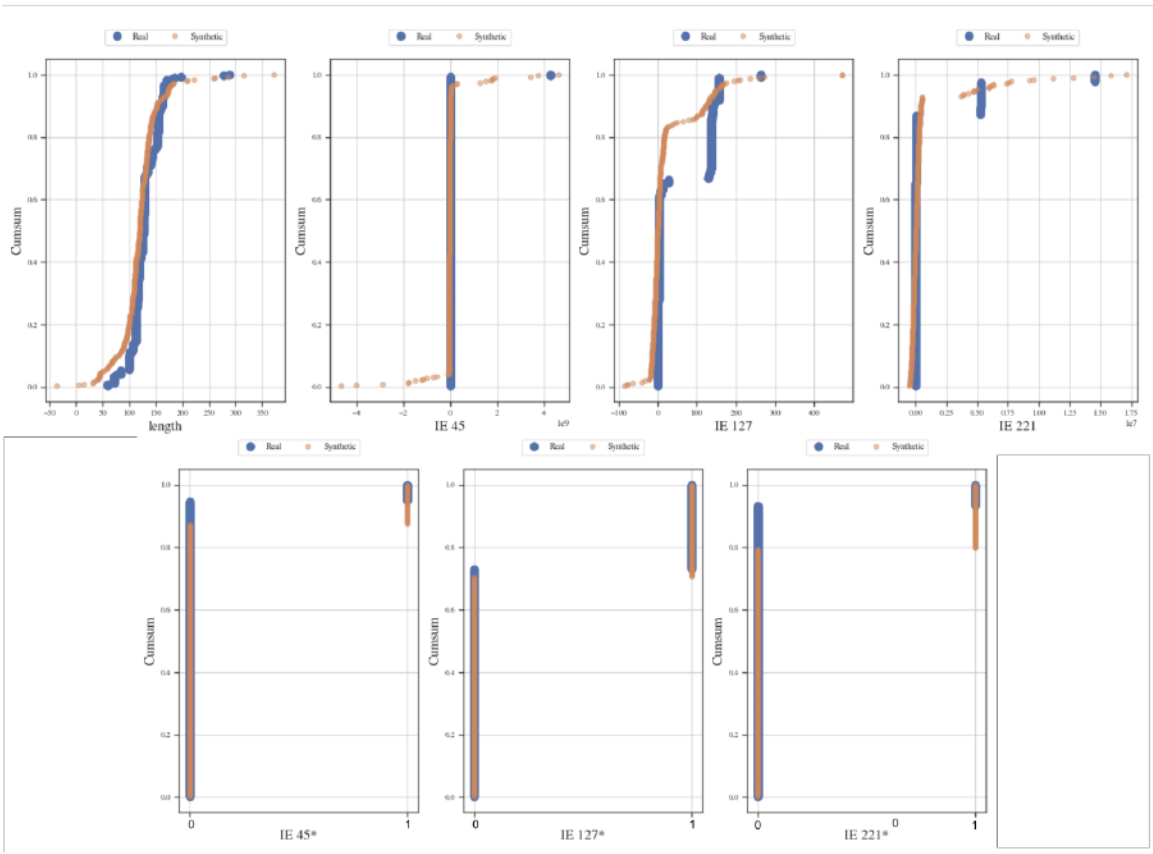


Figure 4.8: Cumulative sum comparison of real versus GAN-generated synthetic data across all features

discontinuous features compared to the diffusion model, which shows greater adaptability to non-continuous data patterns. Figure 4.9 illustrates the distributions of the length feature, IE 45, IE 127, and IE 221. It is evident that the length feature appears more Gaussian and continuous compared to the others, which is likely why the GAN performs better with it.

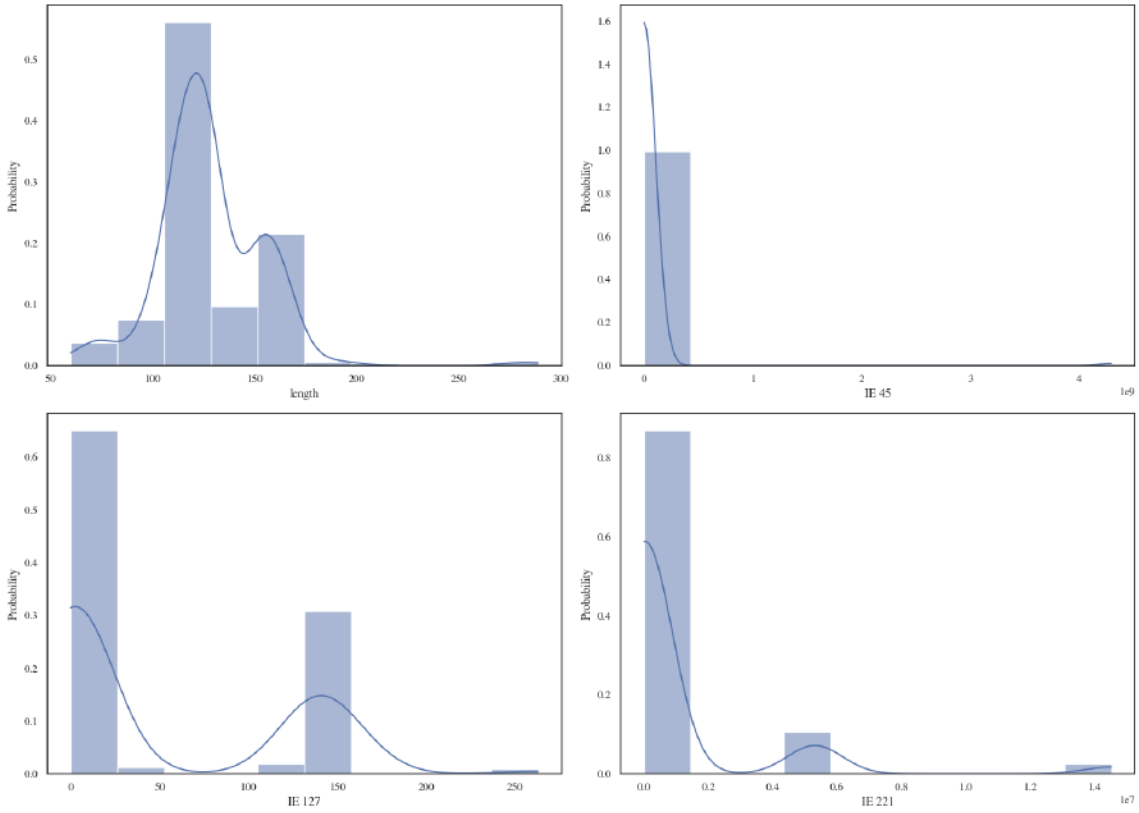


Figure 4.9: Distributions of the continuous variables in the dataset

Once the synthetic data is generated, we integrate 30% of it with the real data and reapply the DBSCAN clustering algorithm. The clustering results, presented in Table 4.6, demonstrate a notable improvement, particularly in terms of the average silhouette score and absolute error, indicating an enhanced clustering performance.

	V-measure	Silhouette score	Absolute error
Subset 1	0.952	0.789	0
Subset 2	0.931	0.721	3
Subset 3	0.908	0.701	0
Subset 4	0.918	0.721	3
Subset 5	0.920	0.701	3

Table 4.6: Comparison of clustering metrics across the subsets using the mixed (original and diffusion-generated) data

Chapter 5

Conclusion & Future Work

5.1 Key Contributions & Insights

This thesis introduces an innovative privacy-preserving approach to crowd counting, using Wi-Fi probe requests to estimate passenger numbers at bus stops. The research focuses on a strategic selection of specific Information Elements as features for crowd size estimation, demonstrating the feasibility of leveraging Wi-Fi probe request data for accurate crowd counting. The primary contributions of this work are as follows:

- **A selective feature set:** Through a feature selection analysis, we establish that reliable crowd-counting can be achieved with a minimal set of features—specifically, four key attributes: the probe request length, and three distinct IEs: IE 127, IE 221, and IE 45. This selective approach optimizes computational efficiency while maintaining high accuracy in estimating crowd size.

- **Introduction of presence flags:** A novel aspect of our method is the incorporation of presence flags for each IE. These binary indicators signal whether the corresponding IE is present in the probe request. Our results showed how their introduction helped with feature representation, and hence crowd estimation.

- **The use of the Hamming distance:** The clustering algorithm in our study is configured to use the Hamming distance for calculating the distance between probe requests. This approach

outperforms the use of Euclidean distance. The Hamming distance is more suitable for this purpose because it focuses on binary differences, providing a clear separation between data points.

- **Generative data augmentation techniques:** To address the challenge of scarce labelled data and the difficulty of creating such a dataset, this work leverages generative data augmentation techniques, significantly improving the model's performance. Although diffusion models have been infrequently explored for tabular data, they have demonstrated in our research the ability to generate realistic tabular samples. When combined with the original data, these synthetic samples led to enhanced results.

Eventually, our experiments validate the efficacy of the proposed pipeline, with the best-performing model achieving a V-measure of 0.952, an average silhouette score of 0.789, and a precise crowd size estimate. While originally intended for passenger counting in transit hubs, our approach is versatile and should be well-suited for a range of other applications, including event management, retail settings, and safety-critical environments.

5.2 Limitations

This study is subject to several limitations that have yet to be addressed. Specifically, we did not consider scenarios where a single individual owns multiple devices, each emitting Wi-Fi probe requests. This could lead to over-counting, as each device would be detected as a separate entity, thus inflating the estimated crowd size. Another key limitation arises from inherent constraints within the Pintor dataset, which means that while the results are promising, they should be considered in the context of this dataset, especially since both the clustering calibration and the resulting findings are fundamentally dependent on it. The main limitations of this dataset are as follows:

- Most devices use the Android operating system, which, given iOS's significant market share (Statista, 2024), introduces a bias toward Android in the dataset. It is also worth noting that early versions of Android devices typically lack MAC randomization unless explicitly enabled.

- Only 8 out of 22 devices were captured in an anechoic chamber, while the rest were recorded in a ”noisy” environment, the nature of which is unclear. If this environment isn't properly shielded, such as with a Faraday cage, unintentional probe requests from other devices could compromise the dataset, conflicting with its primary goal of device counting.

- Analyzing individual smartphones helps understand device behaviour but may not be ideal for accurate device counting. Merging probe requests from different devices doesn't replicate real-world behaviour, where message transmission depends on nearby nodes and protocols like CSMA/CA. CSMA/CA attempts to mitigate collisions by having nodes listen to the communication medium before transmitting.

5.3 Future Work

The limitations identified in this study offer opportunities for further refinement, with the goal of enhancing the reliability of the proposed crowd-counting pipeline.

Given the importance of data quality in training and calibration, further fine-tuning of the tabular diffusion model could enhance the fidelity of synthetic data generation, which would, in turn, benefit model performance. Exploring alternative clustering algorithms like OPTICS also presents a promising direction for improvement.

To address the challenge of over-counting caused by individuals carrying multiple devices, a practical approach combines statistical adjustments with external data validation. An adjustment factor could be derived from observed data or pilot studies to estimate the likelihood of individuals carrying multiple devices. Moreover, if a secondary crowd-counting system, such as a camera-based solution, is available, it could serve as a baseline for real-time validation. By periodically cross-referencing Wi-Fi-based counts with camera data, adjustments can be fine-tuned, resulting in more accurate crowd estimates.

Finally, testing in real-world environments is crucial for evaluating the feasibility of the solution. Recent research by Paradeda et al. (2019) has pointed out the limitations of Wi-Fi-based crowd-counting systems, particularly in transit applications. The authors argue that inconsistent detection rates and delays in device recognition can lead to unreliable results in real-time contexts.

While aggregated data may seem adequate, disaggregated analysis often reveals significant detection errors. They recommend augmenting Wi-Fi detection with data from sources such as GPS and refining detection parameters to improve accuracy.

Bibliography

Abay, N. C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B., & Sweeney, L. (2019). Privacy preserving synthetic data release using deep learning. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, 510–526.

Abdulrahem, B. (2021). *Crawdad nile/probe-requests*. IEEE Dataport.

Acs, G., Melis, L., Castelluccia, C., & De Cristofaro, E. (2017). Differentially private mixture of generative neural networks. *2017 IEEE International Conference on Data Mining (ICDM)*, 715–720.

Agarwal, Y., Balaji, B., Gupta, R., Lyles, J., Wei, M., & Weng, T. (2010). Occupancy-driven energy management for smart building automation. *Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building*, 1–6.

Barbera, M. V., Epasto, A., Mei, A., Kosta, S., Perta, V. C., & Stefa, J. (2013). Crawdad dataset sapienza/probe-requests.

Bond-Taylor, S., Leach, A., Long, Y., & Willcocks, C. G. (2021). Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, *44*(11), 7327–7347.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2022). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Bravenec, T., Torres-Sospedra, J., Gould, M., & Frýza, T. (2022, July). *Supplementary Materials for "What Your Wearable Devices Revealed About You and Possibilities of Non-Cooperative 802.11 Presence Detection During Your Last IPIN Visit"* (Version 1.0). Zenodo.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

BusPas Inc. (2024). About [Accessed: 2024-11-10]. https://buspas.com/about/

Cai, Y., Tsukada, M., Ochiai, H., & Esaki, H. (2021). Mac address randomization tolerant crowd monitoring system using wi-fi packets.

Calabrese, F., Ferrari, L., & Blondel, V. D. (2014). Urban sensing using mobile phone network data: A survey of research. *Acm computing surveys (csur)*, *47*(2), 1–20.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. *Machine learning for healthcare conference*, 286–305.

Choi, J. W., Kim, J. H., & Cho, S. H. (2012). A counting algorithm for multiple objects using an ir-uwb radar system. *2012 3rd IEEE international conference on network infrastructure and digital content*, 591–595.

Chon, Y., Lane, N. D., Kim, Y., Zhao, F., & Cha, H. (2013). Understanding the coverage and scalability of place-centric crowdsensing. *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 3–12.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10850–10869.

Cunche, M., Kaafar, M.-A., & Boreli, R. (2014). Linking wireless devices using information contained in wi-fi probe requests. *Pervasive and Mobile Computing*, *11*, 56–69.

Dong, Y., Wu, Y., Codling, J. R., Aggarwal, J., Huang, P., Ding, W., Latapie, H., Zhang, P., & Noh, H. Y. (2023). Gamevibes: Vibration-based crowd monitoring for sports games through audience-game-facility association modeling. *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 177–188.

Éditeur officiel du Québec. (2024). Act respecting access to documents held by public bodies and the protection of personal information (cqlr, c. p-39.1) [Consulted on: 2024-11-10]. https://www.legisquebec.gouv.qc.ca/en/document/cs/p-39.1

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016, p. 1–88 (2016, May 4). Retrieved April 13, 2023, from https://data.europa.eu/eli/reg/2016/679/oj

Fenske, E., Brown, D., Martin, J., Mayberry, T., Ryan, P., & Rye, E. C. (2021). Three years later: A study of mac address randomization in mobile devices and when it succeeds. *Proc. Priv. Enhancing Technol.*, *2021*(3), 164–181.

Fuada, S., Adiono, T., Prasetiyo, & Islam, H. W. S. (2020). Modelling an indoor crowd monitoring system based on rssi-based distance. *International Journal of Advanced Computer Science and Applications*, *11*(1).

Gade, R., & Moeslund, T. B. (2014). Thermal cameras and applications: A survey. *Machine vision and applications*, *25*, 245–262.

Groba, C. (2019). Demonstrations and people-counting based on wifi probe requests. *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, 596–600.

Guillen-Perez, A., & Cano, M.-D. (2019). Counting and locating people in outdoor environments: A comparative experimental study using wifi-based passive methods. *ITM Web of Conferences*, *24*, 01010.

Hasan, M., Hanawa, J., Goto, R., Suzuki, R., Fukuda, H., Kuno, Y., & Kobayashi, Y. (2022). Lidar-based detection, tracking, and property estimation: A contemporary review. *Neurocomputing*, *506*, 393–405.

Heurtefeux, K., & Valois, F. (2012). Is rssi a good choice for localization in wireless sensor network? *2012 IEEE 26th International Conference on Advanced Information Networking and Applications*, 732–739.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840–6851.

Hong, H., De Silva, G. D., & Chan, M. C. (2018). Crowdprobe: Non-invasive crowd monitoring with wi-fi probe. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1–23.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). Synthetic data - what, why and how? *ArXiv*, *abs/2205.03257*.

Junior, J. C. S. J., Musse, S. R., & Jung, C. R. (2010). Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine*, *27*(5), 66–77.

Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2023). Tabddpm: Modelling tabular data with diffusion models. *International Conference on Machine Learning*, 17564–17579.

Li, H., Chan, E. C. L., Guo, X., Xiao, J., Wu, K., & Ni, L. M. (2015). Wi-counter: Smartphone-based people counter using crowdsourced wi-fi signal data. *IEEE Transactions on Human-Machine Systems*, *45*(4), 442–452.

Liang, D., Xie, J., Zou, Z., Ye, X., Xu, W., & Bai, X. (2023). Crowdclip: Unsupervised crowd counting via vision-language model. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2893–2903.

Lim, R., Zimmerling, M., & Thiele, L. (2015). Passive, privacy-preserving real-time counting of unmodified smartphones via zigbee interference. *2015 International Conference on Distributed Computing in Sensor Systems*, 115–126.

Liu, W., Salzmann, M., & Fua, P. (2019). Context-aware crowd counting. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5099–5108.

Ma, H.-Y., Zhang, L., & Shi, S. (2024). Vmambacc: A visual state space model for crowd counting. *arXiv preprint arXiv:2405.03978*.

Matte, C., Cunche, M., Rousseau, F., & Vanhoef, M. (2016). Defeating MAC Address Randomization Through Timing Attacks. *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, 15–20.

Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. *International conference on machine learning*, 8162–8171.

Nicodemus, K. K. (2011). On the stability and ranking of predictors from random forest variable importance measures. *Briefings in bioinformatics*, *12*(4), 369–373.

Paradeda, D. B., Junior, W. K., & Carlson, R. C. (2019). Bus passenger counts using wi-fi signals: Some cautionary findings. *Transportes*, *27*(3), 115–130.

Park, J.-g., Charrow, B., Curtis, D., Battat, J., Minkov, E., Hicks, J., Teller, S., & Ledlie, J. (2010). Growing an organic indoor location system. *Proceedings of the 8th international conference on Mobile systems, applications, and services*, 271–284.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*.

Pattanusorn, W., Nilkhamhang, I., Kittipiyakul, S., Ekkachai, K., & Takahashi, A. (2016). Passenger estimation system using wi-fi probe request. *2016 7th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, 67–72.

Pérez-Hernández, A., Barreras-Martín, M. N., Becerra, J. A., Madero-Ayora, M. J., & Aguilera, P. (2024). De-randomization of mac addresses using fingerprints and rssi with ml for wi-fi analytics [Accepted for publication in IEEE Access]. *IEEE Access*.

Pew Research Center. (2024, January 31). Mobile fact sheet. Retrieved November 9, 2024, from https://www.pewresearch.org/internet/fact-sheet/mobile/

Pintor, L., & Atzori, L. (2022a). Analysis of wi-fi probe requests towards information element fingerprinting. *GLOBECOM 2022-2022 IEEE Global Communications Conference*, 3857–3862.

Pintor, L., & Atzori, L. (2022b). A dataset of labelled device wi-fi probe requests for mac address de-randomization. *Computer Networks*, *205*, 108783.

Radford, A. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Robyns, P., Bonné, B., Quax, P., & Lamotte, W. (2015). *Crawdad hasselt/glimps2015*. IEEE Dataport.

Simončič, A., Mohorčič, M., Mohorčič, M., & Hrovat, A. (2023). Non-intrusive privacy-preserving approach for presence monitoring based on wifi probe requests. *Sensors*, *23*(5), 2588.

Singh, U., Determe, J.-F., Horlin, F., & De Doncker, P. (2021). Crowd monitoring: State-of-the-art and future directions. *IETE Technical Review*, *38*(6), 578–594.

Song, J.-y., Guo, Y.-p., & Wang, B. (2018). The parameter configuration method of dbscan clustering algorithm. *2018 5th International Conference on Systems and Informatics (ICSAI)*, 1062–1070.

Statista. (2024, August 5). *Market share of apple iphone smartphone sales worldwide 2007-2024*. Retrieved September 23, 2024, from https://www.statista.com/statistics/216459/global-market-share-of-apple-iphone/

Tan, J., & Chan, S.-H. G. (2021). Efficient association of wi-fi probe requests under mac address randomization. *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, 1–10.

Tirelli, T., & Pessani, D. (2011). Importance of feature selection in decision-tree and artificial-neural-network ecological applications. alburnus alburnus alborella: A practical example. *Ecological Informatics*, *6*(5), 309–315.

Torfi, A., & Fox, E. A. (2020). Corgan: Correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records. *The thirty-third international flairs conference*.

Torfi, A., Fox, E. A., & Reddy, C. K. (2022). Differentially private synthetic medical data generation using convolutional gans. *Information Sciences*, *586*, 485–500.

Truda, G. (2023). Generating tabular datasets under differential privacy. *arXiv preprint arXiv:2308.14784*.

Uras, M., Cossu, R., Ferrara, E., Bagdasar, O., Liotta, A., & Atzori, L. (2020). Wifi probes sniffing: An artificial intelligence based approach for mac addresses de-randomization. *2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 1–6.

Uras, M., Ferrara, E., Cossu, R., Liotta, A., & Atzori, L. (2022). Mac address de-randomization for wifi device counting: Combining temporal-and content-based fingerprints. *Computer Networks*, *218*, 109393.

Vanhoef, M., Matte, C., Cunche, M., Cardoso, L. S., & Piessens, F. (2016). Why mac address randomization is not enough: An analysis of wi-fi network discovery mechanisms. *Proceedings of the 11th ACM on Asia conference on computer and communications security*, 413–424.

Vega-Barbas, M., Álvarez-Campana, M., Rivera, D., Sanz, M., & Berrocal, J. (2021). Aforos: A low-cost wi-fi-based monitoring system for estimating occupancy of public spaces. *Sensors*, *21*(11), 3863.

Wang, J., Gao, Q., Pan, M., & Fang, Y. (2018). Device-free wireless sensing: Challenges, opportunities, and applications. *IEEE network*, *32*(2), 132–137.

Wang, Y., Yang, J., Chen, Y., Liu, H., Gruteser, M., & Martin, R. P. (2014). Tracking human queues using single-point signal monitoring. *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, 42–54.

Xi, W., Zhao, J., Li, X.-Y., Zhao, K., Tang, S., Liu, X., & Jiang, Z. (2014). Electronic frog eye: Counting crowd using wifi. *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, 361–369.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in neural information processing systems*, *32*.

Yaik, O. B., Wai, K. Z., Tan, I. K., & Sheng, O. B. (2016). Measuring the accuracy of crowd counting using wi-fi probe-request-frame counting technique. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *8*(2), 79–81.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., & Yang, M.-H. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, *56*(4), 1–39.

Yuan, Y., Zhao, J., Qiu, C., & Xi, W. (2013). Estimating crowd density in an rf-based dynamic environment. *IEEE Sensors Journal*, *13*(10), 3837–3845.

Zou, H., Zhou, Y., Jiang, H., Chien, S.-C., Xie, L., & Spanos, C. J. (2018). Winlight: A wifi-based occupancy-driven lighting control system for smart building. *Energy and Buildings*, *158*, 924–938.

Zou, H., & Xue, L. (2018). A selective overview of sparse principal component analysis. *Proceedings of the IEEE*, *106*(8), 1311–1320.