Machine Learning for RF Cloud in Proximity Detection and Intelligent Spectrum Management: An Empirical Study

by

Zhuoran Su

A Dissertation

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

in

Electrical and Computer Engineering

by

Feb 2024

APPROVED:

Professor Kaveh Pahlavan, Major Dissertation Advisor

Professor Emmanuel Agu, Co-Advisor

Professor Bashima Islam, Co-Advisor

Professor Donald Richard Brown, Head of Department

Abstract

The RF cloud is a shared Radio Frequency (RF) database supported by various wireless technologies. The RF cloud, constituted by billions of Internet of Things (IoT) devices, Wi-Fi access points, and cellular towers, forms a complex ecosystem transmitting vast quantities of data via RF signals. Beyond the exchange of information packets, the RF cloud encompasses detailed signal characteristics such as the Receive Signal Strength Indicator (RSSI), Time of Arrival (TOA), and Channel State Information (CSI). This rich data environment enables the development of intelligent applications in cyberspace, ranging from enhancing wireless positioning accuracy, advancing motion detection capabilities, building security frameworks, to optimizing spectrum sharing techniques.

Emerging techniques often outperform traditional estimation methods because they leverage Machine Learning (ML) and Deep Learning (DL) to utilize RF signal characteristics more effectively. However, incorporating ML and DL into RF Cloud applications introduces challenges related to complexity, interpretability, adaptability, and the cost of implementation.

In this dissertation, we focus on two novel problems of RF cloud applications, proximity detection and mobility support for spectrum sharing, to mitigate these concerns with theoretical analysis and empirical study validation.

1) We evaluate the efficacy of Machine Learning (ML) algorithms versus classical estimation theory in addressing the proximity detection challenge, in both theoretical foundations and empirical performance evaluation. Utilizing the Mitre Range Angle Structured (MRAS) Private Automated Contact Tracing (PACT) dataset, we compare classical estimation methods and ML algorithms—Support Vector Machines, Random Forest, and Gradient Boosted Machines—on Bluetooth Low-Energy (BLE) Received Signal Strength Indicator (RSSI) data. We contrast the complexity, availability, and precision of RSSI-based BLE and Time of Arrival (TOA)-based Ultra-Wide Band (UWB) technologies for proximity detection during epidemics. We present a detailed performance evaluation for the theoretical precision and confidence limits, and the performance on a new empirical dataset collected from diverse environments. Our findings reveal the UWB TOA algorithm's superior precision and confidence, albeit with the advantages of BLE RSSI in smartphone integration and minimal computational requirements.

2) We conduct an empirical study on RF interference intensity (RII) in licensed and unlicensed bands along a downtown Worcester, MA route to demonstrate the spatiotemporal RII behavior, utilizing a mobile spectrum monitoring system. This study informs an ML approach to predict channel availability for vehicular network spectrum access with fourteen RII features. We present the accuracy in predicting channel availability and regenerating RII as a validation of the proposed theoretical foundations, highlighting the potential of explainable ML in intelligent spectrum access and mobility support in the context of next-generation wireless networks.

This comprehensive analysis showcases the potential of utilizing classical theoretical foundations for enhancing the interpretability, adaptability, and efficiency of ML in proximity detection and spectrum management, offering significant implications for

public health safety measures during epidemics and for the wireless communication efficiency of next-generation wireless networks.

Acknowledgments

I would like to express my sincere gratitude to everyone who has supported me during the course of this dissertation journey. First and foremost, my utmost appreciation goes to my research advisor, Professor Kaveh Pahlavan, for his invaluable support, guidance, and patience throughout my Ph.D. journey. It has been a great honor to work with Professor Pahlavan in the CWINS lab. In addition to the skills and knowledge I have learned from him, his words of insight and sagacity encouraged me and held me in good stead. The life lessons and attitude imparted to me by him are priceless treasures that will illuminate my path in the years to come.

I must extend my deepest appreciation to Professor Emmanuel Agu and Professor Bashima Islam for their exceptional guidance and support as co-advisors. Their insightful suggestions and mentorship have been crucial in honing my research and presentation skills, profoundly influencing both the direction and essence of my dissertation.

I would also like to thank the members of my dissertation committee, Professor Seyed Zekavat, Professor Donald Brown, and Professor Alexander Wyglinski, for their insightful comments and suggestions, which have been invaluable to the completion of this dissertation. Their willingness to give their time so generously has been very much appreciated.

A special note of appreciation is reserved for Dr. Yishuang Geng and Dr. Nader Moayeri, who served as external examiners during my dissertation defense. Their thorough evaluations and constructive critiques have been invaluable to this dissertation. Their willingness to participate and contribute their expertise is deeply appreciated.

I would like to extend my heartfelt thanks to my colleagues, Dr. Julang Ying, Pengyu Zhang, Haowen (John) Wei, Ziheng (Leo) Li, Zhenyuan Lei, and Shiyu Cheng in CWINS Lab, and my friends Dr. Zhouchi Li, Dr. Jianan Li, Fei Li, and Xiao Zhang, for creating a warm environment and providing invaluable contributions to my life and research. When I felt anxious, worried, hesitated, or confused, they always encouraged me and spread their energy so that I can carry on to fulfill my dreams.

Words can not express my gratefulness to my dear parents for their care and support and their important responsibility for enabling me to reach this point in my life. They have provided unwavering support and encouragement from the beginning. Their faith in me has been a constant source of motivation and strength.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The RF cloud refers to an invisible, pervasive network formed by wireless technologies, including billions of IoT devices, Wi-Fi access points, and cell towers [1]. It transmits a large amount of data through RF signals, encompassing not just the content of information packets but also detailed signal characteristics like RSSI, Time of Arrival TOA, and CSI. This rich data environment enables the development of intelligent applications in cyberspace, ranging from wireless positioning, motion detection, and security measures to spectrum sharing, by leveraging the unique propagation traits of RF signals. These cyberspace applications involve collecting RF data from multiple sources, processing this data to identify patterns or anomalies, and then applying algorithms to extract meaningful insights for each application, such as triangulating a device's location, recognizing specific gestures, or verifying a device's authenticity based on its unique RF signature.

The first part of this dissertation investigates the effectiveness of UWB TOA and BLE RSSI technologies, two promising technologies widely leveraged in enabling accurate proximity detection, a critical component in managing public health crises like COVID-19. The urgency of accurate social distancing measures has underscored the need for reliable technology that can be widely implemented in smartphones. This work aims to establish a theoretical framework for comparing UWB TOA and BLE RSSI, supported by empirical data. It seeks to determine the more precise and feasible technology for measuring social distance in real-world scenarios, considering factors

like accuracy, complexity, and device compatibility.

The second part of this dissertation exploits the mobility support for spectrum sharing. The rapid evolution of autonomous vehicles and the Internet of Everything (IoE) has intensified the need for innovative spectrum management in wireless communications, particularly for cellular 6G and Wi-Fi 7. Recent years have seen cellular operators like AT&T, T-Mobile, and Verizon extensively deploy mid-band 5G radios at 2.6GHz and 3.5 – 6GHz frequencies, overlapping with Wi-Fi 6 and 7 operations. Concurrently, negotiations are underway to share the 3.2 – 3.8GHz spectrum between the wireless industry, the Pentagon, and other licensees. However, current models for mid-band spectrum sharing fail to account for the complexities introduced by mobility, risking interference and service disruptions. Therefore, empirical research into the statistical characteristics of mid-band RF interference is crucial. Such studies will present the development of intelligent spectrum management strategies, ensuring efficient and effective use of these bands in an increasingly connected world.

This dissertation aims to bridge the gap between traditional theoretical foundations and rapidly evolving applications in two key areas of wireless communication: RF-based proximity detection and mobility support for dynamic spectrum sharing. It seeks to advance understanding in these fields by integrating classical theoretical models with the latest technological advancements. This work is poised to offer novel insights and practical solutions, contributing to public health strategies and technological advancements in proximity detection during epidemics, and the efficient utilization and management of spectrum resources in the era of 6G, Wi-Fi 7, and beyond.

1.1 Contribution

The contribution of this dissertation in two major parts is listed as follows:

- PART I: Chapter 3 and 4.
  - We collected an empirical proximity detection dataset following the structure of the existing PACT dataset with empirical experiments at eight distances

in three flat environments and one non-flat environment encompassing both Line of Sight (LOS) and Obstructed-LOS (OLOS) situations. We analyzed the effect of changing the environment and the effects of various postures (eight angles) of the person carrying the sensor and four on-body locations of the sensor.

- The extracted spatial, temporal, frequency-domain, and statistical features from BLE RSSI data are effective for proximity detection, delivering comparable performance across different scenarios.
- We presented a comparative analysis between classical estimation theory and advanced feature-based ML algorithms, offering a novel approach to evaluating the accuracy of BLE RSSI-based proximity classification.
- We presented an extensive comparison between a ML-based approach using BLE RSSI data and a two-way-ranging technique using UWB TOA measurements, highlighting the strengths and limitations of these approaches in real-world settings.
- We introduced a novel performance metric for smartphone proximity detection, focusing on the confidence in accurately estimating a 6 ft social distance. This metric is defined by the probability of correctly predicting whether an estimated distance is less than or greater than 6 ft, based on the actual measured distance being within or out of the 6 ft threshold.
- We derived bounds on the confidence based on traditional estimation theories for both BLE RSSI and UWB TOA, as functions of distance. These theoretical bounds are computed using the Cramér-Rao Lower Bound (CRLB), a method previously employed in distance-ranging error estimation. This theoretical foundation offers a practical framework to evaluate the suitability of techniques for proximity detection challenges.
- We compared theoretical achievable bounds and validated this theoretical basis using empirical data gathered in practical scenarios, 1) for social distance range estimation using BLE RSSI with the empirical results obtained

using two theoretical RSSI behavior models and three ML classification algorithms, 2) for UWB TOA obtained using a two-way ranging algorithm with BLE RSSI-based approach using a Gradient Boosted Machines (GBM) classification algorithm.

- PART II: Chapter 5 and 6.
  - We presented the need to investigate RF Interference Intensity (RII) in mid-band frequencies, specifically targeting both the 1.9GHz licensed and 2.4GHz unlicensed bands. This investigation is crucial for understanding and managing interference in these commonly used frequency ranges.
  - We introduced a vehicular interference monitoring system to collect interference data across licensed and unlicensed bands within the 1.9 to 2.5GHz spectrum. The empirical measurement was conducted during rush hours in downtown Worcester, MA, resulting in 20 individual test drives with approximately 100,000 spectrum snapshots in the crowded mid-band. The spectrum monitoring system and dataset offer insights into understanding the temporal and spatial behavior of RII.
  - We extracted and analyzed fourteen statistical features of interference from RII, categorized into four classes: intensity, correlation properties, spectrum occupancy, and Doppler spectrum. These characteristics describe RF signal propagation and are extensively utilized in areas such as wireless localization and RF-based motion or behavior detection.
  - By applying a Support Vector Machine (SVM) for predictability analysis, we identified the significance of features for both licensed and unlicensed bands in forecasting channel availability. Our findings indicate that channel occupancy is the key factor in predictability for licensed bands, while average interference intensity plays a major role for unlicensed bands. We achieved a predictability rate of 96% for unlicensed bands and 85% for licensed bands.

– We applied frequently used regression ML models: LR, KNN, RF, and XGB, inputting GPS coordinates to regenerate RII. This approach provides us with a tool to analyze the RII behavior in spatial aspects with irregularly sampled data. We found the average RII is subject to a fixed average received power reflecting the architecture of the fixed civil infrastructure surrounding a location. The RII regeneration model achieved an average RMSE of 5.57 dB for licensed bands and 3.41 dB for unlicensed bands.

– We conducted a comparative evaluation of ML algorithms, including SVM, RF, and XGB, using 14 RII features to model the predictability of channel availability across various times and days. The study emphasizes the effectiveness of ML in modeling channel availability with RII features, demonstrating that additional data collection beyond certain thresholds does not significantly enhance predictability.

1.2 Dissertation Outline

The remainder of this dissertation is organized as follows: Chapter 2 introduces the background of this dissertation, including the traditional modeling method and the development of empirical study for RF proximity detection and dynamic spectrum sharing.

Chapter 3 presented an extensive comparison between classical estimation, ML algorithms, and TOA-based ranging methods, highlighting the strengths and limitations of each technology in proximity detection during epidemics. To compare different approaches, we established the CRLB-based theoretical bounds of the proposed performance metric, the confidence on detecting the social distance of 6 ft. We conducted rigorous experimental validation under varied conditions to reinforce the theoretical models with practical observations.

Chapter 4 The proposed theoretical foundations underscore the potential of both BLE and UWB technologies in enhancing real-time proximity detection, offering insights for effective public health during epidemics.

Chapter 5 presented a thorough empirical study of Radio Frequency Interference Intensity (RII) in crucial mid-band frequencies, including both 1.9GHz licensed and Wi-Fi 2.4GHz unlicensed bands. We built the mobile spectrum monitoring system with an ultra-wideband programmable spectrum analyzer and a laptop integrated with GPS. We conducted data collection in a typical urban area with a pre-selected route. We identified and classified fourteen distinct RII features into four major categories. We established the theoretical foundations for RII modeling.

Chapter 6 We utilized ML to refine RII reconstruction and channel availability prediction, which is essential for enhancing mobility support in intelligent spectrum management, especially within vehicular networks.

Chapter 7 proposed the conclusions of this dissertation and the discussion of future works.

Chapter 2

Background

2.1 ML for RF Cloud Applications

Due to more effective utilization of RF signal characteristics, emerging approaches aided by Machine Learning (ML) and Deep Learning (DL) generally surpass traditional estimation algorithms. However, utilizing ML and DL in RF Cloud applications raises concerns, including:

- **Complexity and Interpretability:** Some of the most powerful ML and DL models are often considered "black boxes" due to their complexity, making it difficult to understand how they arrive at specific decisions. This lack of interpretability can be a significant issue in applications where transparency and trust are crucial.
- **Adaptability:** ML models are designed for specific tasks and can struggle with tasks that are even slightly different from what they were trained on. This lack of adaptability can limit their usefulness in dynamic environments.
- **Cost:** Developing, training, and deploying ML models can be expensive, requiring specialized hardware and software, as well as skilled personnel. The ongoing costs of updating and maintaining these models can also be significant.

In this chapter, we delve into the theoretical foundations of the RF cloud, detailing

the RF signal characteristics and the utilization of spectrum resources. Additionally, we provide a visionary perspective on the integration of ML applications within the RF cloud. This exploration investigates how ML algorithms can utilize the extensive data from wireless technologies to improve application performance, also examining the feasibility of employing classical theories to guide ML models. This work seeks to blend traditional theoretical insights with modern ML capabilities, aiming to mitigate concerns about applying ML to applications within the RF cloud framework by:

- Integrating RF propagation theory, time and frequency domain characteristics of RF signals, classical estimation theory, and feature importance evaluation to define performance bounds that address the concerns of complexity and interpretability in ML-based RF cloud applications. This comprehensive approach enhances the understanding and management of ML models, offering clearer insights into their decision-making processes.

- Optimizing the use of spectrum resources so the RF cloud can accommodate more devices, expanding its database across different scenarios. This strategy not only enhances network efficiency but also addresses the adaptability challenge in ML, as a larger and more diverse dataset improves the learning and generalization capabilities of ML models in dynamic environments.

- Conducting empirical evaluations to determine the minimum dataset size necessary for training effective ML models, thereby reducing the cost of compiling an extensive real-world database. This strategy refines the RF cloud database by preserving only crucial data for future ML models, thereby optimizing storage solutions and improving data utilization efficiency across diverse ML application scenarios.

2.2 Proximity Detection for RF Cloud

Proximity detection is a technology designed to sense and identify the presence of objects or individuals within a certain distance without any physical contact. It relies

on various sensing methods, including RF, infrared, and ultrasonic, and supports the applications including:

- **Security and Access Control:** RF proximity detection is used in security systems for access control to buildings, rooms, or secure areas. By detecting an authorized RF ID, the system can automatically unlock doors or trigger alarms if unauthorized access is attempted.

- **Retail and Customer Experience:** In retail, RF proximity detection can enhance customer experience through personalized marketing and promotions. When a customer with an RF-enabled device, like a smartphone, is near a product or section, the system can push relevant offers or information to their device, benefiting both merchants and customers.

- **Healthcare Tracking:** In healthcare facilities, RF tags can be used to monitor the location of patients, especially those with special needs, dementia, or infants, to ensure their safety and well-being. It also helps in managing the location of medical equipment and staff.

- **Epidemic Control:** The transmission risk of diseases like COVID-19 significantly increases when individuals are closer than 6 feet for over 15 minutes, a situation described as "Too Close for Too Long" (TCTL). RF proximity detection technology offers a method to identify high-risk encounters and thereby help mitigate the spread of epidemics through timely and accurate contact tracing without raising significant privacy concerns or disrupting daily activities.

Traditional Bluetooth's long scan duration limited its use in proximity detection, whereas Wi-Fi RSSI typically samples between 0.25 Hz and 2 Hz. BLE, however, supports a faster advertising rate of 10 Hz [3], significantly enhancing scan efficiency for proximity applications. Supported by most smart devices, BLE offers advantages such as compactness, lightness, affordability, and energy efficiency, making it a highly suitable and promising technology for RF proximity detection applications. In the broadcasting of BLE communication, advertising packets can be sent out in one way

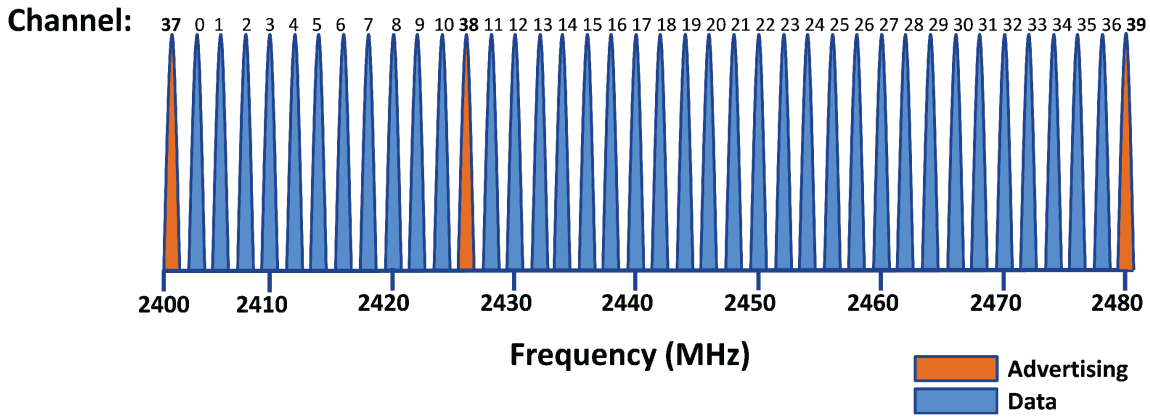


Figure 2-1: BLE channel division. (Source: [2])

on three advertising channels (See Fig. 2-1), from a single device to any scanning or receiving device in the range of coverage. The advertiser periodically sends advertising packets to any device able to receive them [2]. A scanner continuously scans available advertising packets from an advertiser (See Fig. 2-2). Broadcasting packets in BLE technology serve two primary functions [2]: firstly, they distribute advertising packets to applications that don't require a continuous connection, enhancing efficiency. Secondly, they enable a master device to send connectable advertising packets, facilitating the discovery of slave devices for potential connection. In proximity detection, BLE's connectionless broadcasting enables the measurement of the RSSI, offering a key factor for estimating the proximity between devices without requiring a direct connection. Previous studies like [4] have explored how variations in advertising and scanning intervals affect power consumption and the accuracy of proximity detection.

Classical RSSI-based proximity detection usually uses the path-loss model for estimating the distance between two devices. The path-loss is defined as the reduction of the power of an electromagnetic wave as it propagates through space. A BLE path-loss model models RSSI as a function of the distance between the transmitter and receiver. The standard path-loss model includes three primary parameters: reference RSSI at one meter, distance-power gradient, and standard deviation of shadow fading. However, its simplicity limits accuracy in complex environments, prompting researchers to develop refined path-loss models to improve proximity detection's effectiveness

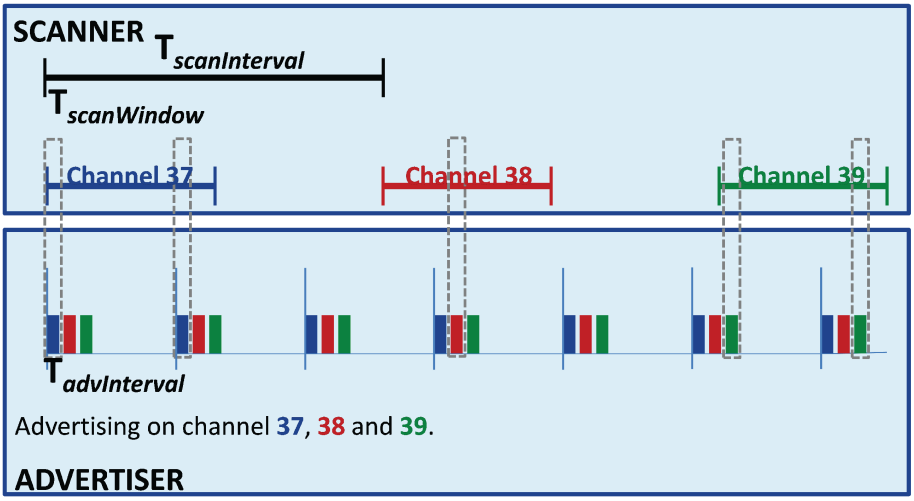


Figure 2-2: BLE advertising and scanning mechanism. (Source: [2])

in response to increasing demands [5, 6, 7]. The performance of classical proximity detection is bounded by CRLB, a fundamental concept in statistics and estimation theory that provides a lower bound on the variance of unbiased parameter estimators. Its application is extensive in the RF sensing domain, particularly in RF-based localization technologies such as Wi-Fi [8, 9, 10], UWB [11], Cellular[12, 13], and BLE [14].

However, analyzing the reliability of BLE RSSI-based ranging presents a complex challenge, primarily due to the significant variability in RSSI signals. This variability arises from the complexity of multipath indoor radio propagation, which leads to signal attenuation, fading, and interference from other devices operating within the unlicensed 2.4 GHz ISM (Industrial, Scientific, and Medical) bands. Previous studies have undertaken real-world measurements and characterized proximity detection using BLE RSSI across various scenarios, indicating the inherent complexities in accurately measuring proximity in dynamic indoor environments.

Proximity detection typically classifies individuals as 'non-proximity' or 'proximity' based on a distance threshold. Utilizing powerful ML classifiers can significantly enhance this process, offering a sophisticated improvement over classical methods. Classical range estimation focuses solely on the spatial characteristics of RSSI, neglecting its temporal and frequency domain properties. By considering these additional

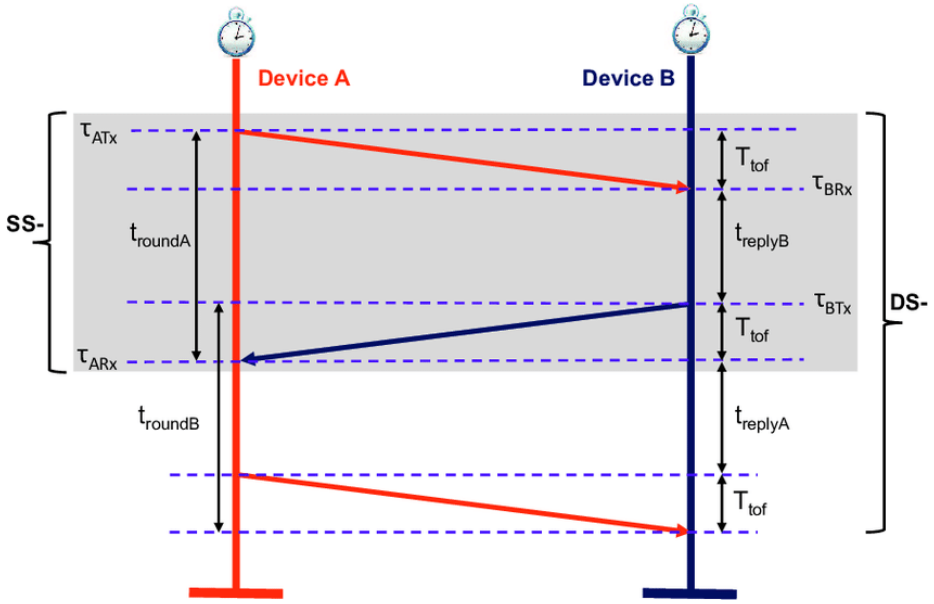


Figure 2-3: UWB Two-way Ranging. (Source: [24])

signal features, ML can achieve more accurate range and distance estimations, thereby enhancing confidence in the results. Model-based ML approaches, particularly in RSSI-based motion and gesture detection [15, 16, 17] , have proven effective in reducing algorithm complexity and computational time, offering a more sophisticated yet efficient solution compared to traditional methods. Table 2.1 shows some model-based ML applications in related works.

Table 2.1: ML Applications in Related Works

Related Work	RF Signal	Application	Classifier
[4]	BLE RSSI	Proximity Detection	Decision Tree
[18]	BLE RSSI	Proximity Detection	SVM
[19]	BLE RSSI	Proximity Detection	Linear Regression
[20]	BLE RSSI	Proximity Detection	1D CNN
[21]	BLE RSSI	Proximity Detection	KNN, CART, RF
[22]	Wi-Fi RSSI	Proximity Detection	Naive Bayes
[23]	BLE RSSI	Positioning	RF, XGB, SVM, KNN, DT

UWB technology, recognized for its precision in TOA measurements, emerges as an alternative to the traditional RSSI-based BLE ranging systems. UWB's advantage lies in its ability to deliver precise distance measurements without relying on complex machine learning algorithms that necessitate extensive training and substantial amounts of labeled data. The availability of cost-effective UWB devices and the sup-

port for UWB positioning in existing 5G cellular networks underscores its potential for widespread adoption. Two-way ranging is a commonly used UWB ranging algorithm without synchronization among all the mobile nodes [25], which meets the low-energy consumption goal for the IEEE 802.15.4 standards. Two-way ranging measures the round trip TOA by calculating the time difference from when device A sends a ranging message to when it receives a response from device B (See Fig. 2-3). This method accurately determines the distance between the two devices by timing the message exchange. In ranging applications, UWB TOA achieves an impressive accuracy of up to 10cm. In contrast, classical BLE RSSI ranging achieves accuracy within a few meters. However, compared with BLE, UWB TOA proximity detection requires higher power consumption for its operation, and it can be more susceptible to interference from obstacles, leading to reduced accuracy in cluttered environments.

2.3 Channel Availability Analysis for RF Cloud

The RF signal radiated from billions of active and passive wireless devices for communications and positioning creates an RF cloud, causing co- and cross-channel interference among all wireless devices. Empirical understanding and modeling of the stochastic behavior of RF cloud interference are needed to manage and regulate the precious spectrum sufficiently, a unique natural resource shared among all wireless devices of the universe. The importance of intelligent spectrum management and regulation was first experimented with worldwide by the FCC's innovative release of Instrument Scientific and Medical (ISM) unlicensed bands in May 1985 [26]. The unlicensed ISM bands became the ground for the growth of popular Wi-Fi and Bluetooth technologies and pioneering commercial application of spread spectrum, orthogonal frequency division multiplexing (OFDM), multiple-input multiple-output (MIMO) antenna systems, and mmWave technologies, which later became adopted for 4 – 7G cellular networks operating in super expensive licensed bands [27]. Today, the exponential growth of IoT devices, the demand for smartphones for higher data rates, and their (IoT and smartphones) ability to be ubiquitous and mobile have heightened the need for a

new paradigm in spectrum management and regulation for sharing and managing spectrum resources in cellular 6G, Wi-Fi 7, and beyond. However, none of the current characterizations of wireless spectrum sharing take the effect of mobility into account.

The need for monitoring the mid-band (1 – 6GHz), where the most popular licensed 5G bands and ISM unlicensed bands reside, is gaining more importance as cellular operators such as AT&T, T-Mobile, and Verizon have installed mid-band 5G radios and popular Wi-Fi 6 – 7 operate in the same frequency bands instead of the mmWaves (24-40GHz) [28]. T-Mobile covers around 260 million customers in the US with its mid-band 5G network and expects to increase to 300 million. Similarly, Verizon covers 200 million in the mid-bands and expects to reach 250 million by the end of 2024. Therefore, mid-bands are becoming the favorite of today's wireless industry because RF propagation in these bands supports coverage of significant distances and passes through the building walls [29]. However, mid-bands are crowded with underutilized military and satellite bands, which can be shared with the wireless industry to facilitate the growth of 6G cellular, Wi-Fi 7, and beyond to maintain the US leadership in this vital industry for secure nationwide communications. We need a new paradigm for intelligent spectrum management and regulations by better understanding the interference in the mid-bands in the presence of mobility. That demands fundamental empirical RF interference propagation research in these bands to characterize the statistical behavior of interference from mobile and stationary devices from the standpoint of a mobile device.

Fig. 2-4 shows a simplified overview of the current infrastructure for wireless communications of mobile devices and the role of major emerging wireless technologies: 6G, Wi-Fi 7, BLE, and wired Ethernet. The top right corner of the figure shows a snapshot of spectrum utilization overlay of interference in 1.9 – 2.5GHz in a 20-minute drive in downtown Worcester, Massachusetts [31]. The 1.9 – 2.5GHz portion of the mid-band spectrum includes the most crowded 1.9GHz licensed bands for 5G and 2.4GHz ISM bands for Wi-Fi [32]. We observe two distinct segments of the bands: the highly utilized bands on the corners and the under-utilized bands between them. This simple observation raises three fundamental questions:

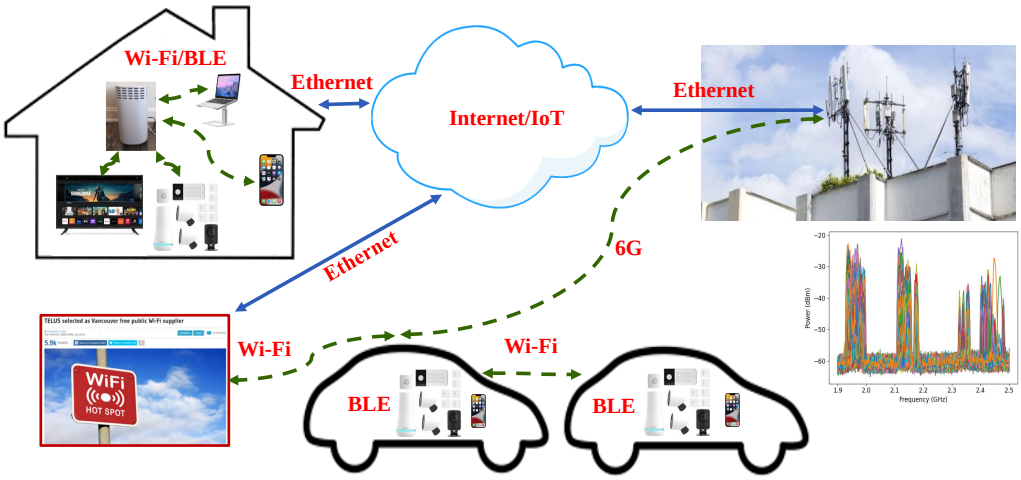


Figure 2-4: Simplified overview of the current infrastructure for wireless communications, the role of major emerging wireless technologies: 6G, WiFi, and BLE, and a snapshot of spectrum utilization overlay of interference in 1.9 – 2.5GHz in a 20 minutes drive in streets of downtown Worcester, MA [30].

- How does the statistical behavior of the interference differ in the highly utilized mid-bands for licensed 5G and unlicensed Wi-Fi communications in the presence of both stationary and mobile devices?
- What causes the under-utilization in the mid-band spectrum while the wireless communication industry demands more bands to support the need and growth of smartphones and IoT devices?
- How can we optimize the overall utilization to support the communication needs of the growing number of stationary and mobile devices in the existing mid-band?

The answers to these three questions are highly profound and deeply complex as they depend on the following:

- The evolution of frequency administration since the inception of FCC in 1934.
- The diverse need for spectrum management and regulation in active communications and passive radar monitoring,

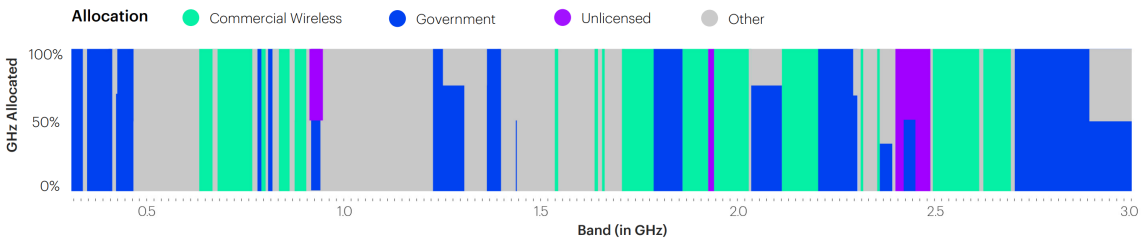


Figure 2-5: Mid Band Spectrum Allocation. (Source: [32])

- The diverse pattern of spectrum utilization in military and commerce for satellite and terrestrial applications,
- The unprecedented exponential growth of stationary and mobile wireless devices since the 1980s.

In the past decade, frequency administration agencies and the research community have initiated considerable efforts to address several aspects of these complex questions. In 2015, the FCC released the Citizen Broadband Radio Service (CBRS) bands at 3550 – 3700 GHz as a three-tiered access spectrum and a framework to accommodate commercial services to share the bands with the military counterpart. Then, a few yearlong disputes on L-bands (1.7 – 2.1GHz) among the global positioning system (GPS) community and wireless communications industry began [27], and the wireless communications industry is continually lobbying the Pentagon and White House on sharing other mid-band military segments in 3.2 – 3.8GHz.

The research community began studying the creation of spectrum fingerprints or radio maps in small areas surrounding a software-defined radio under "spectrum cartography" [33]. In this field, a series of fixed sensors surrounding the radio transmission antenna monitor the interference in irregular sample locations. Then, they interpolate the radio map of interference of the area surrounding the antenna, benefiting from different traditional and machine learning algorithms. More recently, the interference radio map of small areas surrounding a source antenna with the help of RFID monitoring devices spread around the source was interpreted as the digital twin of the interference map [34]. As this interference map works with the assumption of prior knowledge about the positioning of the RFIDs and their stationary range

of interferences [33, 34], they fail to capture the stochasticity of IoT devices in/on mobile vehicles or humans. Mobile monitoring beyond a single fixed radiation RF monitoring source [35] is required to create a radio interference map of a metropolitan area, where the variability of the devices in time is considered. For digital twining of the time-varying interference with sequential empirical samples of the spectrum taken with a mobile device, we need to resort to machine learning algorithms to reconstruct the pattern of interference in time, frequency, and space based on samples of empirical interference gathered from monitoring of the spectrum with a mobile device. As a result, the authors began studying the complexity, feasibility, and benefits of creating a digital twin of the interference in the mid-bands with empirical mobile monitoring of the spectrum by driving in the city streets [31, 35]. This scalable empirical mobile digital twining approach can create an interference radio map of large metropolitan areas and consequently cover the entire nation to demonstrate the real mid-band interference statistical behavior and bandwidth utilization.

These studies will also encourage wireless standardization organizations to include statistical interference models in their future recommendations to facilitate the comparative performance of emerging intelligent spectrum management algorithms. Besides, it will promote the industry to design RF cloud databases to cover all major cities to assist in discovering solutions to optimize mid-band utilization with innovative spectrum management and regulations. Scientific understanding of interference will help find solutions for spectrum sharing among the wireless communications industry and other industries such as positioning, astronomy, and agriculture. It will further enable these industries to resolve their long-time standing disputes, resulting in more comfort for the operation of billions of IoT devices and smartphone users to enjoy faster and more affordable wireless communications.

Chapter 3

PART I: Performance Evaluation of Proximity Detection Using BLE RSSI

The risk of COVID-19 transmission increases when an uninfected person is less than 6 ft from an infected person for longer than 15 minutes. Infectious disease experts working on the COVID-19 pandemic call this high-risk situation being Too Close for Too Long (TCTL). Consequently, the problem of detecting the TCTL situation in order to maintain appropriate social distance has attracted considerable attention recently. One of the most prominent TCTL detection ideas being explored involves utilizing the BLE RSSI to determine whether the owners of two smartphones are observing the acceptable social distance of 6 ft. However, using RSSI measurements to detect the TCTL situation is extremely challenging due to the significant signal variance caused by multipath fading in indoor radio channels, carrying the smartphone in different pockets or positions, and differences in smartphone manufacturer and type of the device. In this study, we utilize the Mitre Range Angle Structured (MRAS) Private Automated Contact Tracing (PACT) dataset to extensively evaluate the effectiveness of Machine Learning (ML) algorithms in comparison to classical estimation theory techniques to solve the TCTL problem. We provide a comparative performance evaluation of proximity classification accuracy and the corresponding confidence levels using classical estimation theory and a variety of ML algorithms. As the classical estimation method utilizes RSSI characteristics models, it is faster to compute, is

more explainable, and drives an analytical solution for the precision bounds proximity estimation. The ML algorithms, SVM, Random Forest, and GBM utilized thirteen spatial, time-domain, frequency-domain, and statistical features extracted from the BLE RSSI data to generate the same results as classical estimation algorithms. We show that ML algorithms can achieve 3.60%∼19.98% better precision, getting closer to achievable bounds for estimation.

3.1 Introduction

With the threat of COVID-19, a highly infectious virus, maintaining social distancing is an effective way to prevent infection. Specifically, the risk of COVID-19 transmission increases when an uninfected person is less than 6 ft from an infected person for longer than 15 minutes (also called Too Close for Too Long (TCTL)). If the list of people who are TCTL to each smartphone user can be detected and tracked passively, they can be notified if the smartphone user tests positive for COVID-19. Existing opportunistic RF positioning technologies can be used to track the infected smartphone user's daily motion trajectory. The owners of neighboring smart devices can then be notified so that they can maintain social distance or get tested if they are in found to have been TCTL. Although the tradeoff between the benefits of COVID-19 mitigation using contact tracing and the intrusion on users' privacy remains a difficult social-political problem, scientific research in this area has recently gained momentum. With its short range and low energy consumption, the use of the ubiquitous BLE signal has attracted significant attention. This led the Massachusetts Institute of Technology (MIT), Boston, MA, to lead the Private Automated Contact Tracing (PACT) consortium [36] to make available several high-quality BLE RSSI datasets, which were gathered in a variety of proximity scenarios. Their goal was to challenge the research and development community to discover a solution to this timely and important problem. The reliability analysis of RSSI-based BLE ranging is a complex problem because of the significant variance in the measured RSSI signal due to the complexity of the multipath indoor radio propagation causing extensive signal

attenuations, fading, and interference from other devices operating in unlicensed 2.4 GHz ISM bands. In prior work, real-world measurement studies and characterization of proximity detection using BLE RSSI have been conducted in various scenarios [37]. Some prior work has proposed approaches to improve RSSI-based proximity estimation by integrating data from other sensors, including light [38], accelerometer and gyroscope [39], and user-sensed motion [40]. Some other authors have incorporated information on the place type [41], user context [42], sensed crowd [43], social context [40], [44], social circles [45], indoor-outdoor detection [46] and place co-location [47]. A modified path loss model has also been utilized [48]. Beyond proximity, other authors used RSSI to estimate the mutual orientation between users [49] and the energy consumption of BLE RSSI proximity detection [4]. In this chapter, we present the results of our extensive comparative performance evaluation of classical estimation theory and Machine Learning (ML) algorithms for social distance estimation using the BLE RSSI data. We utilized the MITRE Corporation Structured Angle dataset of the PACT project to share generate results and make our observations from this experiment. We begin by describing the MITRE Range Angle Structured (MRAS) PACT dataset followed by a review of RSSI features that are useful for distance estimation. Then, we present the classical estimation theory, which facilitates faster proximity computation in a more logically explainable manner, and ML algorithms that can be used to estimate user proximity with all the RSSI features. Finally, we provide our quantitative comparative performance evaluation of traditional and ML algorithms to solve the social distance estimation problem using the BLE signal. For the classical estimation theory results, we present the method for computing the confidence associated with the distance estimated using the BLE RSSI behavior models. We also derive bounds on the confidence of range estimation using the CRLB. For the ML estimations, we classified thirteen spatial, time, frequency, and general statistical features of the BLE RSSI using three different algorithms: Support Vector Machine (SVM), Random Forest and GBM. The final outcome of this extensive study is the comparison of theoretical achievable bounds for social distance range estimation using BLE RSSI with the empirical results obtained using two theoretical

RSSI behavior models and three ML classification algorithms. The RF cloud around wireless devices present an opportunity for designing novel cyberspace applications. The RF cloud contains features of the signal that reflect the multipath characteristics of the environment at each location. As a device moves, these multipath characteristics change rapidly opening an opportunity for other devices to observe these variations in characteristics and relate them to a location-dependent cyberspace application [50], [51]. The PACT project is a new opportunistic cyberspace application focused on an opportunistic proximity check application benefiting from the RF cloud of the BLE. The Center for Wireless Information Network Studies (CWINS) at the Worcester Polytechnic Institute (WPI), Worcester, MA has previous engagement in the PACT project and is now exploring systematic research in this field. Short term, the proximity detection BLE RSSI application can be investigated. Longer term research could involve extending the BLE signal by including other sensors. We build on our prior RSSI based positioning and motion and gesture detection research [15, 16, 27] for the current time. In future, we are planning to extend BLE by incorporating other opportunistic wireless signals including those from Wi-Fi and Ultra-wideband devices to increase the precision of range estimation.

3.2 The Pact Proximity Datasets And Measurement Scenarios

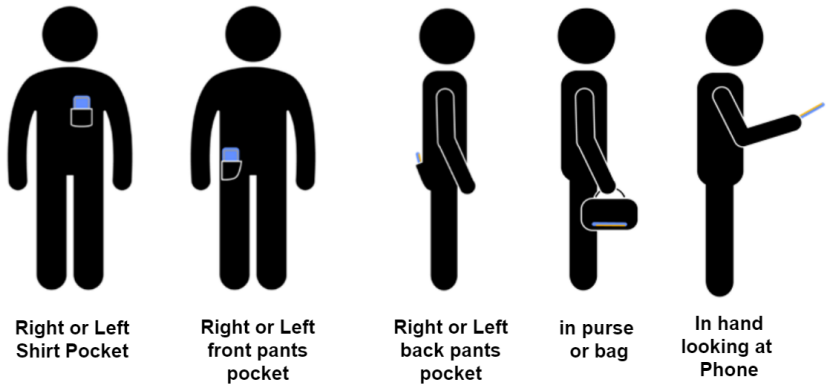
There are seven datasets made publicly available by the PACT consortium [36]. Compared with the other datasets, MRAS dataset is well documented. Moreover, it contains measurements in various testing scenarios at different distances, which are relevant to our study goals of comparing the performance of classical and ML algorithms using various features extracted from BLE RSSI measurements. The MRAS dataset also contains different environment and tester pose settings. Environment settings specify the properties of testing area, such as the room size and the tester's location in the room. Tester settings defines the way devices are used by testers, in

which way they hold the smartphones, and the poses of testers. Fig. 3-1 shows the location of device and 8 selected relative distances between testers for the MRAS database. Fig. 3-1a shows the five scenarios emulating real life scenarios for position of the smartphone: in hand, in purse, in shirt pocket, in front pants pocket, and in the back pants pocket. Fig. 3-1b shows the BLE RSSI measurement scenarios for short range of operation of up to 15 ft. The eight stationary locations for measurements begin at 3 ft, are increased at intervals, and end at 15 ft. The distances are identified with respect to a person who holds the smartphone with BLE beacons. The RSSI measurement data are collected by another person (a receiver) positioned at the eight labeled distances. In each test location identified in Fig. 3-1b, 5-10 seconds measurements of BLE RSSI containing 300-400 samples of the RSSI are measured:

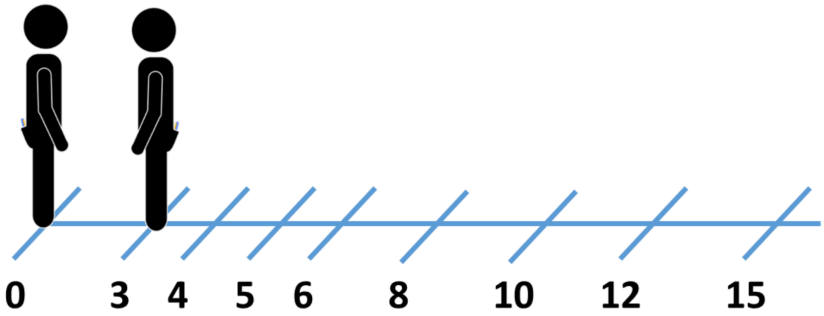
$$s(k) = \text{RSSI}(t)\big|_{t=kT_s}; k = 1, ..., N, \tag{3.1}$$

where N is the number of samples in a location and T_s is the time interval between two adjacent RSSI measurement samples.

Table 3.1 summarizes various operation scenarios reported in the MRAS measurement database for collecting 300-400 samples of RSSI in each stationary dataset. The first two rows capture variations in the room size and locations of testers in the area. The detailed room size setting is not provided in the dataset, but a description of the scenarios is available. For example, the entrance to the bathroom of apartment is defined as small room, the kitchen is defined as medium room, and large living room is defined as large room [37]. The next two rows identify the types of the smartphone and the location of the smartphone on the tester body. The last row identifies testers pose that is either "sit" or "stand" at the marked location. These datasets were collected using three versions of Range-Angle Collection Protocol [36]: Short, Mid and Full. The Full protocol consists of 40 datasets with RSSI measurements at eight different distances shown in Fig. 3-1b and we used these datasets for our performance evaluation for different proximity algorithms. We did not include the Short and Mid versions, which had only two different distances of 3 ft and 8 ft and did not offer



(a)



(b)

Figure 3-1: PACT measurement scenario for the MITRE Range Angle Structured dataset, (a) five scenarios for location of the smartphone, (b) eight distances for measurements of the RSSI data base. (Source: PACT website)

adequate diversity in measurement distances.

In our selected MRAS dataset, multipath fading characteristics and variation in the environment caused close to 30 dB difference in the values of RSSI in each measurement set and up to 30 dB variations in average RSSI in individual sets. The RSSI measurements in each location, defined by (Eq. 3.1), are post-processed before feeding them into classical and ML algorithms respectively. In the classical estimation algorithms, the training data for RSSI behavior at each location Eq. 3.1, are averaged at each distance. The average RSSI at each location is then defined by:

$$P_r = \frac{1}{N} \sum_{k=1}^{N} s(k). \tag{3.2}$$

This data post-processing for classical algorithms associates a single average RSSI

Table 3.1: Scenarios For MITRE-Range-Angle-Structured Dataset

Scenarios	Settings
Description of the areas	Small Room, Medium Room, Large Room, Hallway, Outside
Location in the area	Center open, Center congested, Near wall open, Near wall congested
Type of Device	iPhone XR, iPhone 11, iPhone 8, iPhone6, iPhone 7, iPhone XS, iPhone 6 Plus
Device location on-body	In hand, In purse, Shirt pocket, Front pants pocket, Rear pants pocket
Tester Pose	Standing, Sitting

measurement, P_r, to each location.

For ML techniques, the RSSI measurements at each location is grouped in overlapping windows of RSSI measurement vectors of length L, whose elements are defined by

$$y(n) = \{s(k+n); k = 0, 1, ..L-1\}; n = 1, ..., N-L. \tag{3.3}$$

This processing associates an $N-L$ set of L dimensional vectors with each location. We utilized these post-processed RSSI data to perform our comparative performance evaluations for various classical and ML algorithms. Our performance criterion is the confidence in the decision made by the algorithm for the task of detecting the social distance of 6 ft using BLE RSSI measurements gathered using a smartphone at a given location.

3.3 Features of RSSI Short Range Fading

Motion in the environment affects RF propagation in multipath indoor and urban areas and causes fading in the measured RSSI, which seriously challenges the precision of RSSI-based ranging [52]. The channel impulse response for two wireless devices communicating with a range r in a multipath, indoor, or urban area with N-paths, is represented by [53]:

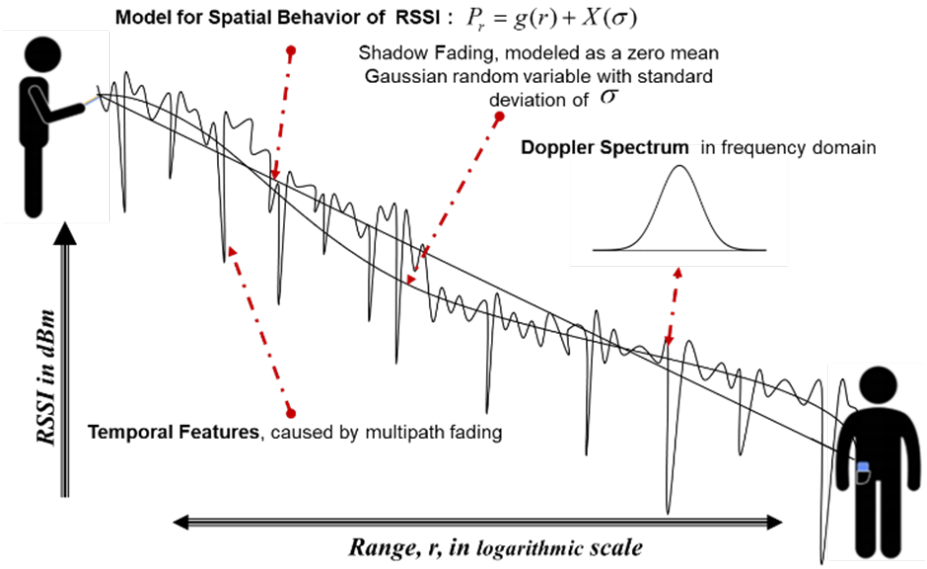


Figure 3-2: Variation of the received power in dB as a function of the logarithmic distance between the transmitter and the receiver and how we approach to model them for different purposes.

$$h_r(\alpha_i; \tau_i; \theta_i) = \sum_{i=1}^{N} \alpha_i e^{j\theta_i} \delta(t - \tau_i),$$

where (α_i, τ_i, θ_i) are the magnitude, time of arrival, phase, and DOA of the i-th path.

$$\begin{aligned} \text{RSSI}(t) = \sum_{i=1}^{N} |h_r(\alpha_i; \tau_i; \theta_i)|^2 &= \left| \sum_{i=1}^{N} \alpha_i e^{j\theta_i} \delta(t - \tau_i) \right|^2 \\ &= \left| \sum_{i=1}^{N} \alpha_i e^{j\theta_i} \right|^2 \end{aligned}$$

We can easily measure this RSSI from a transmitting wireless device without any synchronization with the source. Multipath arrival of the signal in indoor and urban areas, where the applications discussed in this chapter operate, causes extensive fluctuations of the amplitude of the received signal in time. Fig. 3-2 illustrates the variation of the amplitude in dBm (RSSI) as a function of the logarithmic distance between the transmitter and the receiver, r, as a receiver moves away from a transmitter. This figure also shows how we approach the modeling of these variations of the RSSI

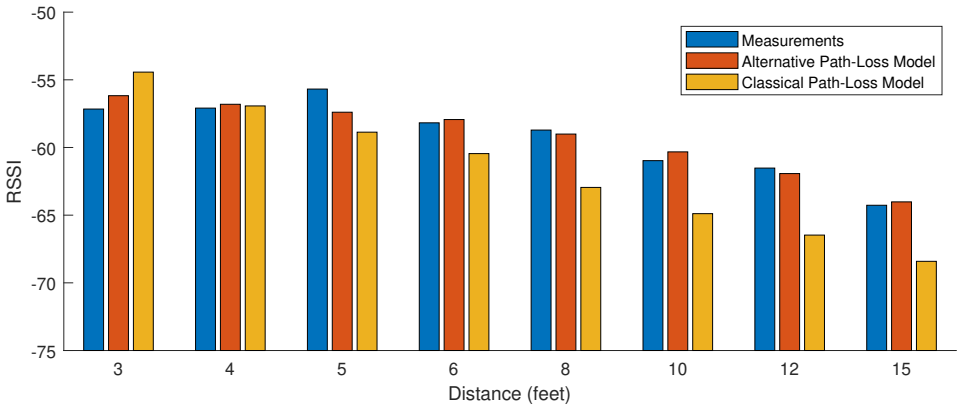


Figure 3-3: RSSI Estimation using traditional RSSI behavior model and the alternative BLE specific model for a set of MRAS RSSI data.

for different applications. The instantaneous RSSI in a multipath environment always varies over time and with small local changes in distance or movement of objects located around the transmitter and the receiver antennas. The average of the RSSI decays as the distance increases and we use an RSSI model to predict the average received RSSI for calculating the coverage and interference of wireless networks and for RSSI-based cyberspace applications [53]. The distribution function of temporal changes in the signal is modeled with a few distribution functions to analyze the error rate of wireless modems. The Fourier Transform of these changes is referred to as the Doppler Spectrum, which reflects the speed of movement of objects or the device in the environment of operation. As the objects scattered in the area or the wireless devices move in the environment or we change the frequency of operation, characteristics of the multipath features fluctuate drastically and cause fading in measured RSSI. In the wireless communication literature, this phenomenon is discussed under temporal, frequency-selective, and spatial fading [54]. In this body of knowledge, RSSI features in space, time, and frequency are modelled using a few physical parameters that can be measured. These features can be utilized to improve the reliability of estimates generated by RSSI-base range estimation techniques.

3.3.1 RSSI Spatial Features

In classical RSSI-base ranging we use the average RSSI in dBm for calculating the distance between an antenna and a device, r. The traditional method to model how the RSSI is related to the distance from the transmitter is to use linear regression and least square estimation to calculate the parameters of the model using empirical data [52]. The traditional statistical linear regression model for the spatial behavior of RSSI in dBm is:

$$\text{RSSI} : P_r = P_0 - 10\alpha \log_{10}(r) + X(\sigma), \tag{3.4}$$

in which r is the distance, X is a Gaussian random variable with variance σ, representing the shadow fading effects, α is the distance power gradient of the environment, and P_0 is the RSSI at a reference distance from the transmitter. Shadow fading represents variations of the RSSI from the linear regression line in dB caused by objects shadowing radio propagation paths between the transmitter and the receiver. We can use the traditional Least Square (LS) method of statistical modeling to estimate the RSSI spatial behavior model parameters, (P_0, α, σ), using measured RSSI data in different scenarios provided by the PACT (section 3.2) [52].

An alternative model for short range BLE RSSI is also reported in the literature [55], which we tested on the PACT database. Based on empirical measurements of BLE, this model suggests that the RSSI has an additional sinusoidal component:

$$\text{RSSI} : P_r = P_0 - A\cos\left(\frac{2\pi r}{\lambda}\right) + 10\alpha \log(r) + X(\sigma). \tag{3.5}$$

Therefore, in addition to traditional model parameters, (P_0, α, σ), this model has two new parameters (A, λ), A is the amplitude scale of the sinusoidal part and λ is its spatial wavelength, which we can also estimate using the LS algorithm. In this chapter, we expand the effective range of this alternative BLE specific model to about 4.5 m (15 ft) and compare the results with produced using the traditional linear regression model described by Eq. 3.4. Fig. 3-3 shows the difference between classical linear regression RSSI model described by Eq. 3.4 and the alternative BLE-specific

RSSI model described by Eq. 3.5 using a set of Mitre Corporation MRAS PACT data in eight distances. The BLE-specific RSSI model on the average provides a slightly better fit to data for predicting the measured RSSI values. In section V.A we compare the performance of classical range estimation algorithms when we used the MRAS RSSI database with both RSSI models.

3.3.2 RSSI Features in Time Domain

In RSSI ranging, we measure a sequence of RSSI values in time at each location and estimate the range of these collective measurements. Because the person measuring the BLE RSSI at a location has slight body motions and the objects in the environment also move, the measured RSSI fluctuations in amplitude even when the transmitter and the receiver are held in specific locations. In the multipath RF propagation literature, these fluctuations are referred to as short range multipath fading and their characteristics are modelled for performance evaluation of wireless communications techniques [53]. By using an ML system designed for ranging, we can benefit from the physical parameters of these fading models as features for training the algorithm. Traditionally in data science we use the mathematical statistics as features of these signals, the new features extracted from our understanding of the behavior of RF propagation in multipath environment can potentially improve the performance of the system. These features have been demonstrated to be very instrumental in RSSI-based gesture and motion detection [15, 16]. In this study, we evaluate the effectiveness of thirteen features of RSSI in estimating the social distancing between smartphone users utilizing BLE RSSI. Table 3.2 is a summary of the radio propagation and statistical features that we have selected to train the ML algorithms in our study. We calculated these features for all BLE RSSI measurements, $y(n)$, defined by Eq. 3.3 to form a vector that is used to train the ML algorithms in the section. 3.4.2. We have divide them into time-domain (section 3.3.2), frequency domain (section 3.3.3), and traditional statistical features (section 3.3.4). The time domain RSSI features benefit from classical radio propagation modeling of these fluctuations, which includes fading rate, average fade duration, coherence time, and shape of the fading distribution,

which we describe in the remaining subsection of this section.

Table 3.2: Summary of Thirteen Features of The RSSI For Training Machine Learning Algorithms

Feature Name	Formula	Description
Time Domain Features		
Fade Duration	$\tau(\rho) = \frac{e^{\rho^2} - 1}{\sqrt{2\pi}\rho B_{\text{rms}}}$ (6a)	The average fading duration
Level Crossing Rate	$N(\rho) = \sqrt{2\pi}\rho B_{\text{rms}} e^{-\rho^2}$ (6b)	The average number of downward crossings of 2dB threshold per second
50% Coherence Time	$R_{\text{yy}}(m) = \frac{\sum_{n=1}^{L-m} \{y(n) - m_y\}\{y^*(n+m) - m_y^*\}}{\lvert r(n)\rvert \times \lvert r(n+m)\rvert}$ (6c)	Representative of speed of fluctuations in signal in a location
Rayleigh Parameter	$f_{\text{ray}}(r) = \frac{y}{\sigma^2}\exp(-\frac{y^2}{2\sigma^2}), \text{y} \geq 0$ (6d)	Representative of speed of fluctuations in signal in a location
Frequency Domain Features		
Energy	$E = \int D(\lambda) d\lambda$ (7a)	The energy of Doppler spectrum
Laplacian Best Fit	$D(\lambda) = \frac{a}{1 + b\lambda^2}$ (7b)	b is the bell shape fit parameter
Rms Doppler Spread	$B_{\text{rms}} = \sqrt{\frac{\int \lambda^2 D(\lambda) d\lambda}{\int D(\lambda) d\lambda}}$ (7c)	The second central moment of Doppler spectrum
General Statistical Features		
Mean	$\frac{1}{L}\sum_{n=1}^{L} y(n)$ (8a)	The central value of a window
Y_{p2p}	$\{y(n)\}\lvert_{\max} - \{y(n)\}\lvert_{\min}$ (8b)	Peak-to-peak changes in RSSI of one location
Standard Deviation	$\sqrt{\frac{1}{L}\sum_{n=1}^{L}(y(n) - \bar{y})^2}$ (8c)	A measure of variations about the central value
Interquartile range	$\text{IQR} = Q_3 - Q_1$ (8d)	A measure of variability, based on dividing a data set into quartiles
Skewness	$E[(\frac{(y(n) - \bar{y})}{\sigma})^3]$ (8e)	A measure of the asymmetry about the mean value of data set
Kurtosis	$E[(\frac{(y(n) - \bar{y})}{\sigma})^4]$ (8f)	A measure of whether the data set is heavy-tailed or light-tailed relative to a normal distribution

Crossing Rate and Duration of the Fades

Fig. 3-4a shows a sample of the fluctuations of RSSI sequences, $y(n)$, over time caused by small-scale temporal fading characteristics of the channel. Two interesting features of short-range fading for RSSI measurements are the fading rate and fading durations. We can calculate the rate of fluctuation of the envelope of the RSSI caused by multipath fading with these parameters. It is well known that in Rayleigh fading channels the threshold crossing rate, $N(\rho)$, and average duration of fade, $\tau(\rho)$, are related to the rms Doppler spread B_{rms} (see section 3.3). Fig. 3-4a shows the definition of the fade rate and the duration of the fade as well as equations relating them together on a sample of MRAS-measured data. Defining the normalized crossing threshold as, $\rho = A/A_{\text{rms}}$, in which A and A_{rms} are the threshold level and RMS amplitude of the RSSI, respectively, these relations are given by Eqs. 6a and 6b [53] in the top two rows of the time-domain features in Table 3.2. Given a set of data in a location (Fig. 3-4a) we find the fading rate and fade duration for the signal and use these values as features to train the ML algorithm.

Coherence Time

Another feature to determine the speed of fluctuations in values of RSSI is the coherence time of the signal [53]. The coherence time is the width of the correlation function of the samples of the RSSI in a location. For L samples of RSSI defined by sequence $y(n)$, Eq. 3.3, the normalized autocorrelation function is given by (6c) in Table 3.2, in which

$$\begin{cases} m_y = \frac{1}{L}\sum_{n=1}^{L} y(n) \\ r(n) = \sqrt{\frac{1}{L}\sum_{n=1}^{L}[y(n) - m_y]^2} \end{cases}.$$

As shown in Fig. 3-4b, the value of the plot at the intersection with the 50% line is used as the coherence time. We can use the coherence time as another time-domain feature of the small-scale fading in a location to train an ML algorithm for ranging.

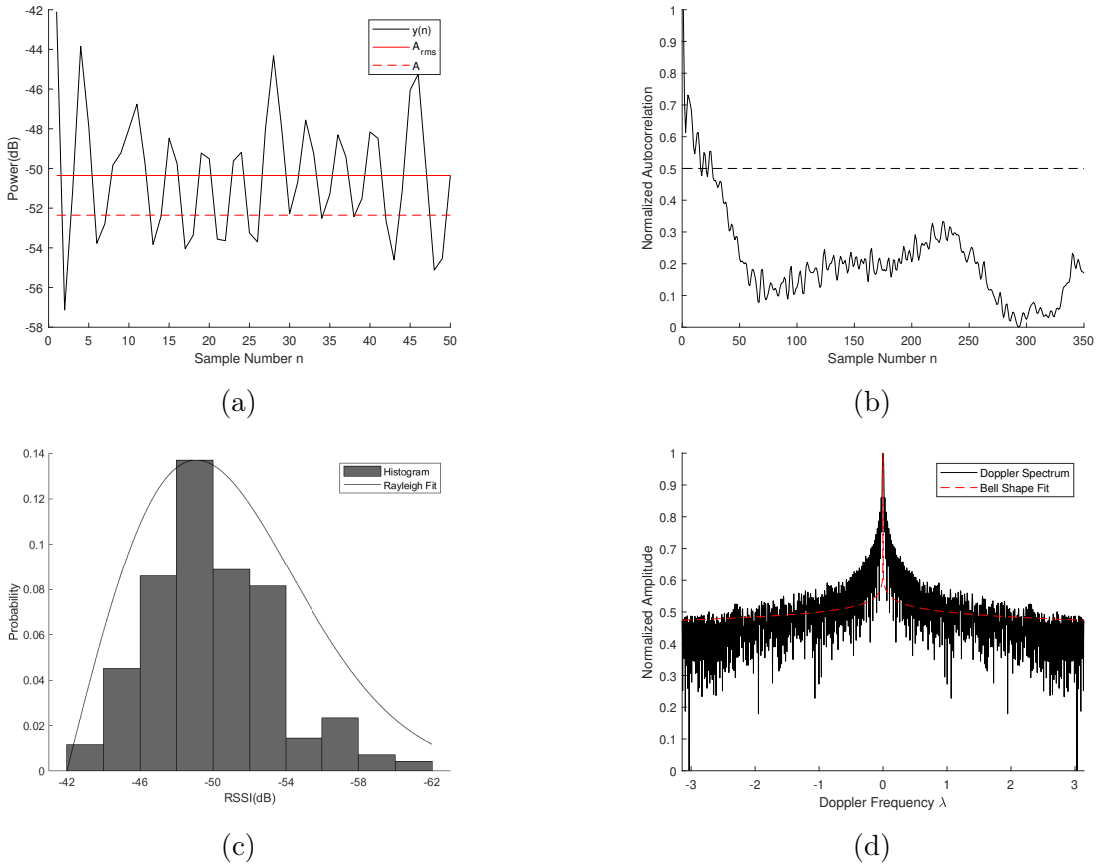


Figure 3-4: Summary of time- and frequency-domain features extracted from samples of MRAS RSSI data a) level crossing rate and fade duration and their relation to RMS Doppler spectrum for a sample, b) 50% Coherence Time using the autocorrelation function, c) Rayleigh fit for the distribution of amplitude fluctuation, and d) Laplacian fit to Doppler Spectrum.

Shape of Fading Distribution

Multipath fading results in fluctuations in the signal amplitude because of the addition of signals with different phases arriving from multiple paths. This phase difference is caused by the signals traveling different distances along multiple arriving paths. Since the phase of the arriving signals changes rapidly, the received signal amplitude undergoes rapid fluctuation that is often modeled as a random variable with a Rayleigh distribution given by Eq. 6d in Table 3.2 [53], where σ_{ray}, is the standard deviation of the Rayleigh distribution function. To model these fluctuations, we can generate a histogram of the amplitude of the received signal in time and fit it to a Rayleigh

distribution function. Fig. 3-4c shows a sample Rayleigh fit to the MRAS data. By fitting the Rayleigh distributions to a set of MRAS data, $y(n)$, we can determine σ_{ray}, the parameters defining the distribution of that data and then associate that parameter as a feature per-location for training the ML algorithm.

3.3.3 RSSI Feature in Frequency Domain

Traditionally, RSSI of RF signals have been used in Doppler radars and for GPS signals to measure the speed to correct the estimated range for moving objects. In recent years, using RSSI signals in time and in frequency with intelligent algorithms has attracted attention in new emerging fields for big data such as gesture [56] and motion detection [15, 16]. Parameters associated with the Doppler spectrum can also be used for range estimation using ML algorithms. Doppler spectrum, $D(\lambda)$, is the magnitude square of the Fourier transform of variation of the signal in the time domain. For a discrete sequence, $y(n)$, using the Fast Fourier Transform (FFT), we can calculate samples of the Doppler spectrum function as:

$$\begin{cases} Y(k) = \text{FFT}\left[y(n)\right] \\ D(k) = D(\lambda)|_{\lambda=k/T_s} = |Y(k)|^2. \end{cases}$$

We extracted three features from the empirical Doppler spectrum obtained from a set of data at a location, $y(n)$. The middle part of Table 3.2 summarizes these frequency domain features. These parameters are the energy of the signal, E, defined by Eq. 7a in Table 3.2, the RMS Doppler spread, defined by Eq. 7b, and the shape of the Doppler spectrum in indoor areas, defined by Eq. 7c. The RMS Doppler spread is the normalized second moment of the $y(n)$, reflecting the speed of motions in the environment. We used this parameter in the previous section to calculate the fade rate and duration as well. According to the IEEE 802.11 standard organization model for the RSSI, the Doppler spectrum shape follows a Laplacian distribution in indoor areas, shown in Eq. 7b of Table 3.2 [54]. By normalizing amplitude to one, $a = 1$. Fitting with empirical data results in a single parameter, b, which reflects the speed of

BLE RSSI fluctuations at a given location. Fig. 3-4d shows a sample of the FFT and the best fit Laplacian function for the MRAS RSSI data. In this way, we extract three parameters to represent the frequency domain characteristics of each data sequence y(n), $\{E,\ B_{\text{rms}},\ b\}$, and use these three features along with the four time-domain features and the following six statistical features for training ML algorithms using the MRAS data.

3.3.4 Statistical Features of RSSI

The features of the signal that we referred to so far in this section have physical meanings that we borrowed from the multipath RF propagation literature [54]. The set of RSSI data in a location can also be treated as a mathematical sequence from which we calculate statistical features that are then fed as inputs to ML algorithms. In this study, we have included six common statistical features of the RSSI samples at a location, shown in the last six columns of Table 3.2, for training the ML algorithms. The two top rows of statistical features are mean and peak-to-peak changes of the RSSI sample at a location. Another traditional RSSI feature is the Interquartile Range (IQR), which shows the spread of RSSI. Skewness and Kurtosis are parameters depicting the shape of RSSI distribution.

3.4 Proximity Detection Algorithms

The objective of this study is to investigate the accuracy of classical estimation theory algorithms for COVID-19 proximity detection and compare their performance with the results obtained using ML algorithms. Classical algorithms use the empirical measurements to model the behavior of RSSI with Eq. 3.4 and Eq. 3.5, then calculate the confidence of the estimation based on the parameters of these models. This approach enables faster computation in a more logically explainable manner and it also enables us to calculate the CRLB of the performance achievable by any algorithm. ML algorithms benefit from all the spatial, time, frequency, and traditional statistical features of the RSSI shown in Table 3.2, solving the same problem and providing their

confidence level on these estimates. In the remainder of this section, we describe the details of these two classes of algorithms that we have used in this study.

3.4.1 Classical Estimation Algorithms

Classical estimation theory provides methods for modeling, estimating, and calculating the performance bounds of an estimator. In the classical estimation theory terminology, estimation of the range using the RSSI, P_r, defined by Eq. 3.2, is referred to as estimation of a single parameter, the range r, using observation of the function of the parameter, $g(r)$, in additive Gaussian noise, the shadow fading $X(\sigma)$,

$$O : P_r = g(r) + X(\sigma). \tag{3.6a}$$

In our problem, we have a traditional RSSI linear regression model, Eq. 3.4, and its alternative BLE-specific model, Eq. 3.5 [57]:

$$\begin{cases} g_1(r) = P_0 - 10\alpha \log_{10}(r) \\ g_2(r) = P_0 - A\cos(2\pi d/\lambda) + 10\alpha \log(r). \end{cases} \tag{3.6b}$$

When we establish the model, the classical estimation theory provides us with tools for systematic estimate of the range and the analysis of the accuracy of the estimation. Given an RSSI value, we can estimate the distance and calculate the confidence in the accuracy of that estimation in observing the social distance. In addition, classical estimation theory provides tools for the calculation of the variance of the estimate using CRLB on the accuracy of a single measurement and optimal confidence expected from estimation using any algorithm.

Empirical Range Estimation and Confidence

In classical estimation theory, the optimal estimate of the range, $\hat{r}$, for an average RSSI measurement of a device, P_r, defined in Eq. 3.2 taken at a specific range, r, is found by solving:

$$\hat{r} = g^{-1}(O) = g^{-1}(P_r). \tag{3.7a}$$

For a traditional linear regressive model, Eq. 3.4, we have a closed-form answer for the problem:

$$\hat{r} = g^{-1}(P_r) = 10^{-\frac{P_r - P_0}{10\alpha}}. \tag{3.7b}$$

For the alternate BLE-specific model, Eq. 3.5, we find the numerical solution to

$$P_r = P_0 - A\cos(2\pi r/\lambda) + 10\alpha \log(\hat{r}). \tag{3.7c}$$

If the estimated range is less than or equal to the admissible social distance of 6 ft, given that the device was also within the 6 ft range, or when the estimated range is more than 6 ft and the device range is also more than 6 ft, we are confident that the algorithm works properly. Therefore, the confidence on the estimate of the classical algorithms for BLE RSSI measurements at a given distance r is calculated from [52]:

$$\begin{aligned} \gamma(r) &= \Pr\left\{[\hat{r} \leq 6/P_r \leq P_6] \cap [\hat{r} > 6/P_r > P_6]\right\} \\ &= 1 - \frac{1}{2}\text{erfc}\left(\frac{|P_6 - P_r|}{\sqrt{2}\sigma}\right), \end{aligned} \tag{3.7d}$$

where $P_6 = g(6)$ is the expected RSSI measured at 6 ft distance obtained from RSSI behavior model and σ is the standard deviation of the shadow fading. For our empirical analysis of the classical estimation methods, we have used Eq. 3.7d to calculate the confidence on any set of test data. We will explain these in more detail in the introduction to section V and Fig. 3-6.

Bounds on Ranging and Confidence

Another power tool from classical estimation theory is the CRLB, which is a bound on the variance of the ranging error, and it is the inverse of the Fisher Information Matrix (FIM) of the dataset [52]:

$$\sigma^2(r) = \text{CRLB} \geq \text{FIM}^{-1} = \frac{\left[g'(r)\right]^2}{\sigma^2}, \tag{3.8a}$$

σ, is the standard deviation of shadow fading at the location, and $g(r)$ is the function representing the model for the two models of RSSI behavior we studied. Substituting the two models given in Eq. 3.6b in to Eq. 3.8a, we have:

$$\sigma_1(r) = \sqrt{\text{CRLB}} \geq \frac{\ln 10}{\sqrt{N}10}\frac{\sigma}{\alpha}r \tag{3.8b}$$

$$\sigma_2(r) = \sqrt{\text{CRLB}} \geq \frac{\sigma}{\sqrt{N}(\frac{2\pi A}{\lambda}\sin(\frac{2\pi r}{\lambda}) + \frac{10\alpha}{\ln 10 \cdot \alpha})}. \tag{3.8c}$$

Equation Eq. 3.8a provides bounds on the variance of the estimate using the two RSSI behavior models at a given location. In our COVID-19 social distancing problem, we are interested in measuring our confidence in a distance estimate. The confidence is the probability that the estimated range is less than or equal to the admissible social distance of 6 ft given that the device was, in fact, within the 6 ft range; or the probability that the estimated range is more than 6 ft and the device range is also more than 6 ft. If we assume that distance measurement error is a zero mean Gaussian random variable, we can model the distance estimate by

$$\hat{r} = r + \eta\left[\sigma(r)\right],$$

where $\eta\left[\sigma(r)\right]$ is the measurement noise calculated from the CRLB of Eq. 3.8a. Therefore, given a distance, r, and assuming that the distance measurement error is a zero mean Gaussian random variable, we can calculate our confidence in making a measurement in one side of 6 ft and estimating it on the correct side from:

$$\begin{aligned}\gamma(r) &= \Pr\left\{[\hat{r} \leq 6/r \leq 6] \cap [\hat{r} > 6/r > 6]\right\} \\ &= 1 - \frac{1}{2}\text{erfc}\left(\frac{|6-r|}{\sqrt{2}\sigma(r)}\right),\end{aligned} \tag{3.9}$$

where $\gamma(r)$ is the bound on the confidence of estimating the distance using RSSI observed at a distance r, and $\sigma(r)$ is the variance of estimation defined by Eq. 3.8a. Eqs. 3.8b and 3.8c demonstrate the bounds on estimating a location from classical

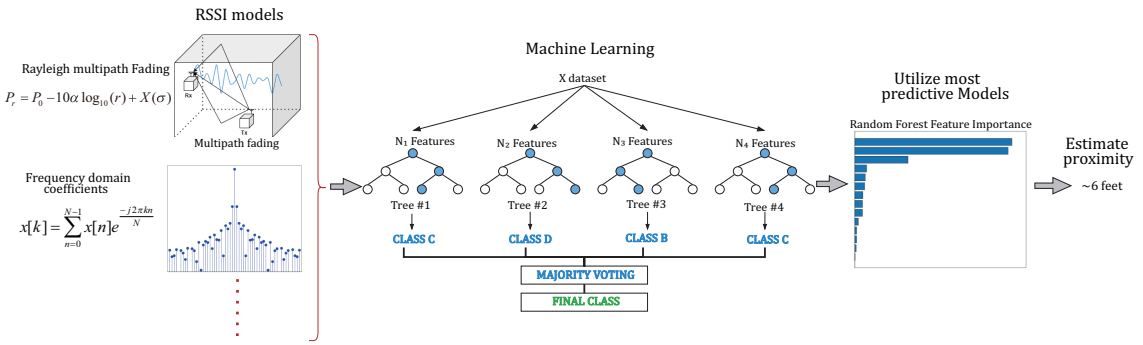


Figure 3-5: Overview of hybrid model-based ML proximity detection approach.

models for RSSI behavior, an ideal expected confidence, while Eq. 3.2 demonstrates the confidence on actual measurements. Eq. 3.7d an algorithm for calculation of confidence from the empirical data. In section V, we use these equations to calculate the bounds on the confidence of range estimation as well as the confidence of the RSSI measurement-based ranging using empirical data (see Fig. 3-6 in that section).

3.4.2 Machine Learning Algorithms

In classical range estimation, only the spatial characteristics of RSSI measured values are utilized without consideration of its temporary characteristics in time and in the frequency domain. ML algorithms can benefit from other features of the signal, providing a better estimate of the range and distance and improving the confidence of the result of estimation. In this chapter, we intend to compare these two approaches quantitatively. In practice, the ML approach is more computationally sophisticated, and the classical approach is more analytically complex but simpler to implement. The classical approach relies on modeling, which enables faster computation in a more logically explainable manner. Moreover, classical approaches generalize better to new, previously unseen scenarios. While prior work has typically used either the classical or ML approaches, we explore combining both methods, creating a third hybrid approach that uses classical models and parameters as inputs to the ML algorithms, facilitating model-based ML algorithms. In a sense, this is the best of both worlds, integrating the temporal characteristics in time and frequency as well as model parameters (see Fig. 3-5). Model-based ML has been shown to be effective in RSSI-based motion and

gesture detections to reduce the complexity and computational time of the algorithms [15, 16, 17]; in this chapter, we have examined them for the proximity range estimation for COVID-19 social distancing with BLE signals.

In the remainder of this section, we review the three ML algorithms that we have considered in this study and used in our comparative performance evaluation among classical: Random Forest, Gradient Boosted Machines, and Support Vector Machines.

Random Forest

Random Forest is an ML classification algorithm that is an ensemble of K classifiers $M_i,\ldots M_K$, where each classifier is a decision tree created using a different sub-sample of the entire dataset [17]. The final classification is obtained by majority voting of the K decision trees M_1 $M_2,\ldots M_K$. For a new test point $\boldsymbol{x}$, the class predicted by the Random Forest model $\boldsymbol{M}^K$ using majority voting is:

$$\boldsymbol{M}^K(\boldsymbol{x}) = \arg\max_{c_j} \{v_j | j = 1, ...k\},$$

where v_j is the number of trees created from the different dataset sub-samples, which predict the class of $\boldsymbol{x}$ as c_j. That is,

$$v_j(\boldsymbol{x}) = |\left\{\boldsymbol{M}^K = c_j | t = 1, ...K\right\}.$$

Gradient Boosted Machines

We explored classification using XGBoost, a high-performance implementation of GBM also called Gradient Boosted Trees [58]. The GBM classification model is an ensemble model that uses K additive functions to predict the output:

$$\hat{y} = \phi(\boldsymbol{x}_i) = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i), f_k \in F,$$

where F is the space of regression trees created from different subsets of the input dataset. To learn the set of functions utilized by the model, the following regularized

objective is minimized

$$L(\phi) = \sum_i l(\hat{y}, y_i) + \sum_k \Omega(f_k),$$

where l is a differentiable convex loss function, which measures the difference between the prediction $\hat{y}$ and target y_i, and

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||\boldsymbol{w}_{\text{score}}||^2,$$

in which T is the number of leaves in the tree, γ and λ are regularization parameters, and $\boldsymbol{w}_{\text{score}}$ is the score of corresponding leaves.

Support Vector Machines(SVM)

SVM is a ML classification algorithm that tries to discover a hyperplane that maximizes the margin between the target classes in feature space [59] and it is based on the theory of maximum linear discriminants. For two classes to be classified, SVM finds peripheral data points in each class that are closest to the other class (called support vectors). For a dataset D with n points $\boldsymbol{x}_i$ in a d-dimensional space, a hyperplane function $h(\boldsymbol{x})$ can be defined as

$$h(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = w_1 x_1 + w_2 x_2 + ... + w_d x_d + b,$$

where $\boldsymbol{w}$ is the weight vector. Overall, n points, the margin of the linear classifier can be defined as the minimum distance of a point from the separating hyperplane given as:

$$\delta^* = \min_{\boldsymbol{x}_i} \left\{ \frac{y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)}{||\boldsymbol{w}||} \right\}.$$

The SVM classifier finds the optimal hyperplane dividing the two classes by solving the minimization problem with objective function:

$$\min_{\boldsymbol{w} i b} \left\{ \frac{||\boldsymbol{w}||^2}{2} \right\},$$

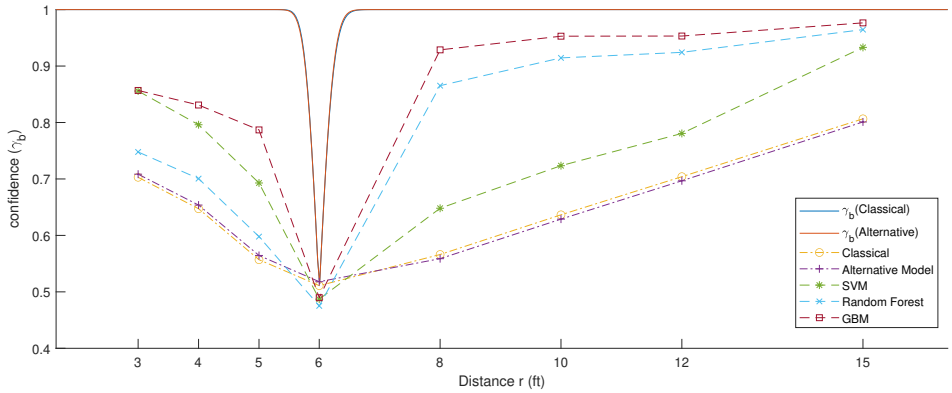


Figure 3-6: Bounds on confidence on estimation as a function of distance for MRAS RSSI database (top lines) versus performance of classical and alternative RSSI behavior modeling range estimation as well as SVM, Random Forest, and GBM ML algorithms.

with linear constraints:

$$h(\boldsymbol{x}) = y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, \forall \boldsymbol{x}_i \in D.$$

Then, the class of a new point $\boldsymbol{z}$, is predicted as:

$$\hat{y} = \text{sign}(h(\boldsymbol{z})) = \text{sign}(\boldsymbol{w}^{\text{T}}\boldsymbol{z} + b).$$

Confidence Calculaion

To compare the performance of ML Classifiers with that of the classical estimation theory, it is necessary to calculate the confidence of the classifications generated by the SVM, Random Forest, and GBM classifiers. To calculate this confidence, we split the training data into 8 groups based on the distance between the transmitter and the receiver. Then we calculated the confidence at each distance from:

$$\gamma(r) = \Pr\left\{[\hat{r} \leq 6/r \leq 6] \cap [\hat{r} > 6/r > 6]\right\}, \tag{3.10}$$

which represents the probability of estimating the distance to be at the correct side of 6 ft social distance barrier. These results from ML algorithms are comparable with the

results obtained from Eq. 3.10 for the two classical approaches from the traditional linear regression and the BLE-specific models for RSSI behavior. These experimental results are then compared with the bounds on confidence in Eq. 3.2 obtained from the calculation of the CRLB.

3.5 Performance of Ranging with BLE Signals

In this section, we present the results of applying the algorithms described in section 3.4, using the BLE RSSI features described in section 3.3, on the PACT MRAS dataset that was described in section 3.2. The basic performance criterion we use is the confidence on correctly estimating the social distance of 6 ft between smartphone users. That is, the probability of correctly detecting the distance of a device relative to the 6 ft threshold using RSSI measurements. We begin by calculating bounds on the confidence of RSSI-based ranging (section 3.4.1); then we present the results of classical estimation theory ranging (section 3.4.1), and finally results from ML algorithms (section 3.4.2).

As the first step, the training data is used for calculating the parameters of the traditional RSSI behavior models using the LS algorithm. The model parameters are then used to calculate the CRLB and then the bounds on the confidence. Then, the test data and model parameters are used for calculating results from the test data to determine the confidence on classical methods with the two RSSI behavior models. Finally, we trained the three ML algorithms using the training data and found the confidence on estimation for the test data to compare with results of classical methods as well as bounds on the performance. We refer to the results of calculating the bounds on the test data using spatial RSSI behavior models as classical performance evaluation, and we present those results first.

All the test scenarios shown in Table 3.1 are for the Line-Of-Sight (LOS) propagation condition without any object obstructing the LOS path between the transmitter and the receiver, and the maximum distance is 15 feet (Fig. 3-1b). We begin by presenting the results for our traditional RSSI behavior model described by Eq. 3.4.

We have 40 sets of data for five scenarios in eight distances. We utilized 75% of the data to estimate the parameters of the RSSI behavior model with the LS algorithm. The model is trained for LS estimation with the RSSI averaged at each distance shown in the measurement scenario of Fig. 3-1b. The three parameters of the traditional RSSI regressive model, power at the reference point, distance-power gradient, and standard deviation of shadow fading, were $P_0 = -54.94$ dBm, $\alpha = 1.74$ and $\sigma = 3.78$ dB, respectively. We repeat the same procedure on training data to calculate the five parameters of the alternative BLE-specific model. These parameters, power at the reference point, scale factor, spatial wavelength, distance-power gradient, and the standard deviation of shadow fading were calculated as $P_0 = -55.88$ dBm, $A = -0.93$, $\lambda = 11.31$, $\alpha = 1.51$, and $\sigma = 3.78$ dB, respectively.

3.5.1 Effects of Distance on Confidence

With these parameters of the RSSI behavior models estimated using LS estimation, we calculated the bound on standard deviation of the range measurement error using the CRLB for the two, classical and BLE specific models from Eqs. 3.8b and 3.8c, respectively. These bound are then applied to Eq. 3.9 to determine the bounds on confidence on the estimate as a function of distance, $\gamma(r)$, for the two RSSI behavior models. The solid line on top of Fig. 3-6 shows the plot of bounds on confidence on estimation as a function of range, $\gamma(r)$, for the two path loss models. As shown in this figure, although the alternative model was providing a better estimate of the RSSI values (Fig. 3-3), the confidence for estimating the range from either of the models with the CRLB remains almost the same. The alternative model provided a slightly better performance than the classical linear regression model in Fig. 3-3 because it fits better to BLE data before 3.5 m. With the BLE RSSI, for less than 1.5 m, we are almost 100% confident in the estimate enabling us to overrule the social distance range of 6 ft and for distances of more than 2.5 m we have the same confidence that we are observing the social distance rule. Since models are based on zero mean Gaussian modeling of the noise at the exact distance of 6 ft, the best algorithms can only detect the range with 50% confidence. The bounds on confidence, $\gamma(r)$, plots in Fig. 3-6

Table 3.3: Effect of Environment on Confidence

Device as Receiver	Environment		Confidence (%) Classical RSSI Model	Confidence (%) GBM
	Room Size	Location of Tester		
iPhone 11	Small Room	Near Wall Open	60.01	81.77
	Medium Room	Near Wall Congested	81.46	92.30
		Near Wall Open	61.88	75.88
		Center Open	51.68	78.69
	Large Room	Near Wall Open	64.11	84.76
		Center Open	78.89	93.43
	Outside	Center Open	74.69	84.49
iPhone XR	Medium Room	Near Wall Congested	66.71	89.35
		Near Wall Open	75.37	89.80
		Center Open	74.65	88.33
	Large Room	Near Wall Congested	84.73	94.96
		Center Open	69.08	88.39
	Hallway	Near Wall Open	66.75	87.13
		Center Open	82.85	87.46
iPhone 8	Small Room	Near Wall Congested	62.04	83.71
	Large Room	Center Open	56.55	94.02
	Hallway	Center Open	75.63	94.44
	Outside	Center Open	86.09	97.35
iPhone 7	Medium Room	Center Open	75.15	88.76
	Large Room	Near Wall Open	81.83	89.20
	Outside	Center Open	64.76	89.15
iPhone XS	Medium Room	Center Open	87.80	95.77
		Center Congested	67.22	80.91
	Outside	Center Open	70.69	84.41
iPhone 8 Plus	Medium Room	Center Open	81.85	80.02
	Large Room	Center Open	60.82	77.52
iPhone XS Max	Medium Room	Center Open	50.81	90.39
iPhone X	Medium Room	Center Congested	66.03	86.34
iPhone 7 Plus	Outside	Center Open	64.30	86.99
iPhone 6	Large Room	Center Open	76.05	92.08

show us the best-expected confidence that we obtain from RSSI measurements in our test dataset.

Table 3.4: Effect of User Behavior (Tester's Pose and Location of Phone) on Confidence

Device as Receiver	Pose		Phone on-body Location		Confidence (%)	Confidence (%)
	Tester1	Tester2	Tester1	Tester2	Classical RSSI Model	GBM
iPhone 11	Standing	Standing	In Hand	In Hand	78.08	88.40
				Front Pants Pocket	60.07	81.77
		Sitting	In Hand	In Hand	67.89	87.27
				Front Pants Pocket	56.91	81.61
	Sitting	Sitting	In Hand	In Hand	61.88	75.88
			Shirt Pocket	Shirt Pocket	60.85	82.56
iPhone XR	Standing	Standing	In Hand	In Hand	75.37	89.80
				Front Pants Pocket	74.51	91.49
			Shirt Pocket	Front Pants Pocket	75.18	87.37
			Front Pants Pocket	In Hand	84.74	94.96
		Sitting	In Purse	Front Pants Pocket	60.76	88.81
			In Hand	Rear Pants Pocket	76.05	81.98
				In Hand	61.71	87.78
				Front Pants Pocket	66.75	87.13
				In Purse	65.28	81.79
			Front Pants Pocket	In Hand	82.85	87.46
	Sitting	Standing	Shirt Pocket	Front Pants Pocket	78.37	90.55
iPhone 8	Standing	Standing	In Hand	In Hand	62.04	83.71
				Front Pants Pocket	75.63	94.44
			Front Pants Pocket	Front Pants Pocket	86.09	97.35
		Sitting	In Hand	In Hand	56.55	94.02
iPhone 7	Standing	Standing	In Hand	Front Pants Pocket	81.83	89.20
		Sitting	In Hand	In Hand	67.91	88.30
			Front Pants Pocket	In Hand	61.61	89.99
	Sitting	Standing	In Hand	In Hand	75.15	88.76
iPhone XS	Standing	Standing	In Hand	In Hand	79.34	90.09
			In Purse	Front Pants Pocket	67.22	80.91
iPhone 8 Plus	Standing	Sitting	In Hand	In Hand	80.02	81.85
			Shirt Pockets	In Purse	60.82	77.52
iPhone XS Max	Standing	Sitting	Front Pants Pocket	Front Pants Pocket	50.81	90.39
iPhone X	Standing	Standing	Front Pants Pocket	Front Pants Pocket	66.03	86.34
iPhone 7 Plus	Standing	Standing	In Hand	In Hand	64.30	86.99
iPhone 6	Standing	Sitting	In Hand	In Hand	76.05	92.08

As the next step, we examined the performance of classical estimation models from solving Eq. 3.7b to estimate the distance from the test data. In this part, we use the average RSSIs for each distance, r, in the remaining 25% of the database to solve Eq.

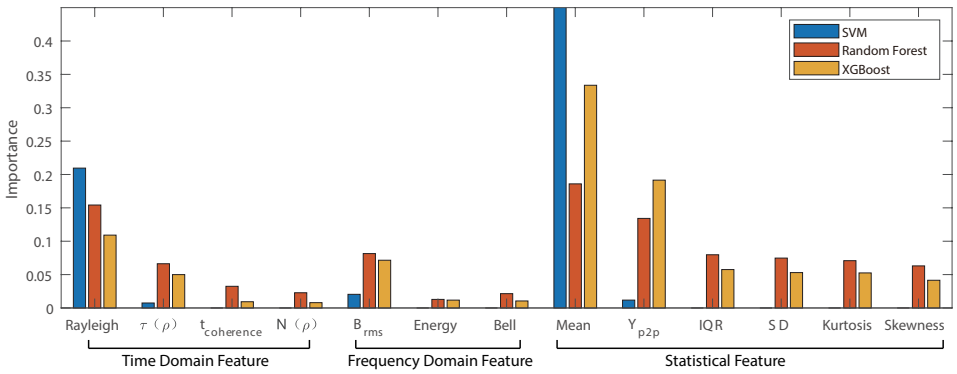


Figure 3-7: Bounds on confidence on estimation as a function of distance for MRAS RSSI database (top lines) versus performance of classical and alternative RSSI behavior modelling range estimation as well as SVM, Random Forest and GBM ML algorithms.

3.7b to find our estimate of the distance, $\hat{r}$, with each of the RSSI behavior models. Then, we empirically calculated the confidence from Eq. 3.10 for any specific data set. The bottom lines of Fig. 3-6 show the performance of the range estimation with traditional and BLE-specific alternative models obtained from empirical studies. The performance of the classical estimation algorithms follows the V-shape of the bounds on the performance, but the quality of the estimate is substantially lower than the bounds. This encouraged us to examine ML algorithms to improve performance. The dashed lines in Fig. 3-6 show the results of applying the three different ML algorithms examined in this chapter (section 3.4.2). The three ML algorithms, SVM, Random Forest, and GBM, will improve the performance over the classical methods when the social distance of the device is 6 ft or larger. However, on or in the proximity of 6 ft, classical models perform slightly better.

3.5.2 Effects of Environment and User Behavior

In the last section, we presented the results of the effects of range on confidence of estimate with classical and ML algorithms, and we compared that with performance bounds that are achievable as calculated using CRLB. Now that we established the framework for the analysis against the bounds and trained our algorithms with the training database, it is possible to explore the relationship between scenarios of

operation and the behavior of the user on the expected performance. In section 3.2, Table 3.1, we partitioned the MRAS database into five different scenarios for the test. We classified the top two scenarios in Table 3.1, describing the area size and relative location in the area, as scenarios related to the effects of the environment, and the last two scenarios describe the location in which the smartphone is carried and the pose of the tester, as scenarios describing the user behavior. In this way, we divide the scenarios into the environment and user behavior and analyze them separately for each of the seven different devices in the middle column of Table 3.1. To make the comparison more focused and clearer, we first compare the performance of the classical RSSI linear regression model with that of the GBM ML algorithm. As shown in Fig. 3-6, the results of confidence analysis with the traditional RSSI model and the BLE-specific model are very close. The average confidence over all distances for the classical method using the traditional linear regression in this figure is 69.60%, and the average confidence for the BLE-specific model is 69.55%. As the alternative model has almost the same average confidence as the classical model and the traditional algorithm offers an easier and more physically explainable method for estimation, we only compared the classical model with the best ML algorithms. The GBM classifier has the highest average confidence of 89.58% for the entire dataset, the average for Random Forest is 84.88%, and SVM has an average of 73.20% in confidence. All three ML algorithms benefited from all thirteen features of the RSSI and performed better than the classical models. Overall, GBM achieves the best results. Therefore, the comparison of GBM and traditional RSSI behavior modeling for our specific problem illustrates the best performance that the classical and ML algorithms can achieve with our existing dataset.

We began our analyses of the effects of various parameters on performance by looking at the results in different environments. Table 3.3 shows the confidence in different environmental settings for all tested smartphones with the classical RSSI model and GBM. The MRAS dataset utilized in this study was collected by multiple currently cohabiting testers and in multiple random scenarios to create a comprehensive dataset. The tests were conducted in various residential buildings or public spaces

representing a variety of architecture and five different space sizes. The testers had to strictly obey social distancing guidelines if the test was conducted in publicly accessible locations [36]. The best indoor results are obtained in medium rooms, near the walls, in congested areas, by iPhone XS (up to 87.80% for classical and up to 95.77% with GBM), and outside, the center of the open areas, by iPhone 8 (86.09% for classical and 97.35% for GBM). The worst indoor result for the classical is obtained in medium rooms, the center of the room, in open areas, by iPhone XS Max (50.81%). The worst indoor result for GBM is obtained in medium rooms, near walls, in congested areas, and by iPhone 11 (75.88%). The worst outdoor result in open areas for classical is obtained by iPhone 7 Plus (64.30%). The worst outdoor result in open areas for GBM is obtained by iPhone XS(84.41%). The average confidence for indoor environment is 70.20% for classical, which is about 4.83% less than that of outdoor (75.03%). The average confidence for indoor environment is 87.37% for GBM which is 2.24% less than that of outdoor (89.61). The average confidences are 61.03% and 82.74% in small rooms, 70.05% and 86.38% in medium rooms, 71.51% and 89.30% in large rooms, 75.08% and 89.68% in hallways for classical and GBM respectively. The average confidence increases with the room size. On average, GBM shows a 17% improvement in confidence over results achieved by the simple classical regressive model. Table 3.4 compares the confidence of estimation using classical estimation approaches with the GBM ML algorithm for different user behaviors and with different devices. The best results are obtained under the scenario that both testers are standing and holding their phones in their front pants pocket, and by iPhone 8 (86.09% for classical and 97.35% for GBM). The worst result for GBM is obtained in the scenario that both testers are sitting and holding their phones in hand, by iPhone 11 (75.88%). The worst result for classical is obtained in the scenario that Tester1 is standing, Tester2 is sitting, and both testers are holding their phones in their front pants pocket, by iPhone XS Max (50.81%). The average confidences are 73.60% (classical) and 88.77% (GBM) if both testers are standing, 61.37% (classical) and 79.22% (GBM) if both testers are sitting, 66.13% (classical) and 86.53% (GBM) if the Tester1 is standing and Tester2 is sitting, 76.76% (classical) and 89.66% (GBM) if Tester1 is sitting and

Tester2 is standing.

3.5.3 Effects of Number of Features in Performance of Machine Learning Algorithms

We studied three Machine Learning algorithms: SVM, Random Forest, and GBM. For comparative performance evaluation of these algorithms, we trained the algorithms using the 13 features shown in Table 3.2, classified into three sub-groups: time-, frequency-domain, and statistical. The ML algorithms, in addition to confidence, produce measures of the importance of the features. Fig. 3-7 shows the feature importance for the ML classifiers that we studied. As shown in the figure, the average RSSI is the most effective feature. This is the feature that we also used for classical estimation modeling. Rayleigh parameter and Y_{P2P} reflecting variations in the RSSI have shown to contribute significantly. Since the MRAS dataset is collected in a static environment and the Doppler Spectrum is related to the speed of moving antenna and moving objects between antennas, the Frequency domain features have shown less contribution to the classification compared with the other two groups. Another traditional approach in ML is to analyze the direct effect of the feature on the performance criteria, which is the confidence in the decision regarding the 6 ft threshold necessary to observe social distance. To implement this procedure, we sort the features for any of the algorithms according to their importance and re-evaluate performance while dropping them one after another. The intuition here is to demonstrate the importance of each feature on performance. Fig. 3-8 shows the result of the gradual removal of features each time we remove the single feature with the highest importance. The confidence of distance estimation using SVM drops significantly after the first three features are removed, demonstrating that the first three features dominate its performance. The performance of the Random Forest classifier drops gradually up to the removal of the first five features. For this algorithm, there is no sharp drop in performance, demonstrating that more features contribute to the model's performance. There is no dominating contribution by certain features.

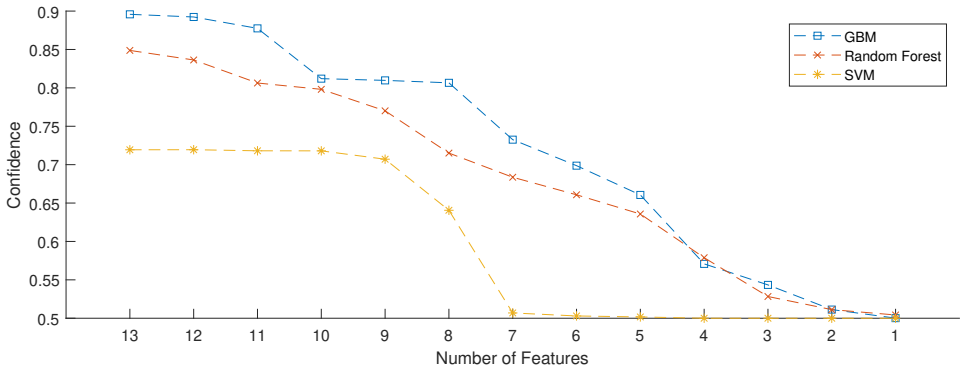


Figure 3-8: Relation between confidence and number features for SVM, Random Forest, and GBM ML algorithms as the most important of thirteen features are eliminated from training the algorithms.

The accuracy of the GBM classifier increases with the number of features and follows a similar gradual performance degradation pattern as Random Forest. As both methods are tree-based ensemble ML methods, the similarity in their performance is expected.

3.6 Summary

The risk of COVID-19 transmission increases if an uninfected person is less than 6 ft from an infected person for longer than 15 minutes (also called Too Close for Too Long (TCTL)). In this chapter, we have presented research, development, and comparative analysis of classical estimation theory methods, which enables faster computation in a more logically explainable manner and novel hybrid model-based ML approaches for proximity distance estimation using the RSSI information radiated from the broadcast channels of the BLE. Our results are based on analyses of the Mitre Range Angle Structured (MRAS) PACT dataset in five different environments, with five different locations for the smartphone, and eight different smartphones. Our analysis methodology provided a framework for the empirical analysis of the estimation confidence when applying both classical estimation theory and ML algorithms for solving the social distance estimation problem with BLE RSSI. We derived bounds on the confidence on estimation using RSSI of BLE as a function of distance. Then, we compared the performance of classical estimation theory ranging algorithms with that

of the ML algorithms against the bound and we analyzed the effects of the environment and user behavior on the performance of the algorithms. The classical estimation theory algorithms using two models for RSSI spatial behavior were compared with three different ML algorithms (Random Forest, GBM and SVM) benefiting from thirteen features of the RSSI. Classical algorithms showed an average confidence of 69.60% in correctly estimating the social distance threshold of 6 ft. The GBM ML algorithm demonstrated that using the thirteen features can increase the confidence in the estimation of this social distance using BLE RSSI with an average confidence of 89.58%, which was 19.98% higher than the average confidence achieved using the classical approach.

Chapter 4

PART I: Proximity Detection: Direct UWB TOA versus ML-Based BLE RSSI

In this chapter, we compare the direct TOA-based UWB technology with the RSSI-based BLE technology using machine learning algorithms for proximity detection during epidemics in terms of complexity of implementation, availability in existing smartphones, and precision of the results. We establish the theoretical limits on the precision and confidence of proximity estimation for both technologies using the CRLB and validate the theoretical foundations using empirical data gathered in diverse practical operating scenarios. We perform our empirical experiments at eight distances in three flat environments and one non-flat environment encompassing both Line of Sight (LOS) and Obstructed-LOS (OLOS) situations. We also analyze the effects of various postures (eight angles) of the person carrying the sensor, and four on-body locations of the sensor. To estimate the range with BLE RSSI, we use 14 features for training the GBM learning algorithm, and we compare the precision of results with those obtained from the memoryless UWB TOA ranging algorithm. We show that the memoryless UWB TOA algorithm achieves 93.60% confidence, slightly outperforming the 92.85% confidence of the BLE RSSI with a more complex GBM algorithm and the need for substantial training. The training process for the RSSI-based BLE social

distance measurements involved 3000 measurements to create a training dataset for each scenario and post-processing of data to extract 14 features of RSSI, and the ML classification algorithm consumed 200 seconds of computational time. The memoryless UWB ranging algorithm achieves more robust results without any need for training in less than 0.5 seconds of computation time.

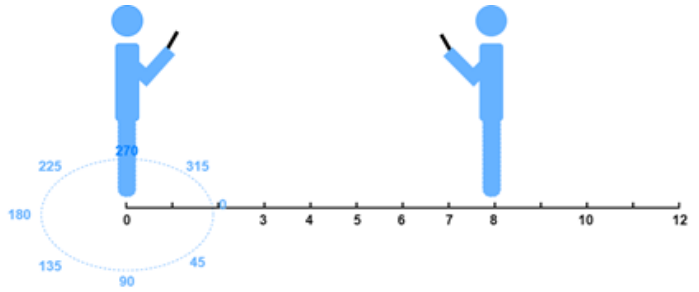
4.1 Introduction

The COVID-19 epidemic revealed the importance of research in opportunistic social distance estimation during the epidemics using the RSSI of BLE wireless technology, which are commonly available in smartphones [48] [38] [20]. Due to the effects of multipath and shadow fading on RSSI-based ranging, direct estimation of distance using RSSI is unreliable as it is compared with TOA-based positioning [60, 27]. However, due to the availability of BLE in smartphones a new trend of research to improve the performance of RSSI-based ranging with machine learning algorithms [61] and hybrid positioning approaches that integrate additional information from various mechanical sensors (e.g., accelerometer and gyroscope) that are built into many smartphones [62] has attracted considerable attention in recent literature. UWB, an alternative emerging popular wireless technology, offers a more precise TOA range without requiring complex ML algorithms that need extensive training and large amounts of labeled training data. Inexpensive UWB devices are already available in the market and the existing 5G cellular networks support UWB positioning. What is lacking in the literature is the comparison of these improved RSSI-based ranging with the UWB ranging. TOA-based ranging with the UWB signals [60, 63] has emerged as an alternative to unreliable, opportunistic RSSI-based BLE ranging. The reuse of existing popular devices, which already have the circuitry required for both UWB and BLE RSSI, could facilitate rapid, wide-scale deployment and curb epidemics quickly. With the recent emergence of low-cost UWB devices [64] and the 3GPP recommendation of TOA base ranging for 5G and beyond [65, 66, 67, 68], TOA-based UWB social distance estimation has become a viable alternative to BLE RSSI.

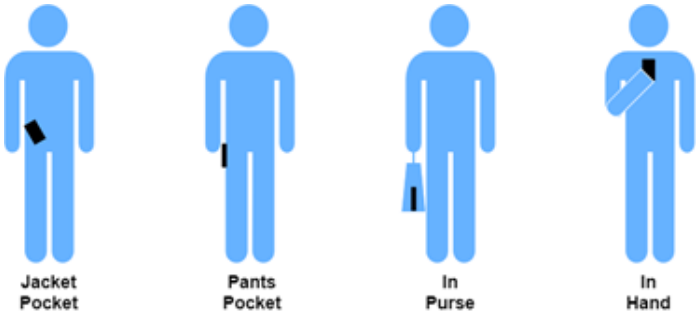
Compared to ML algorithms that require large amounts of labeled training data and an extensive training process, the UWB approach utilizes real-time algorithms that do not require training or memory while still achieving high precision ranging and confidence in proximity estimates.

The objective of this study is to establish a theoretical foundation for comparing the precision and confidence in the protocols for measuring social distance using BLE RSSI and UWB TOA devices, and to validate this theoretical basis using empirical data gathered in practical scenarios. Using the CRLB of RSSI and TOA ranging [60, 63], the theoretical foundations enable derivation and analyses of precision and confidence of estimated social distance. We expand the derivation of confidence for RSSI-based ranging previously presented in [61] to include TOA-based ranging, which we then validate using empirical data gathered in multiple practical scenarios using BLE and DecaWave UWB devices. We perform empirical experiments at eight distances in three flat environments and one non-flat environment encompassing both Line of Sight (LOS) and Obstructed-LOS (OLOS) situations. We also analyze the effects of various postures (eight angles) of the person carrying the sensor, and four on-body locations of the sensor. To estimate range using BLE RSSI, 14 RSSI features were classified using the GBM ML algorithm. The empirical results for TOA ranging using memoryless algorithms are derived from data gathered using inexpensive, off-the shelf UWB DecaWave chipsets. To the best of our knowledge, our analysis is the first to systematically compare BLE RSSI and UWB TOA for social distance estimation using rigorous theoretical foundations.

The rest of this chapter is as follows. In part II, we introduce the data-gathering scenarios we investigated and provide details of our dataset. In part III, we present the relevant theoretical foundations including the calculation of CRLB, as well as the derivation of confidence of proximity detection and corresponding bounds. In part IV, we present experiments to validate the theoretical foundations presented in part II, for the comparative performance evaluation of proximity detection using BLE RSSI and UWB TOA.



(a)



(b)

Figure 4-1: Measurement scenarios used to gather our dataset, (a) eight distances (ft) and eight angles (degrees) for the dataset. (b) four on-body locations for the BLE and UWB devices.

4.2 The Proximity Datasets and Measurements Scenarios

This study builds on our previous study reported in [61], in which we analyzed the existing MITRE Range Angle Structured (MRAS) dataset [37] provided by the Private Automated Contact Tracing (PACT) consortium. The MRAS dataset contains RSSI measurements gathered using BLE devices in various testing scenarios at different distances in flat Line-Of-Sight (LOS) situations. Since the existing MRAS dataset only includes BLE RSSI data, to facilitate a comparison with UWB, we had to conduct experiments to gather UWB TOA measurements. Specifically, in this study, we collected our own dataset to enable comparative performance evaluation of range estimation using BLE RSSI and TOA gathered using DecaWave 1000 UWB technologies. To ensure a fair comparison, we collected data from both device types in the exact same locations, environments, and scenarios. Similar to the PACT dataset,

our dataset scenarios were rich, including various room sizes and on-body positions in which the measurement device was carried/held by the owner of the test device. We also expanded on prior measurement scenarios, and included more diversified situations consisting of LOS as well as Non-Line-Of-Sight (NLOS), and non-flat staircase scenarios. Fig. 4-1a and Fig. 4-1b provide an overview of our measurement scenarios. In Fig. 4-1a, transmitter-receiver distances ranging between 3-12 ft were selected as these have been found to be the most challenging distances when detecting a social distance of 6 feet is the objective (as recommended for COVID-19). Beyond 8 feet, the intervals between the transmitter and receiver were increased in order to generate more significant RSSI differences. Fig. 4-1b shows the four on-body locations in which subjects carried their smartphones. In this study, we also considered two postures, standing and sitting. The transmitter is always positioned at the 0 ft location, while the receiver moves at increments to enable measurements at distances ranging from 3 ft to 12 ft. In order to simulate the effects of shadow fading, at each stationary location, while holding the device, the tester turned 45 degrees clockwise with the same posture. Since we defined the transmitter and receiver being face-to-face as 0 degrees, RSSI and TOA were collected at eight angles at increments of 45 degrees. At each test location shown in Fig. 4-1a, data was gathered for 30-40 seconds, which typically contained 300-400 RSSI samples or 64 TOA samples, which were stored in our dataset as depicted by Eq. 4.1.

$$\begin{cases} p(k) = \text{RSSI}(t)|_{t=kT_p}; & k = 1, ...N \\ \tau(k) = \text{TOA}(t)|_{t=kT_\tau}; & k = 1, ...M \end{cases} \tag{4.1}$$

where $p(k)$ and $\tau(k)$ are the k-th sample in the RSSI and TOA time series collected, respectively. N and M are the numbers of samples collected. Since N has values in the range 300-400 and M is 64 in this study, we selected T_p and T_τ, the time intervals of RSSI and TOA samples to ensure that data was gathered in approximately the same 30-40 seconds interval for both the BLE and UWB devices. Table 1 shows details of our environment settings. We collected our data in 4 environments: a large room (a large laboratory), medium room (a meeting room), hallway (the corridor), and

Table 4.1: Measurement scenarios used to gather data for our dataset

Scenarios	**Settings**
Description of the areas	Medium Room, Large Room Hallway, Stairway
Multipath Scenario	LOS, NLOS
Type of Device	iPhone 7, iPhone 12
Device location (on-body)	In hand, In purse, Pants pocket, Jacket Pocket
Tester's Posture	Standing, Sitting

stairway (an indoor staircase). In each environment, the transmitter (an iPhone 7) and receiver (an iPhone 12) were separated by a wall for the NLOS scenario. For the LOS scenario, the transmitter and receiver were positioned in the center of an environment with no obstacle between them. While collecting data, the two testers were required to hold the phone in specific on-body locations (see the fourth row in Table 4.1). Their posture was either sitting or standing at a given location.

We observed a variation of up to 25 dB RSSI in one scenario. In order to reduce the significant effects of fading on the amplitude of the received signal from which we calculated the RSSI, our study for RSSI included classical and Machine Learning (ML) algorithms. Variations of amplitude do not have drastic effects on TOA measurements and the results of TOA measurement can be utilized directly for range estimation. The RSSI measurements in each location, defined by Eq. 4.1, are post-processed before feeding them in different ways into classical and ML algorithms. For classical estimation algorithms, we utilized the average of RSSI and TOA measurements gathered in each location as defined by Eq. 4.2:

$$\begin{cases} P_r = \frac{1}{N} \sum_{k=1}^{N} p(k) \\ r = \frac{1}{M} \sum_{k=1}^{M} \tau(k) \end{cases} \tag{4.2}$$

This pre-processing of data associates a single average RSSI measurement, Pr, with a single average TOA measurement, τ_r, for each location. For ML techniques, the 300-400 RSSI measurements in each location were grouped in overlapping sets of RSSI measurement vectors of length L, whose elements are defined by Eq. 4.3.

$$S(n, L) = \{P(k + n);\ \ k = 0, 1, ..L - 1\};\ \ n = 1, ...N - L \tag{4.3}$$

This processing associates an $N - L$ set of L dimensional vectors to each location. Similar to [61], our performance criterion is the confidence of the correctness of the decision by the algorithm on whether the inter-subject distance is less than vs. over the social distance of 6 ft. The evaluation was done using BLE RSSI (smartphone) or UWB TOA (DW1000 receiver) measurements gathered at the same location, respectively.

4.3 Theoretical Foundations for Data Analysis

In this section, we introduce theoretical foundations derived from classical estimation theory. We relate the variance of observations (P_r and τ) with the CRLB and then derive equations for the standard deviation of distance estimates as a function of the ground truth. With the assumption that the noise is a zero-mean Gaussian distribution, we are able to derive the upper bound for the performance of the proximity detection problem using the complementary error function (erfc). For RSSI-based ranging, we employ a linear regressive model of the average received powers in the zero mean Gaussian distributed shadow fading, X(σ), with a fixed variance of σ, enabling us to formulate the classical approach based on observation of the average RSSI [61]:

$$O : P_r = P_0 - 10\alpha \log(r) + X(\sigma), \tag{4.4a}$$

Our objective is to estimate the range r. To estimate range using the TOA, we formulate the problem based on direct observation of range by multiplying the average TOA [27]:

$$O :\ \ r = c \times [\tau + \eta(\sigma_\tau)] = r + c \times \eta(\sigma_\tau) = r + \eta(\sigma_r), \tag{4.4b}$$

where c is the speed of light, and $\eta(\sigma_r)$ is the variance of TOA measurement noise determined from the CRLB of the TOA measurement. Eq. 4.4b is a function of pulse shape, bandwidth, W, and the received Signal to Noise Ratio (SNR), which provides

a lower bound on the variation of the estimation [53]:

$$\sigma_1(r) = \sqrt{\text{CRLB}} \geq \frac{\ln 10}{\sqrt{N10}} \frac{\sigma}{\alpha} r \tag{4.4c}$$

Using classical estimation theory, formulating the observation of a function of a noise parameter yields two functions: 1) the traditional RSSI linear regression model [61], and 2) the simple TOA [27] model for distance measurement (Eq. 4.4d).

$$\sigma_r = c \times \sqrt{\text{CRLB}} \geq \frac{c}{2\pi\sqrt{2 \times \text{SNR} \times W \times T_M \times {f_0}^2}} \tag{4.4d}$$

Compared with BLE RSSI, the model for UWB TOA requires more parameters due to the difference in ranging algorithms utilized. Leveraging classical estimation theory, we relate the variance of distance estimates with the variance of RSSI and TOA measurements by introducing the CRLB. In addition, we show that the theoretical analyses works not only for BLE but also for UWB, even though they utilize completely different algorithms to estimate distance.

4.3.1 Maximum Likelihood Range Estimation and CRLB for Ranging Error

In classical estimation theory, the optimal maximum likelihood estimate of the range, is the inverse of $g(r)$, the observation function (Eq. 4.5a):

$$\hat{r} = g^{-1}(O). \tag{4.5a}$$

Therefore, the optimal Maximum Likelihood Estimates (MLE) of range from average RSSI and TOA measurements are given by equation 4.5b:

$$\begin{cases} \hat{r}_{\text{RSSI}} = {g_{\text{RSSI}}}^{-1}(P_r) = 10^{-\frac{P_r - P_0}{10\alpha}} \\ \hat{r}_{\text{TOA}} = {g_{\text{TOA}}}^{-1}(r) = r = c \times \tau \end{cases} \tag{4.5b}$$

The variance of this estimation is the CRLB, which is the inverse of the Fisher Information Matrix (FIM) of the dataset (Eq. 4.6a), calculated in [27]:

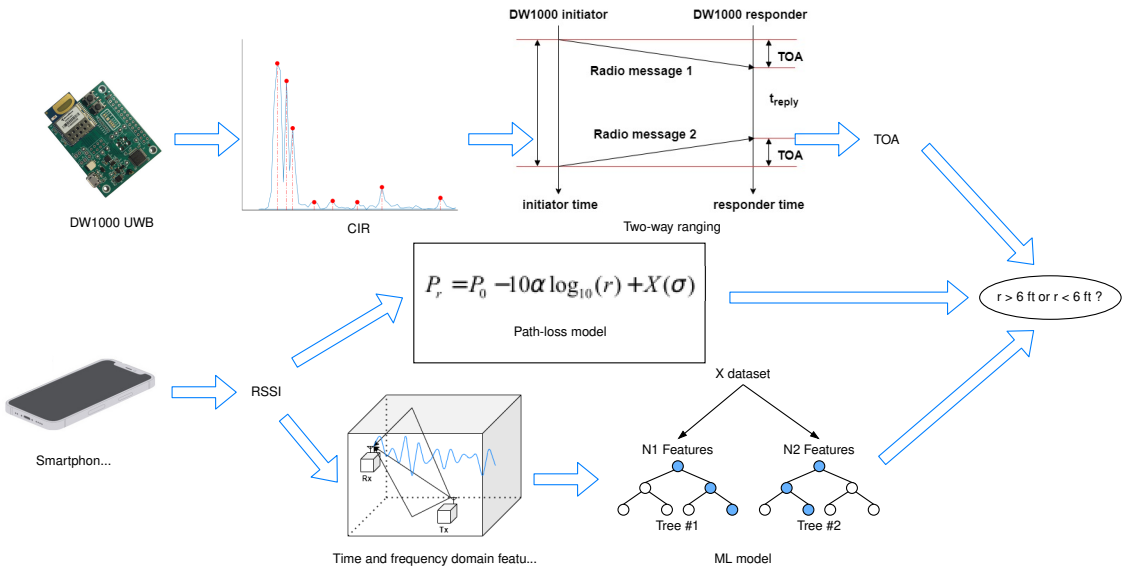


Figure 4-2: Workflow for performance evaluation: UWB TOA was gathered using DW1000, classical path-loss model utilized for BLE RSSI, and GBM for extracted RSSI features.

$$\sigma^2(r) = \text{CRLB} \geq \frac{{\sigma_M}^2}{[g'(r)]^2}, \tag{4.6a}$$

where $\sigma(r)$ is the variance of ranging error, $g(r)$ is defined by Eq. 4.5b, and σ_M is the standard deviation of measurement noise. For the RSSI measurements, the standard deviation of shadow fading, and for TOA, was defined by Eq. 4.4c. With the two proposed models Eq. 4.4c and Eq. 4.4d, and the CRLB given in Eq. 4.6a, the standard deviation of distance estimates is given by Eq. 4.6b.

$$\begin{cases} \sigma_{RSSI}(r) = \sqrt{\text{CRLB}} \geq \frac{\ln 10}{\sqrt{N} 10} \frac{\sigma}{\alpha} r \\ \sigma_{TOA}(r) = c \times \sqrt{\text{CRLB}} \geq \frac{c}{\sqrt{N} 2\pi \sqrt{2\text{SNR}(r) W T_M {f_0}^2}}, \end{cases} \tag{4.6b}$$

where N is the number of samples (300-400 for RSSI and 64 for TOA), and where SNR(0) is the Signal to Noise Ratio at a reference distance.

4.3.2 Derivation of Bounds on Confidence and Validation with Empirical Measurements

When we estimate a range, r, from noisy RSSI or TOA measurements, the characteristics of the noise can be used to calculate the confidence on estimates. Our definition of confidence in this study is the probability of correct prediction that an estimated distance is less or larger than 6 ft when the ground truth is also less or larger than 6 ft (expressed in Eq. 4.7a.

$$\gamma(r) = \Pr\left\{[\hat{r} \leq 6/r \leq 6]\, OR\, [\hat{r} > 6/r > 6]\right\} \tag{4.7a}$$

Confidence is a function of distance and reflects the degree of assurance of proper detection by the algorithm. For RSSI-based ranging, modeled by Eq. 4.4a, and its maximum likelihood estimate given by Eq. 4.5b, the confidence on ranging at a given distance r is given by Eq. 4.7b.

$$\begin{aligned} \gamma(r) &= \Pr\left\{[\hat{r} \leq 6/r \leq 6]\, OR\, [\hat{r} > 6/r > 6]\right\} \\ &= \Pr\left\{[\hat{r} \leq 6/P_r \leq P_6]\, OR\, [\hat{r} > 6/P_r > P_6]\right\} \quad , \\ &= 1 - \tfrac{1}{2}\mathrm{erfc}\left(\tfrac{|\mathrm{P_6} - \mathrm{P_r}|}{\sqrt{2}\sigma}\right) \end{aligned} \tag{4.7b}$$

where P_6 is the expected RSSI at a 6 ft distance calculated using Eq. 4.4a. In this study, we apply the Least Square (LS) algorithm to the collected RSSI data and calculate the empirical parameter (P_0, α, σ) in Eq. 4.4a before estimating P_6. In our RSSI database, there are 300-400 measurements at each location, and we calculate confidence for each of these measurements, which are then averaged over the entire set (dashed blue line in Fig. 4-4). For empirical ranging with TOA measurements, we utilize Eq. 4.5b directly multiplying the measured TOA by the speed of light and then check whether the estimated value is to the right side of 6-feet. These range estimates are averaged over the 64 TOA measurements to determine the empirical value of the confidence at each location in the dataset (dashed red line in Fig. 4-4).

We are also able to calculate confidence bounds on the range estimate as a function of distance from the calculation of the CRLB given by Eq. 4.6b. The CRLB provides

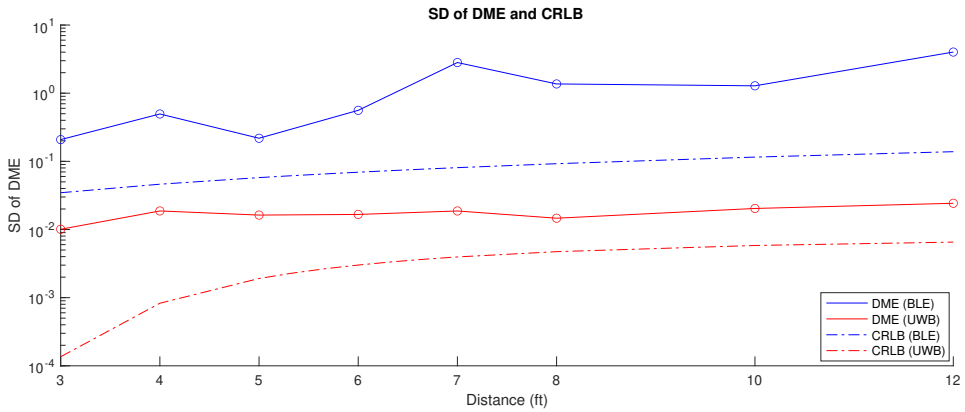


Figure 4-3: CRLB (dashed lines) versus the standard deviation of DME (solid lines)

the estimate of the variance of a parameter from the function relating the parameter to the measurement or observation. In [61], the Distance Measurement Error (DME) of BLE RSSI is given as:

$$\hat{r} - r = \eta\left[\sigma(r)\right] \tag{4.8a}$$

$\sigma(r)$ is the standard deviation of $\hat{r}$ given in Eq. 4.6b. $\eta_{[}\sigma(r)]$ is a zero-mean Gaussian distribution. The confidence bound for BLE RSSI is (Eq. 4.8b):

$$\begin{aligned} \gamma(r) &= \Pr\left\{[\hat{r} \leq 6/r \leq 6] \cap [\hat{r} > 6/r > 6]\right\} \\ &= 1 - \tfrac{1}{2}\mathrm{erfc}\left(\tfrac{|6-\mathrm{r}|}{\sqrt{2}\sigma(\mathrm{r})}\right), \end{aligned} \tag{4.8b}$$

where $\gamma(r)$ is the probability of making correct proximity detection decisions. In Eq. 4.7a, $\gamma(r)$ is the actual probability calculated from real data. However, in Eq. 4.8b, $\gamma(r)$ is the theoretical upper bound for a given distance r.

4.4 Comparison of Ranging with BLE and UWB Signals

In this section, we will first introduce the accuracy of ranging and its theoretical bounds since the confidence of proximity detection is highly related to it. As shown is Section 4.1, the theoretical foundation of this chapter consists of two parts: 1) the

bound of DME which is calculated from Eq. 4.6b, and 2) the bound on proximity detection confidence, which is calculated from Eq. 4.8b. In this section, we present the simulation results of the theoretical bounds as well as the empirical results obtained from BLE RSSI and UWB TOA data. Fig. 4-2 shows the overall structure of the performance evaluation in this section. The UWB TOA is obtained using a two-way ranging approach provided by Decawave. Both a classical regression model and an ML algorithm (GBM), are evaluated on BLE RSSI data. We utilize the confidence defined in section 4.3.2 as the criterion for performance evaluation. The suggested 6 ft social distance is the decision threshold. If the ground truth $r < 6ft$, the confidence is the probability $P\{\hat{r} < 6ft | r < 6ft\}$. Similarly, if the ground truth $r > 6ft$, the confidence is the probability $P\{\hat{r} > 6ft | r > 6ft\}$. We begin by calculating DME for both RSSI and TOA. Then we present the bounds on confidence of RSSI and TOA based ranging, respectively. We also evaluate the performance of BLE and UWB in various environments, which is presented in section 4.4.2.

For BLE RSSI, we first calculate the parameters of the path loss model in Eq. 4.4a by applying the Least Square Estimate (LSE) algorithm. Then, to compute the bound or confidence, we substitute the parameter values calculated into Eq. 4.8b. As shown in Fig. 4-2, we use the GBM ML algorithm to decide whether the distance between two devices is within the 6 ft range. To train a GBM, we extracted 14 features, including frequency-domain and time-domain features. The training set and test set were 80% and 20%, respectively. For UWB TOA, the theoretical confidence bound is also computed using Eq. 4.8b. The empirical confidence at a certain distance is simply calculated as the ratio of the count of correct decisions and the number of total samples. Then we calculated the standard deviation of TOA measurement and plugged the value into Eq. 4.6b to calculate the CRLB of TOA. For all scenarios described in Table 4.1, the parameters for RSSI regression model (Eq. 4.4a) is P_0 = -46.80dBm, α = 2.07 and σ = 3.41 dB. And the CRLB parameters for TOA (Eq. 4.4c) are preset for Deca 1000 transceivers. They are f_0 = 3993.6 MHz, M=499.4 MHz, and T_M = 26 ms. We use the first meter SNR as the reference and assume the SNR is inversely proportional to the squared distance.

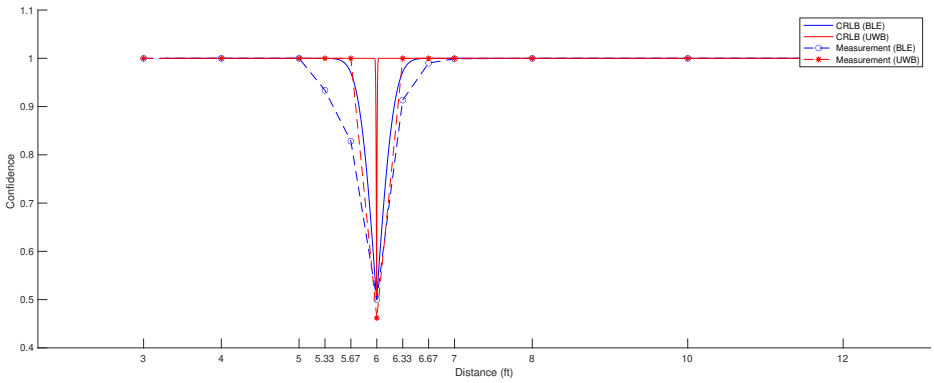


Figure 4-4: Bounds on confidence on estimate as a function of distance (solid lines) versus performance of collected dataset (dashed lines)

4.4.1 Results of Theoretical Foundations Analysis

Since we have already calculated the required parameters for CRLB, we first compare the lower bound of the standard deviation of DME. As shown in Fig. 4-3, the two dashed lines are the CLRB obtained by substituting the parameters into Eq. 4.6b. At all distances, the TOA has a much lower CRLB (the lower dashed line) compared to the CRLB of RSSI (the upper dashed line), which means the UWB TOA is more accurate than BLE RSSI. Next, we calculate the actual SD of DME by calculating 4-4 the standard deviation of the difference between distance estimates and the ground truth shown in Fig. 4-1a. The two solid lines in Fig. 4-3 are the observed results. For both RSSI and TOA estimates, the SD of DME is always above the corresponding CRLB. Fig. 4-3 shows that Eq. 4.6b holds for the estimate generated using the TOA-based, two-way ranging algorithm as well as the GBM-based RSSI estimate.

Using the correct CRLB, we can then analyze the main criterion in this study, the confidence of proximity detection. The theoretical confidence bounds for TOA and RSSI are both calculated using Eq. 4.6b. For RSSI, we input the 14 features into GBM, and it outputs whether the estimated distance is in the 6 ft range. For TOA, we multiply the speed of light by the estimated time and then compare the result with 6 ft to decide whether it is within the 6 ft range. Fig. 4-4 shows the proposed confidence bound (solid lines) and the empirical confidence calculated from the collected dataset (dashed lines). The V-shape bounds show the best performance

of these two approaches on the test dataset. The solid red line is the bound of UWB TOA estimates, and the solid blue line is the bound of BLE RSSI estimates. Both approaches show approximately 100% confidence when the ground truth is far from the threshold ($r > 7ft$ and $r < 5ft$). Theoretically, UWB TOA can achieve much better performance around 6 ft because it has a narrow distance interval with low confidence. To achieve a more precise comparison in low-performance areas, we added two more locations between 5 ft and 6 ft and two more locations between 6 ft and 7 ft. For the distances far from the threshold ($r > 7ft$ and $r < 5ft$), the two-way ranging and GBM have almost the same confidence as the theoretical bounds. For the low confidence range ($5 < r < 7ft$), UWB TOA shows much higher confidence than BLE RSSI. At exactly 6 ft, both algorithms have the lowest confidence. Our empirical result shows that the confidence at all distances is consistently less than or equal to the upper bound calculated from Eq. 4.8b, and that UWB TOA outperforms BLE RSSI when the distance is close to 6 ft.

4.4.2 Effect of Measurement Scenario

In this section, we present the results of BLE RSSI and UWB TOA in various environments and scenarios described in Table 4.1. We use the average confidence at eight locations and eight angles shown in Fig. 4-1a as the performance criteria. As shown in Fig. 4-4, the most significant confidence difference is observed between 5 ft and 7 ft. Since the eight selected locations are distributed between 3 ft and 12 ft, the difference between the average confidences of BLE RSSI and UWB TOA is not very significant (less than 5%). While comparing the performance in different environments, the user's posture and on-body location they carried the device were always "standing" and "in hand", respectively.

We began our performance evaluation as well as analysis of the effect of parameters by comparing the confidence calculated from GBM predictions on BLE RSSI and the confidence from two-way ranging on UWB TOA in different environments. Table 4.2 shows our results in eight environment settings. The best result of BLE RSSI is obtained in the large room with LOS scenario (94.38 %) and the worst result

Table 4.2: Effect of Environment on Confidence

Environment		Confidence (%)	Confidence (%)
Room Size	Multipath Scenario	BLE RSSI (GBM)	UWB TOA
Medium Room	LOS	94.02	94.19
	NLOS	93.30	94.63
Large Room	LOS	94.38	94.43
	NLOS	93.04	94.58
Hallway	LOS	92.18	93.85
	NLOS	91.79	94.34
Stairway	LOS	91.18	94.04
	NLOS	93.68	87.55

Table 4.3: Effect of User Behavior (Tester's Posture and Location of Phone) on Confidence

Posture		Device on-body Location		Confidence (%)	Confidence (%)
Tester1	Tester2	Tester1	Tester2	BLE RSSI (GBM)	UWB TOA
Sitting	Standing	In Hand	In Hand	93.39	93.55
			Jacket Pocket	94.28	94.34
			Pants Pocket	90.28	93.26
			In Purse	93.99	95.12
	Sitting	In Hand	In Hand	94.34	93.75
Standing	Standing	In Hand	In Hand	92.49	93.07
	Sitting	In Hand	In Hand	90.39	93.31

is obtained in the staircase with an LOS scenario (91.18%). For BLE RSSI, the confidence difference is up to 2.7% considering LOS and NLOS and 3.2% for different room sizes. The average confidence of BLE RSSI is 92.84%. The best and worst results of TOA are 94.63% and 87.5% in the medium room and staircase respectively. The confidence difference of TOA is 7.08% and 6.49% for varied room size and multipath scenarios, respectively. UWB TOA has 0.5% more confidence than BLE RSSI. Table 4.3 shows confidence for various postures and on-body locations described in Fig. 4-1b. First, we keep the user behavior fixed (both testers hold their phones in hand) and discuss the effect of user postures. Tester1 is the person who holds the transmitter and Tester2 holds the receiver. For the BLE RSSI approach, confidence is 2.43% higher if Tester1 was standing, compared to results when Tester1 was sitting. The best result is obtained when both testers are sitting, which is 94.34%. and the worst result of 90.3% is obtained when Tester1 is sitting and Tester2 is sitting. Thus, we conclude that

different postures can cause up to 3.95% confidence. However, different postures do not cause huge variations in the result of UWB. As shown in the last column in Table 4.3, the maximum confidence difference for various postures and on-body locations is 0.68%. To compare the effect of on-body location of the device, we kept the postures fixed (sitting and standing). The confidence differences are 4% and 1.86% for BLE RSSI and UWB TOA, respectively. The change of on-body location affected BLE RSSI more, while the change in the multipath environment affected UWB TOA more.

In this section, we presented an empirical comparative performance evaluation of proximity detection for the social distance using the RSSI of the BLE and TOA of the DecaWave UWB devices. To evaluate the performance, we used the confidence on whether a tester is within the social distance (6 feet) as the primary criterion. We provided a novel theoretical foundation with classical estimation theory using the CRLB to develop bounds for the confidence. Then we compared the performance of UWB TOA obtained by a two-way ranging algorithm with a BLE RSSI-based machine learning algorithm against these bounds. We found that for both UWB TOA and BLE RSSI, the empirical result is almost the same as the theoretical bound if the ground truth is far from the boundary (6 feet). However, both the theoretical bound and the empirical result have the worst performance when approaching the boundary. The theoretical foundations show that the average confidence of UWB TOA estimation for the distance between 3-12 feet is 96.98%, which is 1.58% better than utilizing BLE RSSI. To validate the theoretical bounds and evaluate the confidence provided by BLE RSSI and UWB TOA, we collected a novel dataset in fifteen scenarios in order to obtain a fair comparative analysis. For both LOS and OLOS situations, we conducted experiments in three flat environments (medium and large rooms, and corridor) and an environment where the transmitter and receiver are placed at different heights (stairway). The 7 other scenarios involved varying the tester's postures (sitting and standing) and the places where one tester carried the receiver (in hand, jacket pocket, pants pocket, and purse). A machine learning model based on the GBM algorithm was trained using BLE RSSI data for each scenario to estimate the confidence. For the UWB, the TOA is obtained directly using the Decawave two-way ranging algorithm.

UWB TOA outperforms BLE RSSI in almost all environments except the staircase environments with a NLOS situation and both testers were sitting. For BLE RSSI, different postures caused up to 3.95% confidence difference. However, the maximum confidence difference obtained by UWB TOA was only 0.68%. On average, the confidence of UWB TOA was 0.75% better than that of BLE RSSI, which means that the proposed theoretical bound is consistent with the empirical result. Theoretically, the proposed theoretical foundation gives us a relation between the traditional CRLB and the confidence of proximity detection and a bound for confidence as a function of distance. Practically, the theoretical foundation provides an approach to analyze whether a technique is suitable for solving the proximity detection problem.

4.5 Summary

The availability of TOA-based UWB positioning in emerging, inexpensive IoT devices and the 5G cellular networks has created an alternative to RSSI-based BLE positioning for proximity detection for social distancing. The UWB solution operates in real-time without a need for training a complex ML algorithm for RSSI-based ranging. In this chapter, we presented an empirical comparative performance evaluation of proximity detection for social distance using BLE RSSI features classified offline using a complex GBM ML algorithm and TOA data from the DecaWave UWB devices using a simple real-time, memoryless algorithm embedded on the device. To evaluate the performance, we used the confidence on whether a tester is within the social distance (6 feet) as the primary criterion. We provided a novel theoretical foundation with classical estimation theory using the CRLB to develop bounds for the confidence. Then we compared the performance of UWB TOA obtained using a two-way ranging algorithm with BLE RSSI-based approach using a GBM ML algorithm against these bounds. We demonstrated that even without any training, the UWB TOA can outperform BLE RSSI using the GBM ML algorithm. We expect that the next generation of social distance monitoring systems transfer from RSSI based to TOA based technologies.

Chapter 5

PART II: RII Features in Licensed and Unlicensed Bands

In this chapter, we first present the results of an empirical study of the comparative statistics of fourteen interference features in licensed and unlicensed bands in a selected route in downtown Worcester, MA. Then, we benefit from these features to train a machine learning algorithm to predict the availability of the channels for intelligent spectrum access for a vehicle. The main component of the vehicular interference monitoring system is an ultra wideband programmable 26GHz Agilent E4407 spectrum analyzer, interfaced with a GPS device to record the location of measurements, and a laptop to store the results in a centralized database. With this measurement system loaded in a car we drive in a selected path to monitor the interference in 1.9 – 2.5 GHz frequency band, which includes high traffic density neighboring licensed and unlicensed bands. With the empirical monitored interference, we study the statistical behavior of fourteen interference features logically categorized into four classes: interference intensity, correlation properties, spectrum occupancy, and Doppler spectrum in licensed and unlicensed bands. Finally, We benefit from these features to train a machine learning algorithm to predict the availability of the licensed and unlicensed bands for vehicular network access to the fixed backbone network infrastructure.

5.1 Introduction

The exponential growth of IoT devices and the demand for smart devices for higher data rates has heightened the importance of intelligent spectrum management in cellular 5G/6G operating in licensed bands and Wi-Fi devices operating in unlicensed bands [69][70]. As a result, empirical understanding of statistical features of interference for intelligent spectrum management has emerged as an important area of research and development [35]. To understand the interference existing in the spectrum, recently, researchers have monitored the interference in unlicensed ISM bands for Wi-Fi in a fixed location [71], or mobile locations [72].

In this chapter, we present results and analysis of the empirical study of the comparative statistics of interference features in licensed and unlicensed bands in a typical urban area (downtown Worcester, MA) with a programmable interference monitoring system. The monitoring system includes an ultra-wideband programmable 26GHz HP E4407 spectrum analyzer, interfaced with a GPS device to record the location of measurements, and a laptop to store the results in a centralized database. With the empirical monitored interference, we study the temporal, frequency, and spatial behavior of interference in typical urban area scenarios for intelligent spectrum management for a mobile vehicle operating in a congested downtown area.

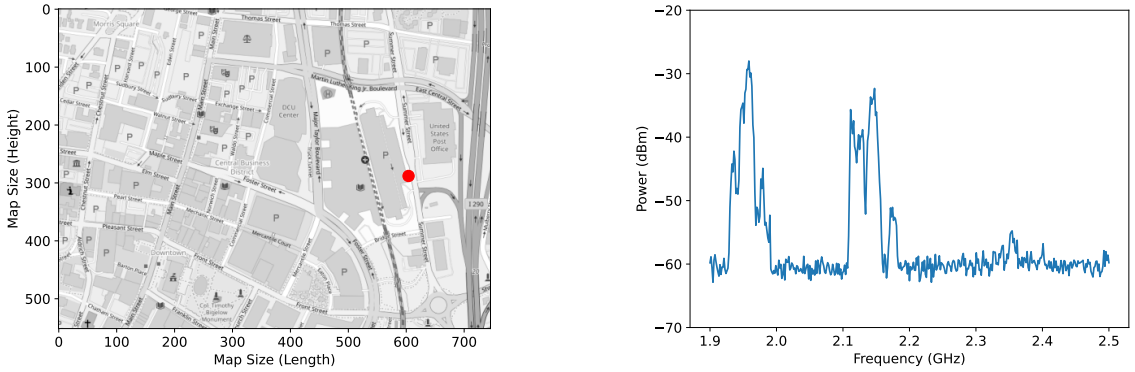
Section 5.2 provides a literature search on interference monitoring for intelligent spectrum management. Section III describes details of the measurement system and the samples of measurements. In section IV, we describe the features of the interference and preliminarily analyze their comparative statistics in adjacent licensed and unlicensed bands. Section V provides conclusions and future directions.

5.2 Background Research in Interference Monitoring and Analysis

Intelligence spectrum management has recently received significant attention in the past few years for efficient utilization of spectrum [69][70]. It is one of the essential

(a) (b)

Figure 5-1: Data collection platform, (a) different components of the data collection platform (b) platform during the data collection with a moving vehicle

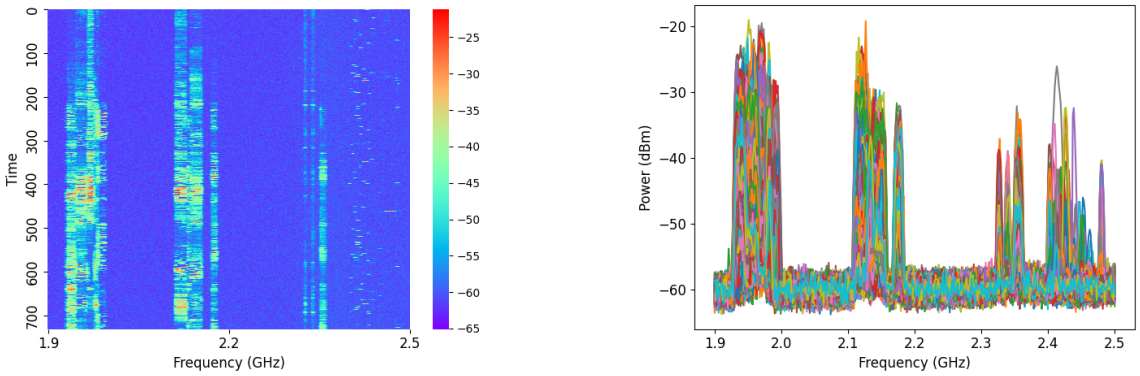


(a) (b)

Figure 5-2: Data collection process in downtown Worcester (a) The location of a recorded measurement: background the map for the entire area and red dot is displayed according to logged GPS coordinates; (b) the interference intensity in 1.9 – 2.5 GHz at this location: an intensity range of -27 – -63 dBm is observed from Cellular Band 2, 66, 30 and unlicensed band (from left to right).

techniques for next-generation communication systems to liberate the spectrum [35]. To understand the meaning and status of interference in an environment, we have to monitor the interference in the interested frequency band. As a background, we will introduce the spectrum monitoring system first.

The spectrum monitoring system can be divided into passive and active monitoring. Active monitoring requires communication between devices. The instruments are usually smartphones and laptops. Thus, they are limited by communication protocols and bandwidth. The advantage is the availability of addresses (MAC, IP, and Cell ID).



(a) (b)

Figure 5-3: Visualized the dataset of complete measurements during a test drive showing the time and frequency behaviors of raw data. A: spectrogram of the dataset, the vertical axis shows time slot numbers, the horizontal axis shows the 401 frequency bins and the minimum, center, and maximum frequencies (left to right). Interference intensity is indicated by the color bar B: congested plot of the same set of data. Active and inactive channels can be distinguished by using a threshold of -57 dBm.

Active monitoring is excellent for analyzing the interference for the same technology. Passive monitoring is typically implemented using a spectrum analyzer, which can scan super-wideband for all technologies. However, detailed information is missing because spectrum analyzers cannot distinguish different technologies. In this work, we passively implement a spectrum monitoring system to benefit from the wide frequency range. Figure 1 provides an overview of passive spectrum monitoring systems described in prior research. Static spectrum measurement provides researchers with a large amount of data over a long time, but the location is fixed. Mobile measurement is more flexible with location and can monitor the spectrum in a large area. However, the data is limited by the number of measurements and the time information. In section 5.3, we will introduce our mobile spectrum measurement system.

Related works successfully establish models for Wi-Fi bands to determine whether the channel is busy. [71] measured the 2.4 GHz and 5GHz Wi-Fi bands and proved the current channel allocation is not efficient enough. 7 different distributions are evaluated using Kolmogorov-Smirnov (KS) distance, Kullback-Leibler (KL) divergence, and Bhattacharyya distance to compare the predictability. Authors in [71] monitored the 2.4 GHz and 5 GHz bands in a railway station with fixed location receivers.

Table 5.1: Cell Tower Information in downtown, Worcester

Cell Tower	Frequency Band Number
eNB ID 48382 (Macro)	2, 12, 66
eNB ID 48645 (Macro)	2, 12, 66, 71
eNB ID 63023 (Macro)	2, 4, 5, 12, 30
eNB ID 60023 (Macro)	2, 5, 14, 66

Table 5.2: Active Licensed Frequency Bands in downtown, Worcester

Band	Frequency (MHz)	Type	Duplex Mode
2	1930 – 1990	Downlink	FDD
66	2110 – 2200	Downlink	FDD
30	2350 – 2360	Downlink	FDD
ISM	2400 – 2500	Unlicensed	TDD

Spectrogram plays a vital role in the analysis of temporal features. To exploit the spectrogram using Deep Learning algorithms, [73] proposes the Q-spectrogram and shows it is better for CNN than the traditional spectrogram using experimental data. The Deep Learning model trained on Q-spectrogram offers 99% accuracy in estimating the Wi-Fi traffic load (five density levels). Instead of using the locations of base stations, [72] monitors 915-928 MHz ISM-band in Melbourne, Australia. The authors stated that the normalized histogram follows a log-normal distribution. The duty cycles are calculated across all the frequency bins as the key parameters for evaluating occupancy. The definition of whether a band is busy varies in the literature, But for Wi-Fi bands, the binarization of data can simplify the modeling. [72] and [71] converted the data into binary patterns and counted the length of the continuous busy and idle durations with certain resolutions. The threshold for binarization also varies in the literature since it depends on the design purpose of the application. To provide a comprehensive understanding of interference in different bands. To comprehensively understand interference in different bands, we follow a similar approach to analyze the interference behavior in 1.9-2.5 GHz.

Table 5.3: Summary of Features of Interference

Feature	Parameter	Equation
Interference Intensity		
Empirical distribution of interference (Beta distribution)	P: the normalized interference Γ: Gamma distribution a and b: shape parameters of Beta distribution	$\frac{\Gamma(a+b)P^{a-1}(1-P))^{b-1}}{\Gamma(a)\Gamma(b)}$ (1)
Average interference intensity (in dB)	N: Number of samples P_I: the power of interference	$\frac{1}{N}\sum_{n=1}^{N} P_I(n)$ (2)
Standard Deviation	N: Number of samples P_I: the power of interference μ_I: Average Interference	$\sqrt{\frac{1}{N}\sum_{n=1}^{N}(P_I(n)-\mu_I)^2}$ (3)
Fade Duration	ρ: normalized threshold B_{rms}: Doppler Spread	$\tau(\rho)=\frac{e^{\rho^2}-1}{\sqrt{2\pi}\rho B_{\text{rms}}}$ (4)
Level Crossing Rate	ρ: normalized threshold B_{rms}: Doppler Spread	$N(\rho)=\sqrt{2\pi}\rho B_{\text{rms}}e^{-\rho^2}$ (5)
50% Coherence Time	E: the expected value operator P_{t_1}: Interference strength at t_1 μ_{t_2}: interference mean σ_{t_2}: standard deviation	$\frac{\text{E}[(P_{t_1}-\mu_{t_1})\overline{(P_{t_2}-\mu_{t_2})}]}{\sigma_{t_1}\sigma_{t_2}}$ (6)
Correlation Properties		
Time Correlations	cov: the covariance operator T_i and T_j: two sample sequences in different time	$\frac{\text{cov}(T_i,T_j)}{\sigma_i\sigma_j}$ (7)
Frequency Correlations	cov: the covariance operator F_i and F_j: two sample sequences in different frequency bins	$\frac{\text{cov}(F_i,F_j)}{\sigma_i\sigma_j}$ (8)
Spectrum Occupancy		
Channel Occupancy	T_{busy}: the time that the channel is occupied T_{total}: the total time	$\frac{T_{busy}}{T_{total}}$ (9)
Empirical distribution of duration (Beta distribution)	T_d: the normalized busy duration Γ: Gamma distribution a and b: shape parameters of Beta distribution	$\frac{\Gamma(a+b){T_d}^{a-1}(1-T_d)^{b-1}}{\Gamma(a)\Gamma(b)}$ (10)
Empirical distribution of inter-arrival time (Beta distribution)	T_a: the normalized inter-arrival time Γ: Gamma distribution a and b: shape parameters of Beta distribution	$\frac{\Gamma(a+b){T_a}^{a-1}(1-T_a)^{b-1}}{\Gamma(a)\Gamma(b)}$ (11)
Doppler Spectrum		
Doppler spectrum shape	λ: the normalized Doppler frequency a and b: shape parameters of Doppler	$D(\lambda)=\frac{a}{1+b\lambda^2}$ (12)
RMS Doppler Spread	λ: the normalized Doppler frequency $D(\lambda)$: Doppler Spectrum	$B_{\text{rms}}=\sqrt{\frac{\int\lambda^2 D(\lambda)d\lambda}{\int D(\lambda)d\lambda}}$ (13)
Energy of Doppler Spectrum	λ: the normalized Doppler frequency $D(\lambda)$: Doppler Spectrum	$\int D(\lambda)d\lambda$ (14)

5.3 Methodology for Empirical Study

This section first presents our data acquisition methodology, platform, and scenarios, for generating a real-world interference database. We empirically study this database to introduce the statistical features for comparing interference characteristics in licensed and unlicensed bands.

5.3.1 Interference Database Creation

In this section, we first describe our data collection setup. Then we describe the measurement scenarios, and finally, we report the data collection procedure.

Data Collection Setup. Our data acquisition platform consists of an Agilent E4407B ESA-E Spectrum Analyzer [74], a GPS receiver [75], and a Laptop. The spectrum analyzer measure 9KHz to 26.5 GHz with 0.4 dB overall amplitude accuracy. Our frequency of interest is 1.9 GHz to 2.5 GHz, which includes the Cellular band (Band 2, 30, and 66) and 2.4 GHz ISM band. Focusing on this frequency allows us to compare the interference for different bands. The antenna connected to the spectrum analyzer is omnidirectional. The GPS receiver provides us with the GPS coordinates of the collected data, allowing us to acquire spatial information. Fig. 5-2 shows the actual measurement in a location on the route. We use a dell m17 series laptop with a Core i7-9750h processor, 16 GB RAM, and 1 TB storage. The spectrum analyzer and the GPS receiver are connected to the laptop with GPIB and USB cables, respectively. Fig. 5-1b shows a labeled version of the setup. During data collection, we place this setup on the back seat of a car, as shown in Fig. 5-1a. We utilize PyVISA [76], a Python package, for easy equipment control and data retrieval.

Measurement Scenario. To collect data from a dense wireless scenario, we choose a 0.8km^2 area near Worcester Common, Worcester, MA, an urban region with a large population. We specifically choose this location due to the presence of at least five cellular towers from any point. Fig. 5-2a is the map of the selected area, and the trajectory is marked as a solid red dot. Here, the relative location is calculated by GPS

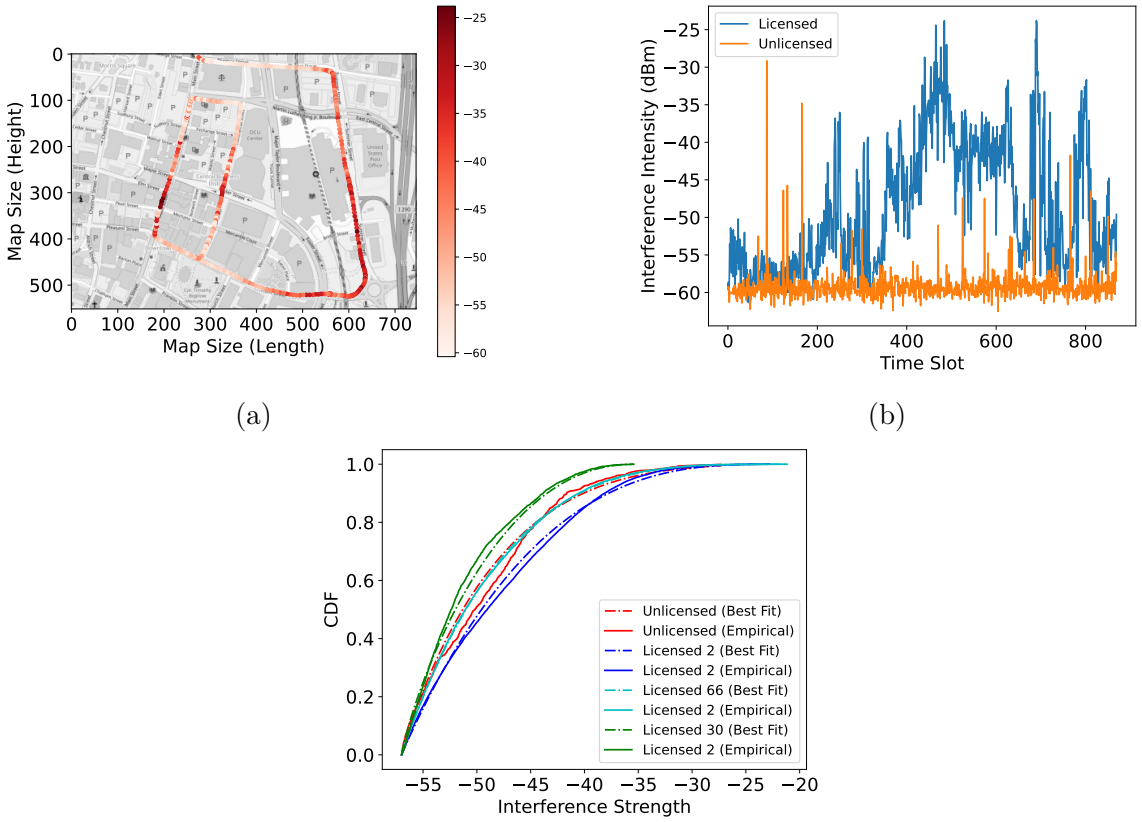
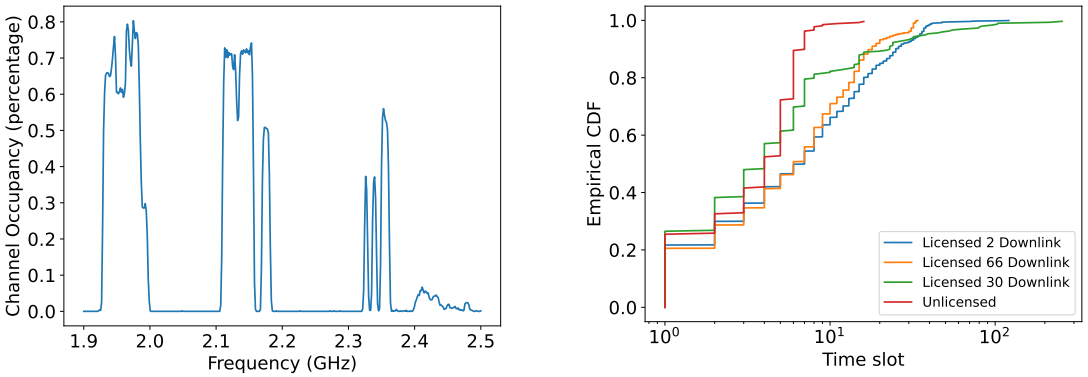


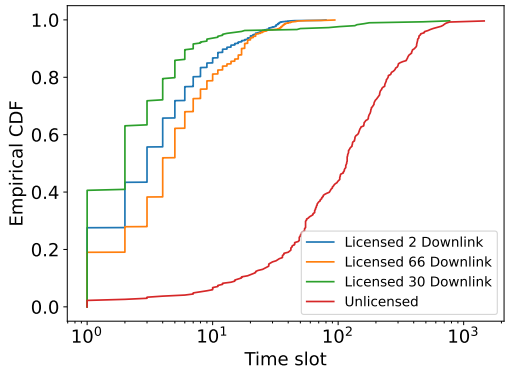
Figure 5-4: Interference intensity features. (a) The trajectory of a vehicle (data from GPS) and the interference intensity in Band 2 are plotted on the map (data from Spectrum analyzer). The shade of dots indicates the interference intensity. (b) Interference intensity comparison of Band 2 (blue) and unlicensed band (orange) (c) Empirical distributions of interference intensity and the best-fits (Beta) of all bands introduced in Table 5.2

coordinates. Fig. 5-2b is the spectrum measurement corresponding to this location. We observed a -27 – -63 noise level during the data collection.

Data Collection Procedure. We perform 5 drives, each lasting between 10 – 15 minutes, depending on the traffic situation. During each drive, we collect about 700 – 1000 measurements. The total collected data site is 2 hours. We set the frequency range to 1.9 – 2.5 GHz and the amplitude to -20 – -70 dBm to collect the data. Each measurement consists of 401 amplitude readings representing the 600 MHz frequency span and the 2.2 GHz center frequency. To compensate for the varying noise level, we normalized the amplitude to make fair comparisons.



(a) (b)



(c)

Figure 5-5: Spectrum occupancy property. (a) Channel occupancy in the entire frequency range 1.9 – 2.5 GHz. The channel occupancy is less than 0.8 for licensed bands, and for an unlicensed band is less than 0.1. (b) Empirical CDF of the busy duration of channels in Table 2. The horizontal axis is the time slots in the logarithmic scale (c) Empirical CDF of the inter-arrival of channels in Table 5.2. The inter-arrival time in an unlicensed band is significantly longer than in licensed bands.

5.3.2 Interference Features

This section describes the interference features for comparing the characteristics of licensed and unlicensed traffic.

Intelligent spectrum management demands understanding the statistical behavior of interference features for implementing machine learning algorithms. Table 5.3 shows the fourteen features that we identify as interference features from the previous works [61]. These features describe the propagation characteristics of RF signals, and they are widely used in wireless localization and RF-based motion/behavior detection areas. Based on the spectrum management problem, we reclassify the features into 4

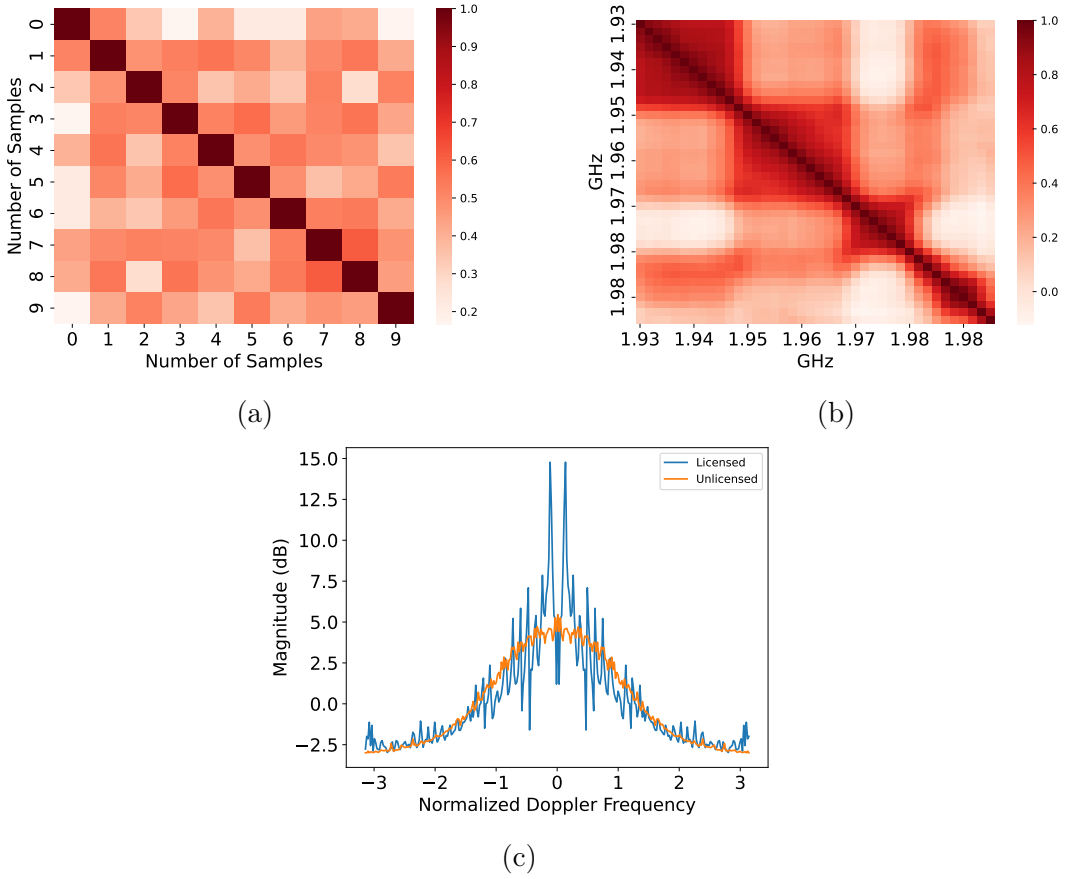


Figure 5-6: Correlation properties and Doppler spectrum features. (a) Time correlation of 10 consecutive 50 samples in Band 2, which is slightly time-correlated (b) Frequency correlation of all bins in Band 2 (c) Doppler Spectrum calculated over consecutive samples in licensed Band 2 (blue) and an unlicensed band (orange). Normalized amplitudes are plotted against the normalized Doppler frequency.

classes instead of using the categories designed for the proximity detection problem in [61]. For the benefit of the analysis, we divide these interference features into four groups – interference intensity, correlation properties in time and frequency, Doppler shape, and spectrum occupancy.

Interference intensity feature group includes the shape of the distribution, which follows the Beta distribution [71]). Eq. 1 in Table 5.3 represents this distribution, and the second column of this table explains the parameters of this distribution. Other features of interference density are average, standard deviation, fade duration, level crossing rate, and coherence time represented by Eqs.2 to 6. Fig. 5-3a and 5-3b show a sample interference intensity of a channel during the measurement time.

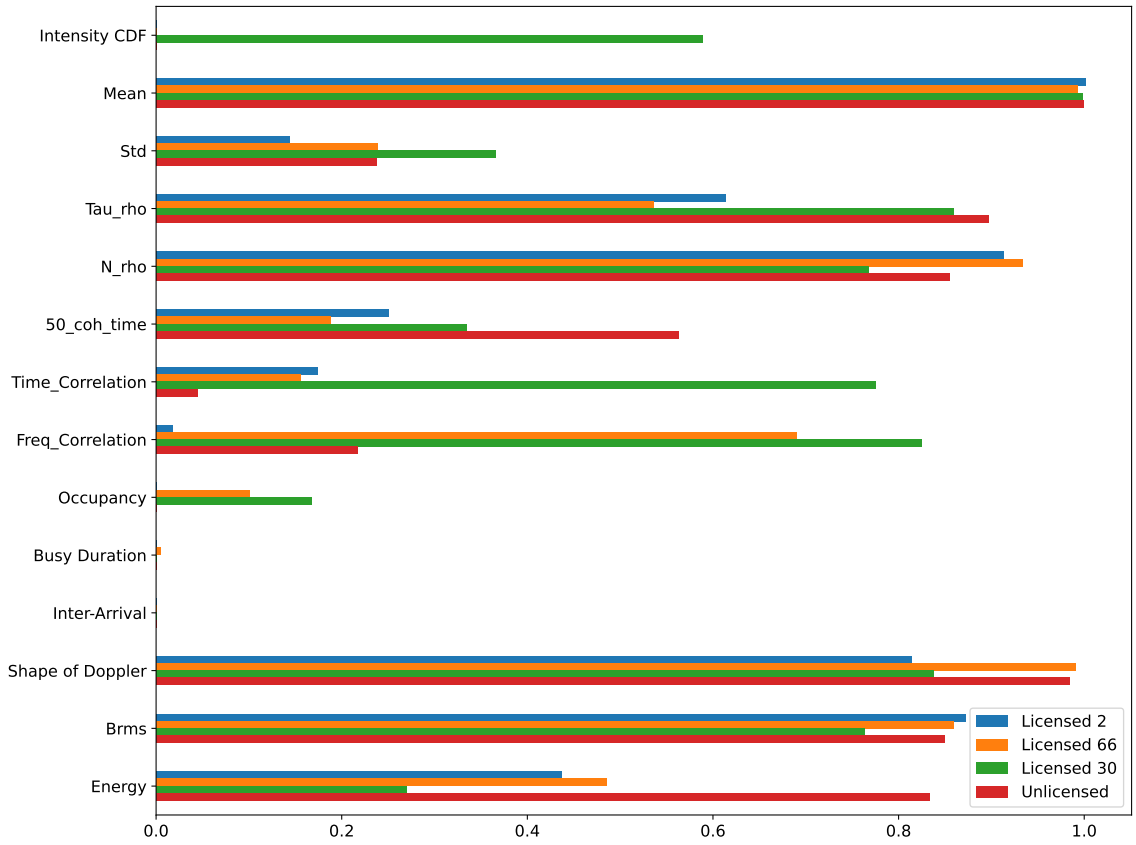
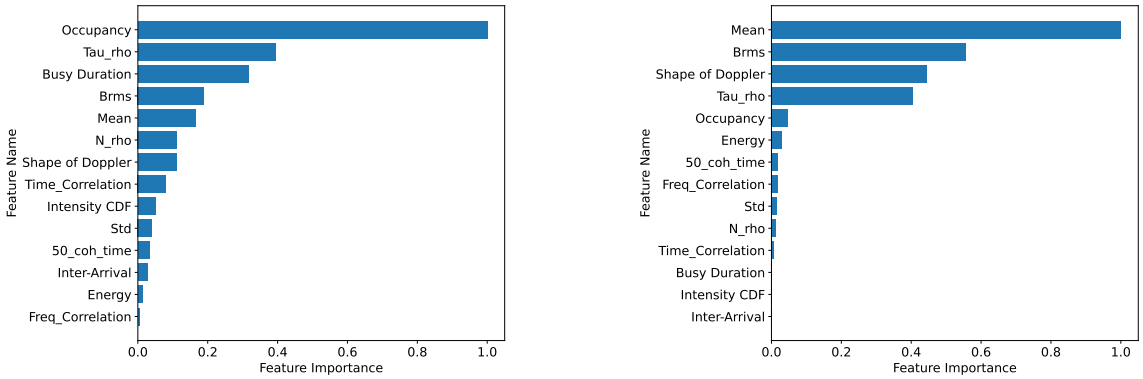


Figure 5-7: A sample normalized features of training data for the SVM classifier, calculated over 50 samples in four bands shown in Table 5.2. Some features are not available because data is not adequate to fit the Beta function.

Correlation properties feature set includes the correlation properties in time- and frequency-domain described by Eq. 7 and 8. Correlation in time is applied to the time-domain behavior of interference shown in Fig. 5-3a, and frequency correlation is applied to the results shown in Fig. 5-3b.

Spectrum occupancy feature group includes the channel occupancy. Channel occupancy or utilization is the ratio of time that the channel is utilized for application to the overall time of the measurement (defined by Eq. 9). Any channel is either active or idle, and channel occupancy is the ratio of active time over the whole duration. Another relevant feature of the spectrum occupancy is the statistics of active and idle times defined by Eqs. 11 and 12.

Doppler spectrum feature group represents the Doppler characteristics of the interference. Doppler spectrum is the Fourier Transform of the interference behavior in



(a) (b)

Figure 5-8: Normalized importance level of the 14 features extracted from empirical interference database in SVM classifiers in, (a) licensed bands, and (b) unlicensed bands.

time[61]. It is characterized by the shape of the spectrum, the RMS Doppler spectrum, and the energy in the Doppler spectrum, which Eqs 12, 13, and 14 define, respectively.

5.4 Results and Discussions

This section discusses the detailed analysis of the comparative interference behaviors between licensed and unlicensed bands for intelligent spectrum management in a vehicular platform. First, we provide an overview of the temporal and spatial behavior of the interference. Then, we demonstrate the performance of the fourteen features grouped into four feature groups described in the previous section. Finally, we present the predictability analysis, benefiting from machine learning algorithms trained interference feature, for intelligent spectrum management in a vehicular platform.

5.4.1 Temporal and Spatial Behavior of the Interference

We analyze the normalized data to understand the temporal and spatial behavior of the interference. First, we observe the spectrogram of the collected signals to identify the differentiating factors between licensed and unlicensed bands. We provide an example of the complete drive in Figures 5-3a and 5-3b. The spectrogram measurement shown in Figure 5-3a starts at 0 sec and ends at 720 sec, and the power is between

-20 dBm and -70 dBm. Note that the location of a moving vehicle changes over time, and thus in Figure 5-3a, different times also represent different locations. We observe that licensed bands experience higher interference at locations closer to the cell tower (around 504 sec) than on the open road (0 – 216 sec). Due to the similar variation in power, we observe a similar pattern of interference at 1.9 GHz – 2.0 GHz and 2.1 – 2.2 GHz. However, we observe discontinuous interference at the 2.4 GHz unlicensed band, and the power level is significantly lower than that of the licensed bands. Fig. 5-3b represents the congested plot of the data and shows the interference between 1.9 GHz and 2.5 GHz. We found the 2.0 GHz – 2.1 GHz and 2.2 – 2.3 GHz mostly inactive bands during the experiment, which shows the potential of spectrum-sharing applications.

5.4.2 Interference Feature Analysis

Statistical Features of Interference Intensity. Fig. 5-4a shows the interference intensity in licensed downlink band 2 (1930-1990MHz) along the route. The range of the interference intensity in this licensed band is [-20, -70] dBm. The color of each dot indicates the interference intensity at that location and time, and the actual strength can be found in the color bar on the right side of the map. Figure 5-4b compares the interference intensity in this licensed band channel, the blue line, with the interference in the unlicensed band at (2407MHz), the orange line. We observe that the licensed band has significantly greater interference intensity during the data-collecting process. The unlicensed band interference demonstrates the opportunistic nature of the distributed random medium access of Wi-Fi with substantial idle time in transmission. The centralized medium access of downlink cellular demonstrates a much more efficient utilization of the channel. Similar to previous works, we observe that our data also fits the Beta distribution.

Figure 5-4c shows the empirical CDF of interference intensity of all bins of three downlink licensed and one unlicensed band, shown in Table 5.2, against best-fit Beta distribution for each of them (Eq. 1).

Other five interference intensity features for intelligent spectrum management of

the empirical data for the three licensed and one unlicensed band are shown in Figure 5-7.

Correlation Properties of Interference. Time and frequency-domain correlation properties are beneficial to predict channel state, or spectrum availability based on past measurements [77]. We calculate the time correlation of short time series with a size of 50 samples in the same frequency bin. For example, Fig. 5-6a shows sample time correlation for 10 consecutive 50 samples of interference intensity in licensed Band 2. The complete comparison of all bins in three licensed and one unlicensed band (Table 5.2) is shown in Fig. 5-7, and they vary between 0-0.8.

To measure frequency-domain correlations, we calculate the correlations of all 40 frequency bins in Band 2 as an example. Fig. 6b is the frequency correlation heat map in the 1930 – 1990 MHz band. According to the frequency correlation, we can divide this band into four groups: 1.93 – 1.95GHz, 1.95 – 1.97GHz, 1.97 – 1.98GHz, and 1.98 – 1.99GHz. We observe that the frequency bins in each group are highly correlated, and the frequency correction between neighboring bins depends on which group it belongs to. Fig. 5-7 shows that licensed bands have a much higher frequency-domain correlation than unlicensed bands.

Channel Occupancy Statistics Fig. 5-5a presents the channel occupancy in 1.9 – 2.5GHz in one unlicensed and three licensed bands shown in Table 5.2. We observe that the channel occupancy of the unlicensed band, centered at 2.45 GHz, is significantly lower(0.1) than the three licensed bands centered at around 2, 2.2, and 2.35GHz (>0.5).

The shape of the duration and inter-arrival time distributions are features for predicting the channel availability. The empirical CDF of busy duration and inter-arrival time are provided in Fig. 5-5b and 5-5c, respectively. We observe that the licensed bands tend to have longer busy duration, while the unlicensed band has much longer inter-arrival time.

The following arguments explain the substantial difference in utilization as we observe them in a vehicle. The Wi-Fi dominantly produces the unlicensed band traffic

with a TDD distributed random medium access mechanism, while cellular traffic in licensed bands benefits from a more efficient FDD centralized medium access control. In addition, coverage of the Wi-Fi is significantly lower than the cellular base stations; as a result, a mobile vehicle observes the traffic associated with the Wi-Fi access points in shorter intervals. A third factor is that cellular base stations are deployed in grids to minimize the interference among them, while Wi-Fi deployments follow a random pattern without interference coordination[78].

Doppler Spectrum Analysis Fig. 5-6c shows the Doppler spectrum of a licensed band (in blue) and an unlicensed band(in orange). The licensed band shows a double-ear shape in the center, while the unlicensed band is flatter. Using Eqs. 13 and 14, we calculated the RMS Doppler spread and the energy of the Doppler spectrum. Figure 5-7 shows that the RMS Doppler spread in the unlicensed band is more than in 3 licensed bands.

5.4.3 Predictability Analysis for Intelligent Spectrum Access

Recent works on predictability analysis of channel availability with auto-regressive predictors in heavy wireless traffic have attracted considerable attention [71, 77]. In this section, we demonstrate the performance of our proposed features for predicting interference. To achieve this goal, we develop a machine learning model that predicts the interference/availability in the licensed and unlicensed bands and can operate as a backbone access of vehicular networks.

We first analyze the importance of each of the fourteen features described above in predicting the availability in both licensed and unlicensed bands. Fig. 5-8a and 5-8b show the normalized feature importance for the licensed and unlicensed bands, respectively. Channel occupancy contributes most to predictability for the licensed band and average interference intensity for unlicensed bands. However, channel occupancy dominates the effectiveness in licensed bands, and three intensity and two Doppler characteristics play significant roles (Fig. 5-8a). Correlation features are less effective for licensed bands. In unlicensed bands, four features, including

channel occupancy, RMS Doppler spread, the shape of the Doppler spectrum, and fade duration, dominate the Channel occupancy (Fig. 5-8b). Spectrum occupancy, correlation characteristics, and four intensity characteristics play minor importance.

Using the 14 features in 4 feature groups, we train two Support Vector Machine (SVM) classifiers, one for licensed and another for unlicensed bands. This classifier takes a 14-dimensional input. The output of the classier is the two classes, where the classes represent the channel that is either busy or idle. We split the dataset of 4000 samples into non-overlapping train and test samples having 75% and 25% of measurements. Consecutively. The test set determines the accuracy of the classifier, and we utilize this accuracy as the predictability metric. We also measure the channel availability, which is the one-channel occupancy. One channel occupancy happens when the measured amplitude is continuously greater than the threshold. We measure this channel occupancy using Eq. 9.

We achieve 96% predictability for unlicensed bands and 85% predictability for licensed bands. Interestingly, despite 90% channel availability (1-channel occupancy) in the unlicensed bands, the licensed bands demonstrate 20% availability. This wide gap is narrowed down in predictability because other features of the interference also contribute to predictability.

5.5 Summary

In this chapter, we introduced an interference monitoring system for vehicular interference monitoring to collect a database of interference in licensed and unlicensed bands in the busy 1.9 – 2.5GHz spectrum in a typical path in downtown Worcester, MA. We extracted fourteen statistical features of the interference in licensed and unlicensed bands divided into four classes: interference intensity, correlation properties, spectrum occupancy, and Doppler spectrum. Then, we benefited from an SVM machine learning algorithm for the predictability analysis of the licensed and unlicensed bands. We demonstrated that the predictability of unlicensed bands is 96% while licensed bands have 85% predictability. The predictability of unlicensed bands was highly correlated

with channel occupancy, while other interference features played more significant roles in predictability in licensed bands.

Chapter 6

PART II: An Empirical Study of Mid-Bands RII for Spectrum Sharing and Mobility Support in 6G Licensed and Wi-Fi Unlicensed Bands

This chapter presents an extensive empirical study of measuring and modeling RII in the most popular segment of the mid-bands in the presence of mobility. We focus on 1.9 – 2.5GHz, which includes popular 6G 1.9GHz licensed and Wi-Fi 2.4GHz unlicensed bands. To capture the spatiotemporal behavior of the RII, we drive a mobile interference monitoring system on a typical route in downtown Worcester, MA, and tag the measurements with their GPS location and timestamps. We repeat measurements over the route ten times in one day and repeat the entire experience the same day of the following week. Benefiting from the classical statistical models for the received signal strength (RSS), we present a novel theoretical foundation for the spatiotemporal behavior of RII and explain the need for machine learning (ML) algorithms for modeling. To train and validate the machine learning models, we collect a large dataset of fourteen statistical features of the RII categorized into four classes (interference intensity, correlation properties, spectrum occupancy, and

Doppler spectrum) in licensed and unlicensed bands. We train three regression machine learning models, Support Vector Machine (SVM), Random Forest (RF), and XGBoost (XGB), to regenerate RII and analyze the availability of the licensed and unlicensed bands for spectrum sharing and mobility support. We demonstrate that our models can regenerate average RII intensity within a standard deviation of 5.76 dB and predict channel availability with 93.79% accuracy for one-step predictions and 91.84% accuracy for duration predictions, respectively.

6.1 Introduction

Today, the importance of mobility support for autonomous vehicles and the emergence of the Internet of Everything (IoE) [79] has heightened the need for a new paradigm in spectrum management and regulation for sharing and managing the physically limited spectrum resources [80] for wireless mobile communications in cellular 6G, Wi-Fi 7, and beyond. In the past couple of years, cellular operators such as AT&T, T-Mobile, and Verizon have heavily installed mid-band licensed 5G radios at 2-6GHz while popular Wi-Fi 6 and 7 operate in unlicensed segments of the same mid-bands [81]. Besides, the wireless communication industry is negotiating with the FCC to share the 3.2-3.8GHz spectrum with the military and other license owners. However, the existing literature in wireless spectrum sharing for mobility support has not received adequate attention to the theoretical foundations and empirical studies of regenerating RII and predicting the channel availability in the mid-band [30].

The existing literature on regeneration of RII, also referred to as cartography [33, 82, 83, 84, 85, 34] relies on synthetic data generated with ray-tracing software[86] or empirical measurement at dedicated fixed locations surrounding a transmitting antenna [87]. Benefiting from various traditional and ML interpolation algorithms, they regenerate the RII in a small area surrounding an antenna location from measured values of RII in limited locations[88]. This approach is not scalable to large metropolitan areas to support mobility for spectrum sharing because it demands an unmanageable number of RII measurement sensors. Considering cost and feasibility, empirical mobile

RII monitoring via extensive street-level driving is crucial for improving large-scale spectrum monitoring and supporting mobility.

Though channel availability for spectrum management has been widely studied, these studies primarily focus on fixed-location network monitoring [88, 89, 90, 91, 71]. The current channel occupancy literature focuses on the RII intensity distribution instead of the channel availability analysis in a mobile scenario [92, 72]. This lack of support for mobility in RII monitoring hinders the support of beam forming with directional antennas and various traffic conditions, causes low spatial resolution of channel availability, and ignores the interference from low transmit power networks.

To bridge this gap between mobility support and channel availability analysis, we aim to model the spatial and temporal behavior of RII. To achieve this goal, a large empirical RII dataset at 1.9-2.5 GHz is first collected using our mobile spectrum monitoring system [31] in Downtown Worcester, MA. Then, we present a theoretical foundation for RII based on classic models for RSS behavior. We empirically demonstrate that traditional statistical models fail to regenerate and predict channel availability in a mobile scenario as they do not consider spatial diversity.

To develop a machine learning model for regenerating RII and predicting the channel's availability, we justify models for spatial shadow fading due to fixed civil infrastructure and temporal multipath fading caused by mobility. The empirical statistical features of RII are concluded for ML by dividing them into four groups: interference intensity, correlation properties, spectrum occupancy, and Doppler spectrum.

This study concludes with the development and validation of ML models to regenerate RII and predict channel availability. For mobility support, the velocity variations of the vehicle result in irregular spatial sampling, indicating varying distances between consecutive samples. Therefore, ML is employed to regenerate RII along the route, compensating for velocity changes through spatial interpolation. In practical applications, if another vehicle traverses a segment of the route, RII data for that area becomes accessible to passengers, regardless of their velocity. For RII regeneration, the theoretical bases are corroborated using Linear Regression (LR), K-nearest Neighbors

(KNN), RF, and XGB algorithms along the driving route. In the analysis of channel availability, 14 statistical features derived from RII measurements serve as inputs to ML classifiers, including SVM, RF, and XGB, to forecast channel availability. The empirical results, showing RII regeneration across the same route on different days, underline the spatial RII modeling's feasibility and highlight the significance of these features in predicting channel availability.

The remainder of the paper is organized into four sections: Section 6.2 details the theoretical foundations of RII behavior, referencing classical models for RSS. Section 6.3 elaborates on the empirical data acquisition system and the analysis of RII features. Section 6.4 shows the outcomes of ML-based RII regeneration and the predictability analysis of channel availability. Finally, Section 6.5 provides a summary and conclusions of the study.

6.2 Theoretical Foundations for the Effects of Mobility on RII

Allowing spectrum sharing without understanding the characteristics of the RF interference and its impact on the new and existing communication services can result in less intelligent spectrum management and regulations. This paper lays the theoretical foundation of RF interference in the mid-bands for mobility support by conducting a pioneering scalable empirical study of the statistical behavior of interference features. This study will shed light on bandwidth utilization, interference intensity, and statistical interference characteristics vital to machine learning for intelligent spectrum management and regulation.

Modeling of RF propagation in urban and indoor areas is an established practice in wireless communications networks, and all cellular and Wi-Fi standards organizations have traditionally recommended models for RF path-loss as well as multipath arrival profiles since the 1990s. More recently, they recommend channel state information (CSI) to support MIMO antenna systems [54]. Path-loss models calculate the RSS,

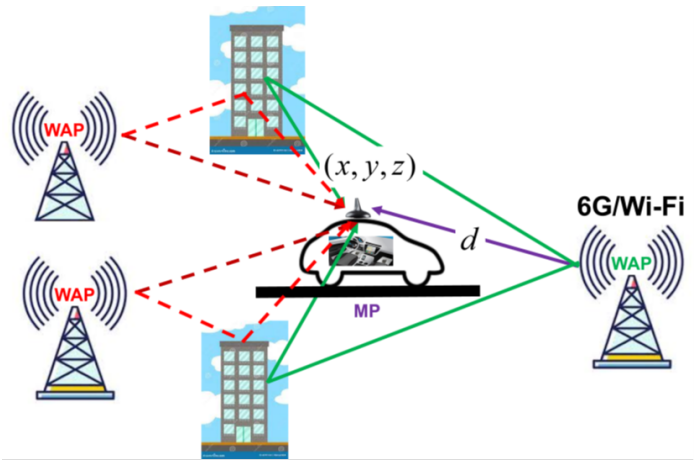


Figure 6-1: Multiple arriving signals through multipath contributing to RSS and multiple interference sources contributing to RII.

which has considerable similarities and differences with the received RII. Traditionally, RSS characteristics are studied through path-loss models, the statistical characteristics of spatial shadow fading caused by the urban fixed civil infrastructure, and the statistical behavior of temporal multipath fading caused by the mobility of the devices.

When an MP drives along a route, the motion route and velocity pattern of the vehicle, as well as the motion of other vehicles driving in close vicinity of the MP, affect the RII or RSS, resulting in slow varying spatial shadow fading and fast temporal multipath fading [54, 93]. By modeling this spatiotemporal fading characteristic, an MP, knowing its own velocity, can predict its average RII in a location identified by its GPS coordinate and predict or anticipate the average RII in the following location that the MP arrives at. To emulate the effects of temporal local mobility, MP can benefit from statistics of temporal multipath fading for the implementation of its intelligent ML-based spectrum management algorithms. Therefore, we proceed by establishing a theoretical foundation for modeling the effects of mobility on long-term spatial shadow fading and temporal local multipath fading.

6.2.1 Mobility and Spatial Shadow Fading Caused by Fixed Civil Infrastructure

Classical path-loss modeling allows for the RSS to be represented through a linear model. However, RII arises from all interference sources within a coverage area, necessitating a more complex modeling approach. Access to the power level, antenna pattern, and location of all interfering sources may not always be feasible. Therefore, RII can be modeled more effectively as a function of physical location, bypassing the need for detailed information on each interference source. Figure 6-1 depicts the contributors to RII for a mobile platform (MP) equipped with 6G and Wi-Fi receivers, navigating through an urban environment. This illustration highlights the influence of wireless access points (WAPs) and the multipath scenarios engendered by the architectural layout of the civil infrastructure on RII. RSS is the cumulative power received at an MP from L-multipath arrivals from a single WAP transmitter located at a distance, d, to establish a wireless communication link. The RSS can be determined by finding the magnitude square of the phasor addition of received signal amplitudes and phases from direct and all contributing L-reflected multipath arrivals [54]:

$$\text{RSS} = \left| \sum_{i=0}^{L} a_i \frac{\sqrt{P_0}}{d_i} e^{-\frac{2\pi d_i j}{\lambda}} \right|^2 . \tag{6.1}$$

Here, d_i and a_i represent the length and reflection coefficient of multipath arrivals, respectively. $d_0 = d$ and $a_0 = 1$ represent the direct path. λ is the carrier frequency wavelength, and P_0 is the RSS at unit distance between the MP and the WAP.

Figure 6-2 provides more details on comparing the formation of RSS and RII through multipath arrivals. Figure 6-2(a) shows a typical RSS spatial behavior in dB versus distance in logarithmic form, which exhibits serious fading conditions. In classical multipath RF propagation literature, we model the RSS spatial behavior in dB versus $10log(d)$, with a best-fit line with the slope of a, referred to as the distance power gradients:

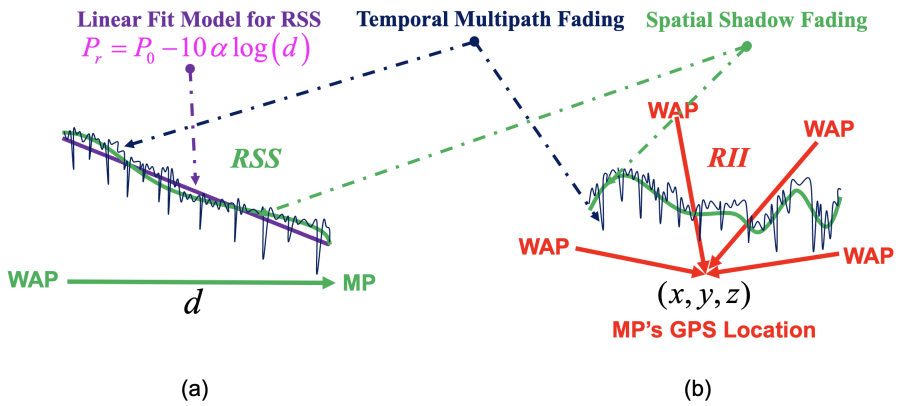


Figure 6-2: Comparison of RSS and RII measurements (a) RSS versus distance (in logarithmic scale) between a base station and a mobile with linear regression, redline, to model the average RSS, (b) RII versus distance traveled by a mobile vehicle along a route with redlines representing places that car stopped.

$$RSS(dB) = P_r(dB) = P_0(dB) - 10\alpha \log d + X(\sigma), \tag{6.2}$$

plus a zero mean Gaussian shadow fading, $X(\sigma)$, reflecting the spatial effects of multipath on average RSS, from the fixed infrastructure of buildings and the roads surrounding the WAP and the MP. The average RSS changes slowly in the scale of significant infrastructure variations as the device moves along. In addition to the slow-varying shadow fading of average RSS, fast temporal multipath fading of RSS arises from changes in multipath conditions caused by the local motion of the MP and other moving objects in the area.

More importantly, Figure6-1 and 6-2 also illustrate the differences between RII and RSS. As shown in Figure 6-2(b), when we measure RII for an MP located at coordinate, (x, y, z), $N\text{-}WAPs$ contribute to the received power each through a different multipath arrival:

$$\text{RII} = \left| \sum_{j=1}^{N} \sum_{i=0}^{L_j} a_{ij} \frac{\sqrt{P_{0j}}}{d_{ij}} e^{\frac{-2\pi d_{ijj}}{\lambda}} \right|^2 . \tag{6.3}$$

Theoretically, the slow-changing spatial variations of RII can be determined by calculating the complex addition of the aforementioned signals. However, in practice,

accessing the power level, antenna pattern, and location of all interfering sources is not always feasible. Therefore, the RII can be modeled with a 3D function of the GPS coordinates of the MP,

$$\text{RII} = f(x, y, z) + X(\sigma), \tag{6.4}$$

and resort to empirical study of the spatial behavior of the RII. As shown in Figure 6-2b, line fitting cannot be utilized to model the spatial behavior of RII for its regeneration. Machine learning algorithms offer themselves as the solution as they benefit from numerous physical and mathematical features of the received interference power to be more effective for regenerating the RSS [61]. Therefore, it is expected to be effective for the regeneration of RII. Machine learning models designed from empirical RII measurements along a route depend on the geometry of surrounding buildings and the location of interference along the route and are specific to that route. These models regenerate the RII along a specific route. To scale them to cover the interference in a metropolitan area or a nation, it is necessary to traverse all routes within these areas, akin to Google's method of covering every street to develop Google Maps from images collected along all routes.

While spatial modeling of the RII is completely irrelevant to the traditional linear regressive models for RSS, the temporal behavior of RII and RSS have amazing similarities.

6.2.2 Mobility and Temporal Multipath Fading

As shown in Figure 6-1, the arrivals of multiple interfering rays at MP from multiple interfering WAPs forming the RII, or multiple rays carrying the signal from a specific WAP to the MP, arrive from all directions with the same probability. Since the multiple arriving signals with equal probability from all directions form a complex Gaussian distribution, the amplitude, and phase of the received signal from Rayleigh and uniform distribution [35]. As a result, the amplitude of the RII, $f(a)$, follows Rayleigh distribution,

$$f(a) = \frac{a}{\Gamma} e^{-\frac{a^2}{2\Gamma}}, \quad a \geq 0 \quad . \tag{6.5}$$

Here, Γ is the mean-square amplitude of the channel gain factor. This distribution changes to a Lognormal distribution:

$$f(a; \mu, \sigma) = \frac{1}{a\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln a - \mu)^2}{2\sigma^2}\right). \tag{6.6}$$

When the channel bandwidth exceeds a hundred MHz, The most popular Doppler spectral density $D(\lambda)$ [1] recommended by 2 – 6G cellular industry follows a double ear shape:

$$D(\lambda) = \frac{1}{2\pi f_m} \left[1 - \left(\frac{\lambda}{f_m}\right)^2\right]^{-1/2}, \quad |\lambda| < f_m. \tag{6.7}$$

Here, $f_m = v/l$ is the maximum Doppler shift for motions with a maximum velocity of v, and l is the carrier frequency wavelength. The most popular Doppler spectrum recommended by Wi-Fi standards holds a Laplacian bell shape:

$$D(\lambda) = \frac{A}{1 + B f_m^2}. \tag{6.8}$$

Here, A and B are the shape parameters.

Utilizing these Doppler spectrum models enables the emulation of temporal interference fluctuations and the assessment of performance using real devices or through analytical evaluations. This aligns with methodologies utilized in the last several decades for the design and performance evaluation of wireless communications systems based on standard models for multipath propagation.

6.3 Empirical Measurement and Modeling of RII

Empirical RF propagation modeling relies on mobile monitoring systems to validate the theoretical foundations of propagation and design objective-oriented practical

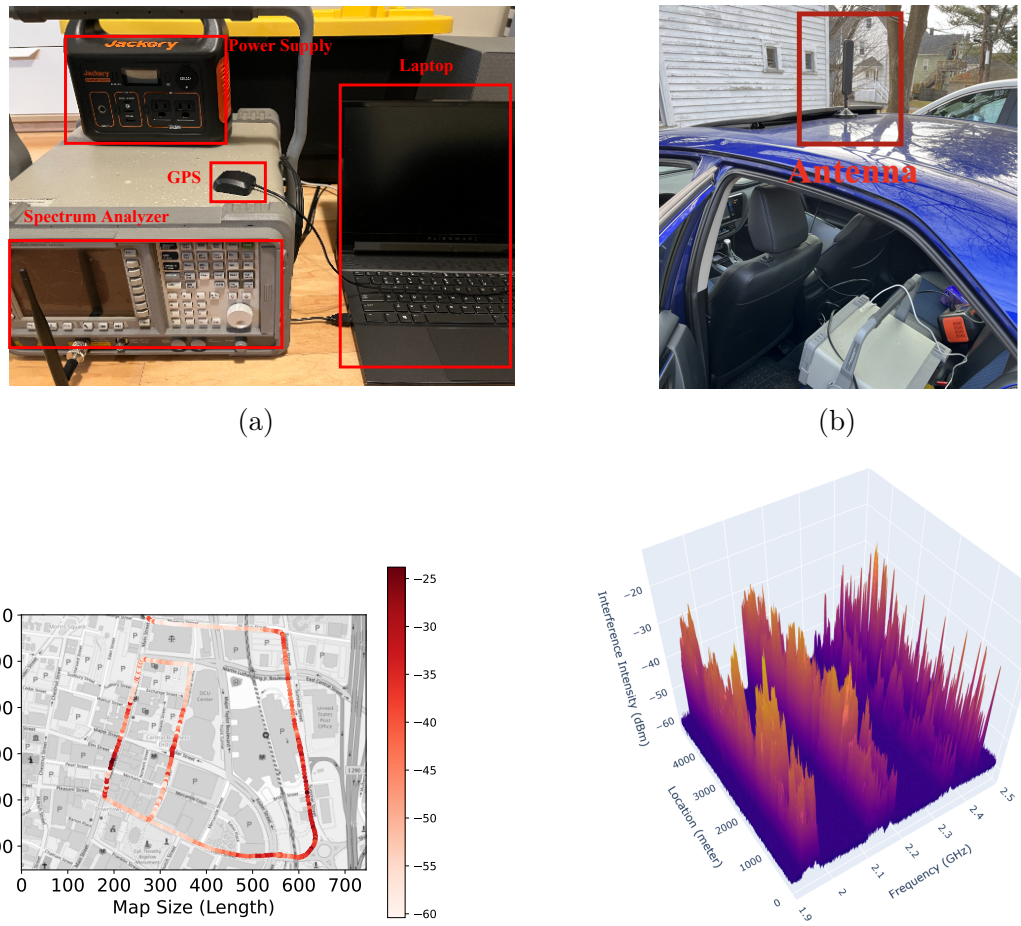
[1] the magnitude square of the Fourier transform of the temporal fluctuating samples

statistical models to regenerate the received signal features. Since the historical Okumura model [94], developed in the late 1960s to model path-loss in urban areas for calculation of the RSS, all models recommended by standard organizations are based on empirical mobile monitoring of RSS from a single WAP. These models regenerate the general statistical behavior of the RSS at different distances for coverage study and cellular deployment of the wireless communications networks [54]. Mobile RII monitoring is a new trend, which measures the interference from multiple sources for near real-time regeneration of RII intensity and traffic pattern under different spectrum regulations everywhere for spectrum sharing and traffic engineering needed for mobility support [31, 30].

6.3.1 RII Empirical Mobile Data Acquisition

Measurement System and the Route: Figure 6-3 shows the overview description of our RII data acquisition system and the measurement route in downtown Worcester, MA. Figure 6-3a shows the elements of our empirical interference data acquisition system on a laboratory bench, and Figure 6-3b shows the system mounted in the back seat of a car with an Ultrawideband (UWB) antenna covering mid-bands mounted on the top of the vehicle. The data acquisition hardware platform consists of an Agilent E4407B ESA-E Programmable Spectrum Analyzer, a GPS receiver, a laptop, and a portable power supply. Our laptop's proprietary Python user interface software controls the spectrum analyzer to measure any spectrum band from 9KHz to 26.5 GHz with 0.4 dB overall amplitude accuracy. The antenna connected to the spectrum analyzer measures different frequency bands so that the antenna pattern accommodates the frequency band of interest. The spectrum analyzer samples the RII in dBm at a rate of ten samples per second, and the GPS receiver provides the latitude and longitude of the location every second. The Dell M17 series laptop with a Core i7-9750h processor, 16 GB RAM, and 1 TB storage stores the results of measurement of the interference in each location tagged with the GPS coordination of that location.

Figure 6-3c depicts the route of the car in downtown Worcester, MA, chosen for RII monitoring. Figure 6-3d shows the associated overlaid measurements of samples of the



(a) (b)

(c) (d)

Figure 6-3: Data collection process in downtown Worcester (a) the data collection platform identifying different hardware elements; (b) the data collection system mounted in the back seat of a car with the UWB antenna installed in the roof; (c) the sample data collection route in downtown Worcester, MA with background map for the entire area. The red dot's intensity reflects the average interference intensity, and location is displayed according to logged GPS coordinates; (d) 3D representation of measured RF interference intensity in 1.9 – 2.5 GHz over the route: an intensity range of -27 – -63 dBm is observed from Cellular Band 2, 66, 30 and 2.4GHz unlicensed band [31].

interference along that route in the 1.9-2.5 GHz segment of the mid-band spectrum, which includes crowded 5G licensed cellular bands 2, 30, and 66, as well as the 2.4 GHz most popular unlicensed ISM bands for Wi-Fi.

This band selection enables us to compare the spatiotemporal characteristics of two popular bands at the extreme regulation ends. A wireless service provider owns the licensed full-duplex (uplink and downlink) bands according to FCC regulations

for Worcester, MA, and assigns segments of frequency and time resources available to individual users according to its proprietary resource allocation strategy. In other words, the service provider owner of the band manages the interference control in its proprietary domain in separate uplink and downlink channels. In the unlicensed bands, multiple agents owning different Wi-Fi access points and other devices operating in the band share the band in a fully liberated manner according to the power limitations regulated by the FCC. Two characteristics of these devices are that they often operate in half-duplex, uplink and downlink channels use the same carrier, and since the medium is shared with multiple agencies, we do not have an interference control mechanism according to the FCC spectrum regulations. However, most Wi-Fi devices operating in the band benefit from carrier sensing at the medium access control, which allows multiple agents to share the spectrum. The empirical study of the interference for these two classes of bands allows a fundamental understanding of the differences in traffic in the two extremes of the regulated, licensed, and unlicensed bands.

Raw Measurement Results: Figure 6-3c shows the scenario for measurement by identifying the route in Downtown Worcester. In performance evaluation of large-scale systems covering metropolitan areas, such as Wi-Fi positioning systems, it is a common practice to select a challenging area, such as the downtown of a city, for comparative performance evaluation of alternative algorithms for a task [95, 96, 97]. The selection of the paths is innovative and subject to patents [98]. In this paper, this tradition is followed. Our measurement aims to study the spatiotemporal behavior of the interference along this route at different times of the day on different days to find the minimum requirement for measurement to characterize the interference along a path. As we explained in our theoretical foundations, the RII is subject to a fixed average received power reflecting the architecture of the fixed civil infrastructure surrounding a location, which we refer to as slow-varying shadow fading and a fast multipath fading caused by the temporal motion of the device and other objects close to the transmitter and receiver antennas. To capture these effects with empirical measurements, we drive along the route in Downtown Worcester ten times during the peak traffic hour (7-9 AM) of one day of the week (Thursday, July 13, 2023), and we repeat the same

experience in another day (Thursday, August 17, 2023). This way, we have twenty sets of snapshots of RII in the 1.9-2.5 GHz segment of the midbands spectrum in 5000 locations along the route taken at 10 different times of a day on two different days. Each snapshot consists of a measurement of the RII in 400 frequency samples between 1.9 and 2.5 GHz. Figure 6-3d shows the 3D plot of all the raw measurements at 7 AM, July 13, 2023, to visualize the RII spatiotemporal characteristics in the raw empirical data.

Post-Processing of the Raw Data: In a mobile RII monitoring along the route, we sample the interference at 10 samples per second while the GPS device is sampling the location coordinates at one sample per second. In addition, each time we drive a route for empirical measurement of RII, we have a specific pattern for the velocity of the mobile that depends on the traffic pattern of the street at that moment. Regeneration or digital twining of the RF interference along the same route for another vehicle with a different velocity pattern, matching the traffic pattern of that moment, demands post-processing of the data. The only parameter that associates RII measurement in two drives is the location, and for that, we must also consider repetitive measurements at one location, such as a traffic light. We interpolate the data in the distance from the route's starting point for spatial digital twining of the RII and benefit from the data in stops at lights for statistical analysis of temporal variations caused by other moving objects around the transmitter contributing to the RII and the receiver.

Since snapshots of the spectrum are taken sequentially in ten samples per second and GPS locations are at one sample per second, we linearly interpolate the location of time samples to have a location fix for all time samples of the spectrum. Then, each data acquisition session along the route collects a set of data where the GPS location tag in Cartesian coordinate drawn from the longitude-latitude of a location is the time for measurement of a snapshot and represents the monitored frequency set of carriers of the spectrum.

$$f(a;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} a^{\alpha-1} e^{-\beta a} \tag{6.9}$$

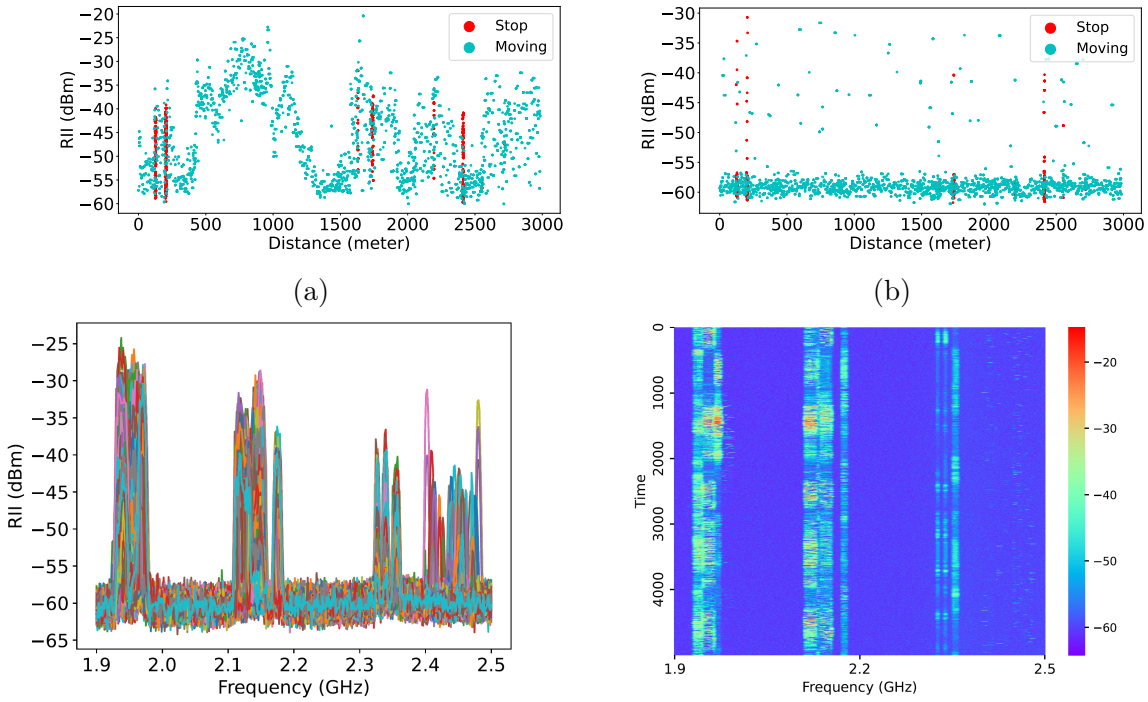


Figure 6-4: (a) Sample measurement of RII in one licensed band along the path, (v) Sample measurement of RII in one unlicensed band along the path, (c) overlay of all 5000 measurements on July 13, at 7 am, (d) spectrogram of the data on the same time relating intensity to time and frequency characteristics.

Figure 6-4 demonstrates three different views of the post-processed data. Figure 6-4a shows a sample measurement of RII at one licensed frequency along the path as a function of distance traveled. The redlines in locations with multiple measurements identify locations where we have stops with multiple measurements in one location. Figure 6-4c shows the overlay of all 5000 measurements on July 13, at 7 AM, demonstrating frequency utilization in different segments of the 1.9-2.5 GHz in the three licensed bands and one unlicensed band. Figure 6-4d shows the spectrogram of the data on the same three unlicensed and one unlicensed segment of the band. We benefit from the post-processed data to analyze the statistical features of the RII.

6.3.2 Statistical analysis of RII Features

The theoretical foundations for RF propagation rely on statistical models for the temporal and long-term fluctuations of the received RF signal level at different frequency bands to enable us to emulate these fluctuations in a repeatable laboratory envi-

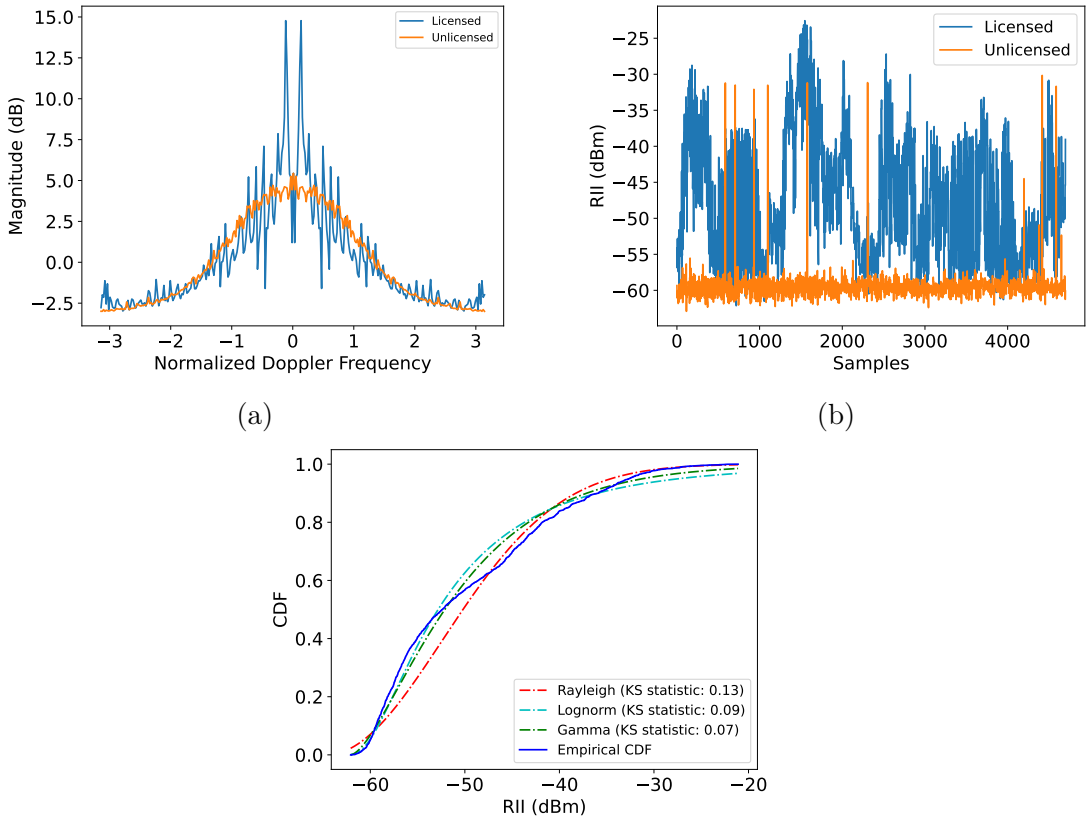


Figure 6-5: Validation of short-term fluctuations of the RII with empirical data, (a) Doppler spectrum, (b) a data set for RII short-time behavior in licensed and unlicensed bands, (c) best-fit distributions to data with Rayleigh, Lognormal, and Gamma-distributions.

ronment for comparative performance evaluation of alternative solutions for wireless communications. Every wireless communication standard organization recommends theoretical statistical channel models of RF signal fluctuations to predict the behavior of the channel in different application environments. However, the current statistical modeling of RF interference behavior for spectrum sharing and mobility support is in its infancy and is insufficient. As explained in section 6.3.1, we cannot model the average RII variations with the classical shadow fading analysis for modeling the RSS. However, theoretically, we expect short-term channel variations due to multipath fading caused by mobility in the environment with Rayleigh and Lognormal distributions for the amplitude and a double ear Doppler spectrum like those for RSS. As the first step in the empirical analysis of RII short-term variations, we benefit from the

empirical data to validate these theoretical observations for licensed and unlicensed bands.

Figure 6-5 shows the validation of short-term fluctuations of the RII with empirical data in licensed and unlicensed bands. Figure 6-5b shows the RII measurements in sample licensed and unlicensed bands. Figure 6-5a shows the empirical Doppler spectrum (add more discussion when you have it for both bands), which follows the double ear shape of the expected Doppler spectrum described by (6.7). Figure 6-5c compares best-fit distributions to data with Rayleigh, Rician, lognormal, and gamma distributions. Despite our expectations, the gamma function in (6.9) fits the RII better than expected Rayleigh distributions of (6.5) and the Log-normal distribution of (6.6), which reflects the imperfection of circular arrival of the interfering signal from different sources contributing to the RII.

In the remainder of this section, we present fourteen features of the RII that we benefit from in our study of ML application to regeneration and predictability of RII. We divide these features into four groups: interference intensity, correlation properties, spectrum occupancy, and Doppler spectrum.

Statistical Features of Interference Intensity. Figure 6-5a shows the interference intensity in licensed downlink band 2 (1930-1990 MHz) along the route. The range of the interference intensity in this licensed band is [-20, -65] dBm. The color of each dot indicates the interference intensity at that location and time, and the actual strength can be found in the color bar on the right side of the map. Figure 6-5b compares the interference intensity in this licensed band channel, the blue line, with the interference in the unlicensed band at 2407 MHz), the orange line. We observe that the licensed band has significantly greater interference intensity during the data-collecting process. The unlicensed band interference demonstrates the opportunistic nature of the distributed random medium access of Wi-Fi with substantial idle time in transmission. The centralized medium access of downlink cellular demonstrates a much more efficient channel utilization. Like previous works, our data also fits the Beta distribution. Figure 6-5c shows the empirical CDF of interference intensity of a downlink licensed band, shown in Table 6.1, against the best-fit Beta distribution for

each of them.

Correlation Properties of RII. Time and frequency-domain correlation properties are beneficial to predict channel state or spectrum availability based on past measurements. We calculate the time correlation of a short time series with a size of 50 samples in the same frequency bin. We calculate the frequency correction between the selected frequency bin and its neighboring bins to measure frequency-domain correlations.

Channel Occupancy Statistics. The channel availability is commonly calculated as the proportion of received power exceeds the pre-selected threshold [88, 89]. Some related works suggest the threshold to be 3-5 dB above the noise floor [90, 88]. Authors in [72] calculate a threshold by setting a fixed 5% false alarm rate. Our threshold is selected as three times the standard deviation of the shadow fading, which is also approximately 3 dB above the noise floor (-57 dBm). The collected RII is first binarized with the threshold, where '1' represents RII $>$ -57, and '0' represents RII $<$ -57. The channel occupancy is then calculated by counting the number of '1' s. We observe that the channel occupancy of the unlicensed band, centered at 2.45 GHz, is significantly lower (0.1) than the three licensed bands centered at around 2, 2.2, and 2.35 GHz ($>$ 0.5).

Doppler Spectrum Analysis. Figure 6-5a shows the Doppler spectrum of a licensed band (in blue) and an unlicensed band(in orange). The licensed band shows a double-ear shape in the center, while the unlicensed band is flatter. Using (6.7) and (6.8), we calculated the RMS Doppler spread and the energy of the Doppler spectrum.

6.4 Machine Learning for Regeneration and Availability

This section discusses the detailed analysis of the comparative interference behaviors between licensed and unlicensed bands for intelligent spectrum management in a vehicular platform benefiting from machine learning algorithms trained by fourteen

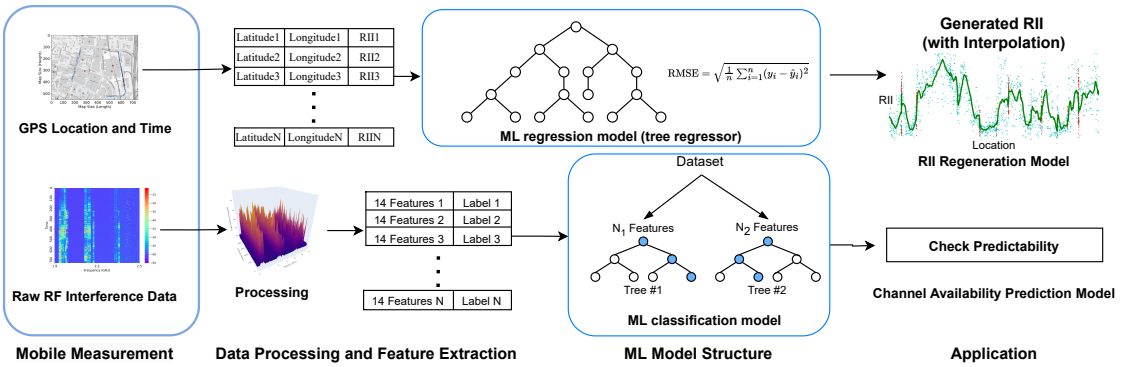


Figure 6-6: Overview of ML-based regeneration and channel availability prediction

interference features of the empirical data. We present the regeneration of RII for interference intensity analysis, followed by the predictability analysis for traffic engineering.

Instead of using traditional path-loss modeling to fit the RII data to a fixed curve, we propose a machine learning-based approach to reconstruct the RII of a given route to address the complexity caused by the fast-changing environment during driving. The top part of Figure 6-6 shows the general structure of the RII regeneration algorithm. This model can take location information as input and predict the average RII at that location. The performance metric is Root Mean Square Error (RMSE) calculated with empirical measurements and predictions. The proposed model takes location as input and predicts the average RII at that location. For training the model, RMSE is leveraged as the loss function where the ground truth is acquired from the empirical measurements.

Benefiting from the theoretical foundations introduced in Section 6.3.2, we propose a channel availability model that takes 14 features in Table 5.3 and outputs a binary prediction. The lower part of Figure 6-6 shows the workflow of this classification model. The raw data from mobile measurements is processed with the Synthetic Minority Oversampling Technique (SMOTE) algorithm to balance the 0 and 1 labels since the licensed and unlicensed bands have significantly more 1s and 0s, respectively (see Figure 6-3a and 6-3b). Then, the 14 features are calculated with a sliding window with a sample size of 200. With input the features, the model will output whether the

channel is available in the next slot.

6.4.1 Regeneration of RII Intensity with ML

Benefiting from ML regression models, the RII can be modeled similarly to the traditional path-loss model. Instead of giving average RSS in a range, the model takes GPS coordinates and outputs the estimated RII. Four frequently used regression models are evaluated: LR, KNN, RF, and XGB. KNN regression model predicts a new data point by calculating the weighted average of the neighbors of the target data point. This work selects the number of neighbors, k, as 5. RF and XGB regression models are selected because of the non-linearity and complexity between the 14 input features and the output RII [99]. These tree-based regression models are obtained by recursively partitioning the training data into small subsets to provide performance evaluation for the RII regeneration. The regression model is trained with a single test drive, which consists of approximately 5000 samples collected in 10 minutes, and tested with the remaining 19 drives (approximately 190 minutes measurements) since our observation shows that the number of drives does not significantly increase the RII regeneration algorithms' performance, because the similarities of average RII on the same route at different times, which will be discussed in the following paragraph. So, one drive is enough to train a regeneration algorithm. For the performance evaluation, we first show the RMSE of trained algorithms in different bands and then provide an interpolation example to show the value of the model in regenerating irregularly sampled data.

Result of RII Regeneration With the RII model, we can regenerate the RII along a given trajectory with a higher spatial resolution. In Figure 6-7, approximately 5000 measurements in 500 unique locations are leveraged as input (blue data points). After applying the RII regeneration algorithm, the pre-defined trajectory is created with a distance resolution of 0.1 meters. The data in red color is the regenerated RII map with 30000 RII samples in 30000 locations. This approach provides us with a tool to analyze the interference behavior in spatial aspects with irregularly sampled RII data. The difference between regenerated RII and the empirical measurement shows

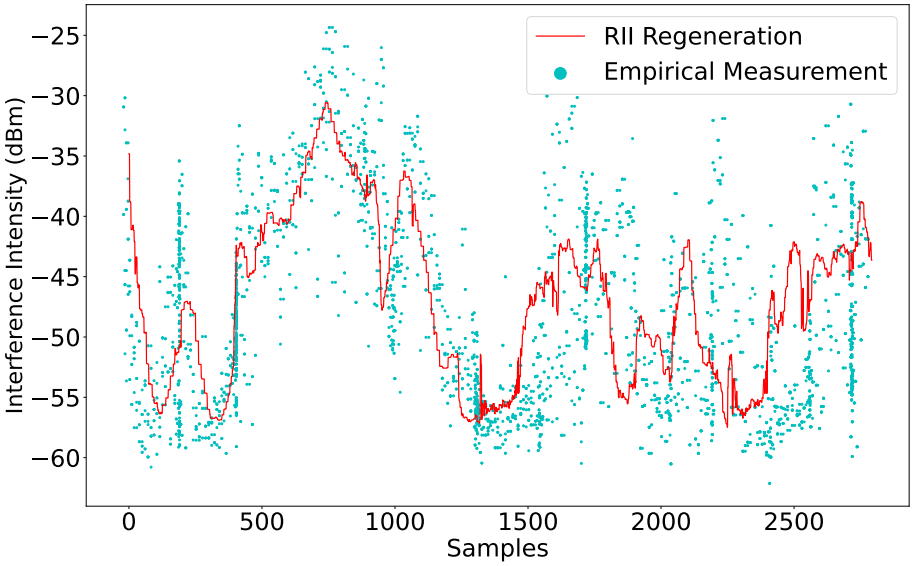


Figure 6-7: Regenerated average RII at each location in a licensed band vs raw data

the algorithm works better for licensed bands because CSMA traffic in ISM bands is difficult to predict and regenerate. For frequency bands with stable transmissions, the RII regeneration performance is better. The model in unlicensed bands needs more data to achieve a similar performance as licensed bands. The detailed RMSE comparison is provided in the following Table 6.1. With the best-performed XGB, the standard deviation of shadow fading is approximately 5.57 dB for RII regeneration.

Evaluation of ML algorithm for RII regeneration. Based on our observation, the error of this regeneration process is highly related to the frequency band and its channel occupancy. Some ML algorithms, like LR, show different performances in licensed and unlicensed bands. Table 6.1 compares the performance of 4 ML regressors in different bands. The RII regeneration algorithms show higher RMSE in frequency bands with heavy traffic (Band 2 and 66). XGB regressor is the best-performance ML model among all evaluated regressors. The RMSE is lower in ISM bands because most of the time, the channel is not occupied or our measurement system cannot capture the traffic due to the limited transmitting power and coverage of devices working in licensed bands, so the regression model is always predicting the RII in -57 - -60 dBm and result in a low RMSE.

Comparison of models trained in different times. To compare the regenerated

Table 6.1: RMSE of RII Regeneration in dB for Different Bands and Algorithms

	LR	KNN	RF	XGB
Band 2	6.0178	4.1494	4.1723	4.1608
Band 66	6.8477	4.5810	4.6227	4.6089
Band 30	3.2747	1.7259	1.7293	1.7229
ISM	2.7865	2.6838	2.6817	2.6776

Table 6.2: Predictability in Different Bands

	SVM	RF	XGB
Band 2	0.9912	0.9963	0.9984
Band 66	0.9866	0.9952	0.9957
Band 30	0.9501	0.9881	0.9862
ISM	0.9117	0.9594	0.9581

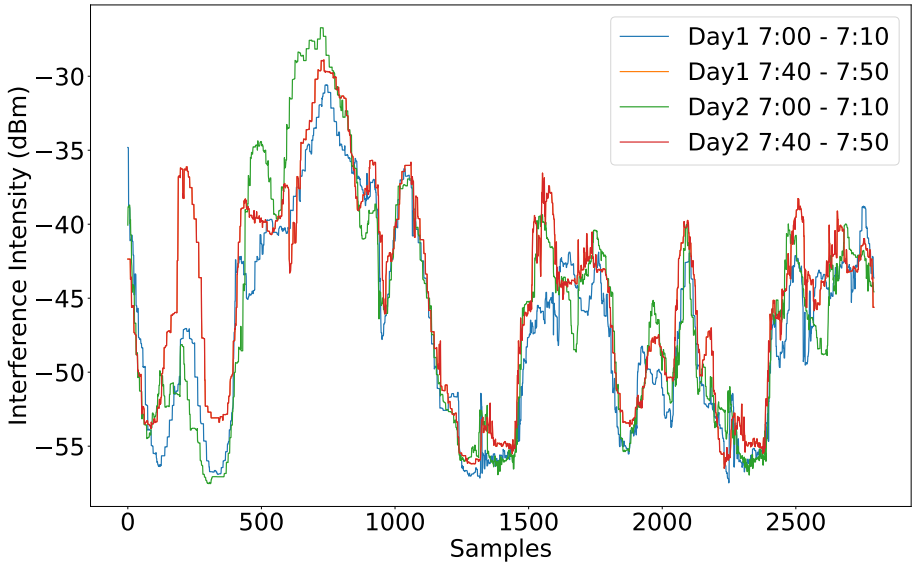


Figure 6-8: RII regeneration in different times (4 Drives)

samples at different times and days, RII ML models are trained with each individual test drive along the route. The regenerated RII models are compared in Figure 6-8. The models show a similar pattern or variations in most locations, which means driving multiple times along the same route is not necessary in practice; a single test drive is enough to capture the average RII behavior in an area. Due to the change in traffic conditions, the car's trajectory changes slightly in test drives. The resulting small-scale variations are also shown in Figure 6-8. Our result validates the theoretical foundations present in Section 6.2. The average RII is subject to a fixed average

Table 6.3: Predictability in Different Times of the Day and Different Days

	Day 1			Day 2		
	SVM	RF	XGB	SVM	RF	XGB
7:00 – 7:10 AM	0.8377	0.9239	0.9355	0.8110	0.9254	0.9288
7:10 – 7:20 AM	0.8347	0.9467	0.9487	0.8285	0.9398	0.9370
7:20 – 7:30 AM	0.8361	0.9180	0.9309	0.8343	0.9071	0.9133
7:30 – 7:40 AM	0.8395	0.9106	0.9158	0.8338	0.9251	0.9338
7:40 – 7:50 AM	0.8341	0.9312	0.9397	0.8314	0.9535	0.9603
8:00 – 8:10 AM	0.8328	0.9207	0.9380	0.8409	0.9339	0.9489
8:10 – 8:20 AM	0.8459	0.9281	0.9453	0.8508	0.9373	0.9413
8:20 – 8:30 AM	0.8627	0.9384	0.9439	0.8318	0.9282	0.9377
8:30 – 8:40 AM	0.8430	0.9284	0.9331	0.8348	0.9308	0.9444
8:40 – 8:50 AM	0.8390	0.9301	0.9400	0.8368	0.9385	0.9410

received power reflecting the architecture of the fixed civil infrastructure surrounding a location, so each time a vehicle passes the same street, the average RII tends to be the same. The remaining difference between each two RII models is due to the slow-varying shadow fading and fast multipath fading caused by the temporal motion of the device and other objects close to the transmitter and receiver antennas.

6.4.2 Predictability Analysis for Short-term Channel Availability

In this section, we present the predictability analysis in different bands, days of the week, and times of the day. The features are extracted from interpolated data and evaluated using three ML classifiers: SVM, RF, and XGBoost.

Table 6.2 shows the performance of the 3 ML classifiers in 4 bands. XGB and RF have comparable performance in all bands, but SVM performs worse in less occupied Band 30 and ISM bands. The classifiers should be carefully selected for applications in intelligent spectrum access based on frequency and channel occupancy. The average predictability for SVM, RF, and XGB is 83.70%, 92.98%, and 93.79%, respectively.

Table 6.3 shows the average performance of four ML classifiers. The predictability in Day 1 is calculated by input data collected in Day 1 to model trained with data from Day 2 and vice versa. All ML algorithms show slightly better performance on the second day. For SVM, the predictability on Day 2 is 1.9% greater than that on

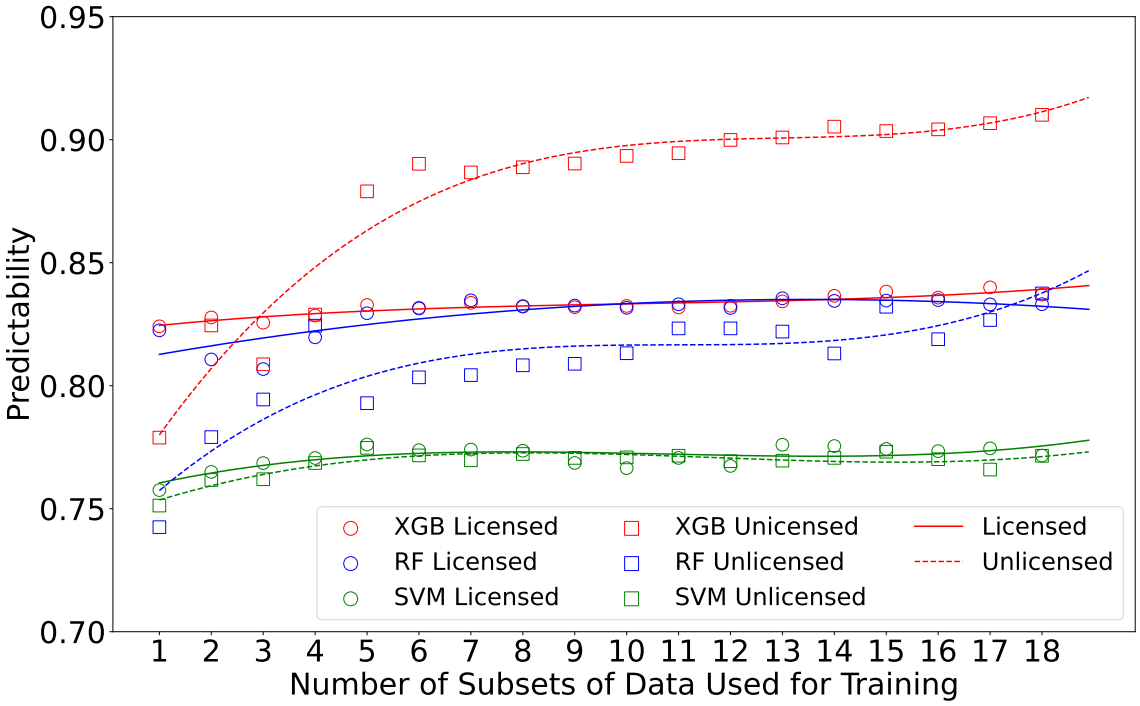


Figure 6-9: Number of drives used for training vs predictability

Day 1; for RF and XGB, the difference is 2.3% and 2.6%, respectively.

Figure 6-9 shows the impact of the number of drives used in the training process. ML models are trained with different numbers of test drives (from 1 to 18). The predictability is calculated by testing the trained models with the same two drives separate from training. For licensed bands, the predictability does not change significantly after training with 2 or 3 drives. The predictability requires more data (5 or 6 drives) for unlicensed bands to be stable. SVM shows about 77% predictability in all models. XGB shows slightly better performance (1%) in licensed bands than in RF but much better in unlicensed bands (7%).

To evaluate the importance of the features described in Table 5.3, a histogram is provided with the normalized feature importance in 4 different bands (See Figure 6-10). For band 2, the only dominant feature is the channel occupancy. For band 66, channel occupancy is still the most contributed feature, but the influence of average RII and fading characteristics are greater. Band 30 has a similar behavior as band 66 with less feature importance. For the ISM band, the three dominant features are channel

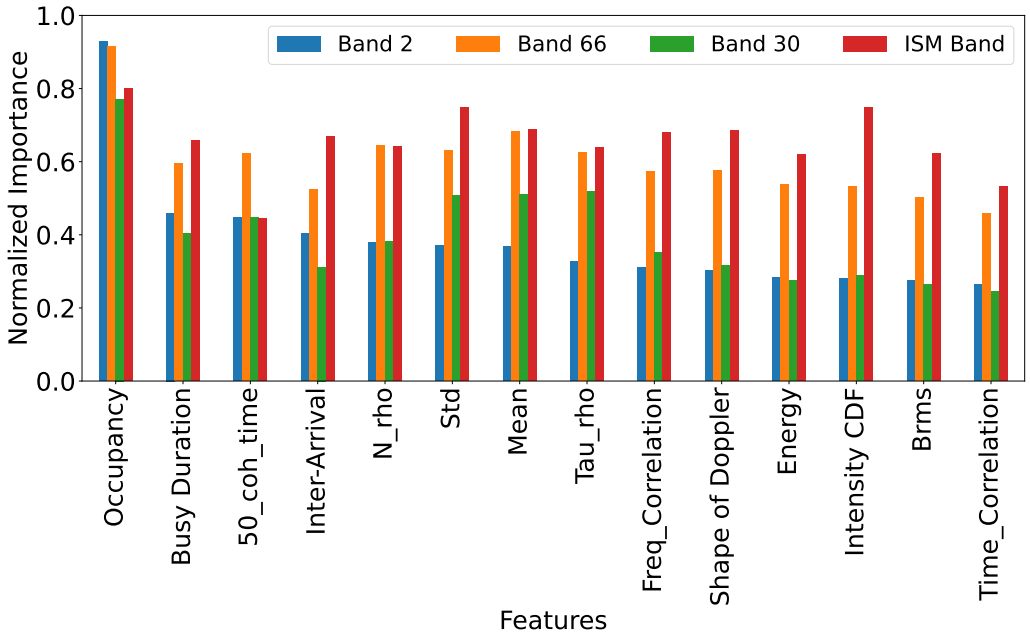


Figure 6-10: XGB feature importance in different bands

occupancy, RII CDF parameters, and standard deviation. However, the other features also have significant contributions to the prediction model. The channel occupancy is the most helpful feature for all 4 frequency bands evaluated.

6.4.3 Predictability Analysis for the Duration of Channel Availability

Since the channel availability duration plays a vital role in intelligent spectrum management, XGB models, leveraging the same 14 features, are trained to compare the actual and predicted idle duration of the channel availability. The model uses data from the preceding time slot with 200 samples to predict the maximum idle duration in the forthcoming slot, capping the output range at [0, 200] to focus on the selected future period. The 200-sample threshold for predictability was chosen because the maximum channel availability duration for the three licensed bands seldom exceeds this limit. Training was conducted on 80% of the dataset (48,000 samples) and testing on the remaining 20% (12,000 samples).

The empirical probability distribution function of the channel availability duration

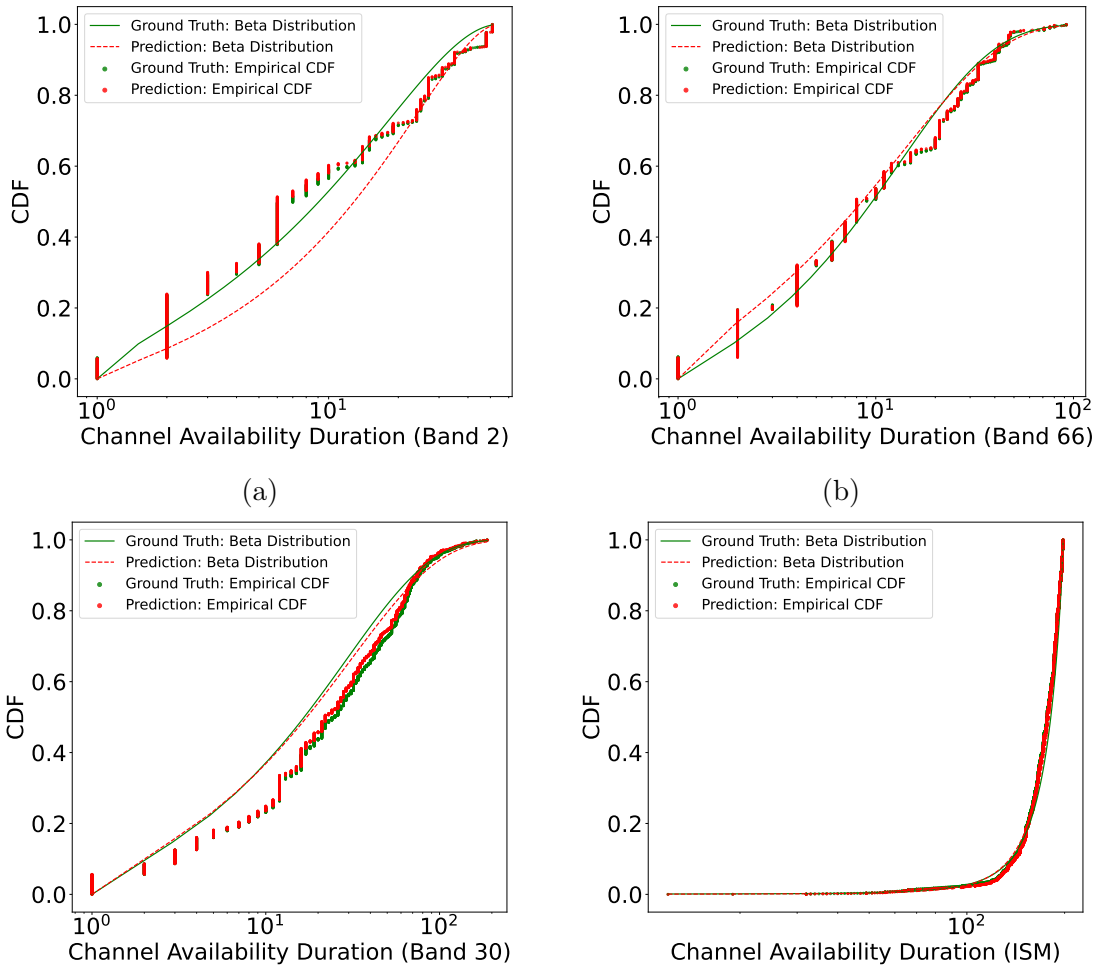


Figure 6-11: Channel Availability Duration Prediction: (a) Band 2 (b) Band 66 (c) Band 30 (d) Unlicensed band

Table 6.4: Channel Availability Duration Predictability

	Band 2	Band 66	Band 30	Unlicensed
p_0	0.8862	0.8713	0.7482	0
p_{200}	0	0	0	0.4210
α	0.6084	0.7986	0.6305	48.5884
β	2.1771	142.1061	11.3632	0.8393
Average Available Duration	1.5574	1.9680	8.2679	182.7671
Standard Deviation	6.3949	7.5145	21.0991	23.6213
Accuracy (XGB Classifier)	0.9809	0.9750	0.8902	0.8275
RMSE	1.4420	1.6916	3.3443	3.2087

D is then expressed by:

$$p_D(d) = \begin{cases} p_0, & \text{if } d = 0, \\ p_{200}, & \text{if } d = 200, \\ Beta(d; \alpha, \beta), & \text{else.} \end{cases} \tag{6.10}$$

Where p_0 is the probability that the channel is completely unavailable and p_{200} is the probability that the channel availability duration is more than one slot. The duration in [1, 199] can be fitted to a sample statistical distribution. The best-fitted distribution for the presented dataset is the Beta distribution, considering all four frequency bands. Table 6.4 presents the result of empirical probability distribution function parameters and the performance of the ML-based channel availability duration prediction. p_0 is 0 for unlicensed bands and p_{200} is 0 for licensed bands.

The empirical CDF plots for four frequency bands are displayed in Figure 6-11, where the alignment of points reflects whether predictions follow the actual distribution. In densely occupied licensed bands such as band 2 and band 66 (Figure 6-11a and Figure 6-11b), the data points align with the fitted Beta distribution at a shorter duration but significantly deviate at a longer duration. Specifically, the model accurately predicts short idle duration (0 – 25 for band 2, 0 – 50 for band 66) but overestimates longer duration (25 – 60 for band 2, 50 – 80 for band 66). Conversely, for the less occupied band 30 and the unlicensed band, the duration of channel availability is more evenly distributed between 0 and 200. For band 30 (Figure 6-11c), deviations from the fitted CDF start at a shorter duration and persist across the range, suggesting a consistent overestimation of idle duration. The unlicensed band (Figure 6-11d) closely follows the expected distribution, with minor discrepancies between 50 – 120.

Overall, the accuracy for channel availability duration prediction is 91.84% and the RMSE is 2.42 time slots. The trained models for licensed bands tend to over-predict the channel availability duration beyond certain thresholds, while the unlicensed band model under-predicts short-term availability. These variations are due to the differences among media access control (MAC) of wireless technologies and spatial

shadow fading from the fixed infrastructure, resulting in an uneven distribution of channel availability across the observed 200 samples. Cellular 5G and 6G technologies adopt mainly frequency-division-duplex central MAC, while Wi-Fi 6 and 7 benefit from time-division-duplex distributed MAC with carrier sensing.

6.5 Summary

To provide mobility support for future spectrum-sharing applications, we have presented the similarities between RSS and RII as the theoretical foundations for RII modeling in two aspects: spatial shadow fading due to fixed civil infrastructure and temporal multipath fading caused by mobility. We leveraged a mobile spectrum monitoring system, a spectrum analyzer installed in a vehicle, to validate the effectiveness of our theoretical foundations. The empirical measurement was conducted during rush hours in downtown Worcester, MA, resulting in 20 individual test drives with approximately 100,000 spectrum snapshots in the crowded 1.9 - 2.5 GHz mid-band. Our results based on ML-based RII regeneration analysis, including LR, KNN, RF, and XGB using 14 RII features into four groups, showed that the spatial behavior of RII can be modeled with an average standard deviation of 5.57 dB in the licensed bands and 3.41 dB for unlicensed bands using best-performed XGB. Next, we compared the models trained at different times and concluded that the average RII in the same route does not vary significantly over time, consistent with our theoretical foundations.

Finally, we are predicting the channel availability in different bands with 14 RII features into four groups. We compare different ML classifiers, including SVM, RF, and XGB. We observe that the most essential feature in the past few seconds is the channel availability. The average predictability obtained with XGB trained using a single test drive is 93.79%, better than RF (92.98%) and SVM (83.70%). We further study the effect of training with different numbers of repeated drives and show that the predictability does not increase significantly after training with 3 and 6 test drives for licensed and unlicensed bands, respectively. For channel availability duration prediction, we achieve an accuracy of 91.84% with an RMSE of 2.42 slots

Chapter 7

Conclusion and Future Work

In this dissertation, we presented theoretical foundations and the empirical validations for two challenging ML-based problems: proximity detection for social distancing and mobility support for spectrum sharing. Our analysis methodology for these two problems established a framework that integrates classical estimation theory with ML algorithms to address the challenges and concerns associated with applying ML techniques to RF cloud applications,

1) For proximity detection, our analysis methodology provided a framework for the empirical analysis of the estimation confidence for classical estimation theory and ML algorithms with BLE RSSI and two-way ranging UWB TOA. We provided a novel theoretical foundation with classical estimation theory using the CRLB to develop bounds for confidence on proximity detection, as a function of distance between two devices. We conducted empirical studies with a new empirical dataset for proximity detection, mirroring the structure of the PACT dataset, across diverse environments and considering various human postures and sensor placements. The GBM ML algorithm demonstrated that using the thirteen features can increase the confidence in the estimation of this social distance using BLE RSSI with an average confidence of 89.58%, which was 19.98% higher than the average confidence achieved using the classical approach. The theoretical foundations show that the average confidence of UWB TOA estimation for the 12 selected distances ranging from 3 to 12 ft is 96.98%, which is 1.58% better than utilizing ML-based BLE RSSI. Furthermore, the proposed

theoretical bounds have been confirmed to align with empirical results.

2) To provide mobility support for future spectrum-sharing applications, we presented the similarities between RSS and RII as the theoretical foundations for RII modeling in two aspects: spatial shadow fading due to fixed civil infrastructure and temporal multipath fading caused by mobility. Our comprehensive study on RII within the 1.9GHz – 2.5GHz vehicular interference monitoring system, we collected an extensive dataset during rush hours in downtown Worcester, MA, which facilitated a detailed analysis of the temporal and spatial behaviors of RII across both licensed and unlicensed bands. By examining fourteen statistical features of interference, categorized into intensity, correlation properties, spectrum occupancy, and Doppler spectrum, we have deepened our understanding of spectrum sharing with mobility support. Utilizing machine learning models, including LR, KNN, RF, and XGB trained with GPS location and empirical RII, we regenerated the RII along the route with an average RMSE of 5.57 dB for licensed bands and 3.41 dB for unlicensed bands. We used SVM, RF, and XGB to predict channel availability in different bands with 14 RII features. We demonstrated that the predictability of unlicensed bands is 96% while licensed bands have 85% predictability, with SVM. The average predictability obtained with the best performed XGB, trained on a single test drive, is 93.79%. We further studied the effect of training with different numbers of repeated drives and showed that the predictability does not increase significantly after training with 3 and 6 test drives for licensed and unlicensed bands, respectively.

Advancements in proximity detection will focus on refining ML algorithms to utilize complex features and sensor-fusion data more effectively. Future research could also develop cross-technology solutions that enhance the interoperability of various technologies like BLE, UWB, and Wi-Fi, paving the way for a universally applicable and robust proximity detection system. These efforts will significantly improve the accuracy and reliability of proximity detection technologies.

For intelligent spectrum management, the expansion of IoT and the 6G networks call for advanced, nationwide management systems capable of supporting extensive data communications. Future research could explore deep learning models such as

LLM and generative models to enhance spectrum cartography, offering sophisticated predictive and analytical tools to predict spectrum usage patterns and optimize frequency allocation. Additionally, collaboration with regulatory bodies is crucial to ensure that policies evolve alongside technological advancements and shifting user demands.

As one of the emerging areas of research, many more applications in the RF cloud are not equipped with solid theoretical frameworks and empirical studies. I hope this dissertation can inspire other researchers to dedicate their efforts to this area.

Bibliography

[1] K. Pahlavan et al. "RF cloud for cyberspace intelligence". In: *IEEE Access* 8 (2020), pp. 89976–89987.

[2] Jacopo Tosi et al. "Performance evaluation of bluetooth low energy: A systematic review". In: *Sensors* 17.12 (2017), p. 2898.

[3] Yuan Zhuang et al. "Smartphone-based indoor localization with bluetooth low energy beacons". In: *Sensors* 16.5 (2016), p. 596.

[4] Alessandro Montanari et al. "A study of bluetooth low energy performance for human proximity detection in the workplace". In: *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE. 2017, pp. 90–99.

[5] Yuri Assayag et al. "Adaptive Path Loss Model for BLE Indoor Positioning System". In: *IEEE Internet of Things Journal* (2023).

[6] Subha Viswanathan and Sreedevi Srinivasan. "Improved path loss prediction model for short range indoor positioning using bluetooth low energy". In: *2015 IEEE SENSORS*. IEEE. 2015, pp. 1–4.

[7] Ali Nikoukar et al. "Empirical analysis and modeling of Bluetooth low-energy (BLE) advertisement channels". In: *2018 17th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*. IEEE. 2018, pp. 1–6.

[8] Qiyue Li et al. "Cramér-Rao Bound analysis of Wi-Fi indoor localization using fingerprint and assistant nodes". In: *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*. IEEE. 2017, pp. 1–5.

[9] Xiaohua Tian et al. "Performance analysis of Wi-Fi indoor localization with channel state information". In: *IEEE Transactions on Mobile Computing* 18.8 (2018), pp. 1870–1884.

[10] Julang Ying, Kaveh Pahlavan, and Xinrong Li. "Precision of RSS-based indoor geolocation in IoT applications". In: *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE. 2017, pp. 1–5.

[11] Liyuan Xu et al. "UWB body motion assisted indoor geolocation with a single reference point". In: *2018 IEEE/ION Position, Location and Navigation Symposium (PLANS)*. IEEE. 2018, pp. 1509–1514.

[12] Christian Mensing and Simon Plass. "Positioning algorithms for cellular networks using TDOA". In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Vol. 4. IEEE. 2006, pp. IV–IV.

[13] Zohair Abu-Shaban, Xiangyun Zhou, and Thushara D Abhayapala. "A novel TOA-based mobile localization technique under mixed LOS/NLOS conditions for cellular networks". In: *IEEE Transactions on Vehicular Technology* 65.11 (2016), pp. 8841–8853.

[14] Bowen Wang et al. "BLE localization with polarization sensitive array". In: *IEEE Wireless Communications Letters* 10.5 (2021), pp. 1014–1017.

[15] Yishuang Geng et al. "Enlighten wearable physiological monitoring systems: On-body rf characteristics based human motion classification using a support vector machine". In: *IEEE transactions on mobile computing* 15.3 (2015), pp. 656–671.

[16] Kaveh Pahlavan, Xinrong Li, and Juha-Pekka Makela. "Indoor geolocation science and technology". In: *IEEE communications magazine* 40.2 (2002), pp. 112–118.

[17] Andy Liaw, Matthew Wiener, et al. "Classification and regression by randomForest". In: *R news* 2.3 (2002), pp. 18–22.

[18] Stefan Kalabakov, Aleš Švigelj, and Tomaž Javornik. "Smartphone Proximity Detection Using WiFi and BLE Fingerprinting". In: *2022 International Balkan Conference on Communications and Networking (BalkanCom)*. IEEE. 2022, pp. 36–40.

[19] Mimonah Al Qathrady and Ahmed Helmy. "Improving BLE distance estimation and classification using TX power and machine learning: A comparative analysis". In: *Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems*. 2017, pp. 79–83.

[20] Sheshank Shankar et al. "Proximity Sensing: Modeling and Understanding Noisy RSSI-BLE Signals and Other Mobile Sensor Data for Digital Contact Tracing". In: *arXiv preprint arXiv:2009.04991* (2020).

[21] Vivek Chandel, Snehasis Banerjee, and Avik Ghose. "ProxiTrak: A robust solution to enforce real-time social distancing & contact tracing in enterprise scenario". In: *Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international symposium on wearable computers*. 2020, pp. 503–511.

[22] Iacopo Carreras et al. "Comm2sense: Detecting proximity through smartphones". In: *2012 IEEE international conference on pervasive computing and communications workshops*. IEEE. 2012, pp. 253–258.

[23] Charu Jain, Gundepudi V Surya Sashank, S Markkandan, et al. "Low-cost BLE based indoor localization using RSSI fingerprinting and machine learning". In: *2021 sixth international conference on wireless communications, signal processing and networking (WiSPNET)*. IEEE. 2021, pp. 363–367.

[24] Cung Lian Sang et al. "Numerical and experimental evaluation of error estimation for two-way ranging methods". In: *Sensors* 19.3 (2019), p. 616.

[25] Yi Jiang and Victor CM Leung. "An asymmetric double sided two-way ranging for crystal offset". In: *2007 International Symposium on Signals, Systems and Electronics*. IEEE. 2007, pp. 525–528.

[26] M Marcus. "Regulatory policy considerations for radio local area networks". In: *IEEE Communications Magazine* 25.7 (1987), pp. 95–99.

[27] Kaveh Pahlavan. *Indoor Geolocation Science and Technology*. Denmark: River Publishers, 2019, pp. 104–109.

[28] Theodore S Rappaport et al. "Millimeter wave mobile communications for 5G cellular: It will work!" In: *IEEE access* 1 (2013), pp. 335–349.

[29] NOKIA. *5G spectrum bands explained — low, mid and high band*. 2023. URL: https://www.nokia.com/thought-leadership/articles/spectrum-bands-5g-world/ (visited on 07/03/2023).

[30] Kaveh Pahlavan et al. "Characteristics of Mid-Bands Interference in Wireless Spectrum Sharing for Mobility Support in 6G, Wi-Fi 7, and Beyond". In: *International Journal of Wireless Information Networks* 30.4 (2023), pp. 1–4.

[31] Zhuoran Su, Kaveh Pahlavan, and Bashima Islam. "An Empirical Study of Interference Features in Licensed and Unlicensed Bands for Intelligent Spectrum Management". In: *2023 IEEE 24th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. IEEE. 2023, pp. 252–260.

[32] Accenture. *Spectrum Allocation in the United States*. Sept. 2022. URL: https://api.ctia.org/wp-content/uploads/2022/09/Spectrum-Allocation-in-the-United-States-2022.09.pdf (visited on 12/12/2023).

[33] Yeduri Sreenivasa Reddy et al. "Spectrum cartography techniques, challenges, opportunities, and applications: A survey". In: *Pervasive and Mobile Computing* 79 (2022), p. 101511.

[34] Gregory D Durgin et al. "Digital spectrum twinning and the role of RFID and backscatter communications in spectral sensing". In: *2021 IEEE International Conference on RFID Technology and Applications (RFID-TA)*. IEEE. 2021, pp. 89–92.

[35] Kaveh Pahlavan. "Understanding of RF Cloud Interference Measurement and Modeling". In: *International Journal of Wireless Information Networks* (2021), pp. 1–16.

[36] *Structured Contact Tracing Protocol*. Boston, MA: MIT, 2020. URL: https://mitll.github.io/PACT/files/Structured%20Contact%20Tracing%20Protocol,%20V.%202.0%20(1.5).pdf.

[37] *Structured Contact Tracing Protocol*. Boston, MA: MIT, 2020. URL: https://mitll.github.io/PACT/.

[38] Shu Liu, Yingxin Jiang, and Aaron Striegel. "Face-to-face proximity estimationusing bluetooth on smartphones". In: *IEEE Transactions on Mobile Computing* 13.4 (2013), pp. 811–823.

[39] Kleomenis Katevas et al. “Finding dory in the crowd: Detecting social interactions using multi-modal mobile sensing”. In: *Proceedings of the 1st Workshop on Machine Learning on Edge in Sensor Systems*. 2019, pp. 37–42.

[40] Davide Giovanelli et al. “Bluetooth-based indoor positioning through ToF and RSSI data fusion”. In: *2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE. 2018, pp. 1–8.

[41] Xuan Bao et al. “Pinplace: associate semantic meanings with indoor locations without active fingerprinting”. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2015, pp. 921–925.

[42] Yiqiang Chen et al. “Surrounding context and episode awareness using dynamic bluetooth data”. In: *Proceedings of the 2012 ACM conference on ubiquitous computing*. 2012, pp. 629–630.

[43] Zhixian Yan, Jun Yang, and Emmanuel Munguia Tapia. “Smartphone bluetooth based social sensing”. In: *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. 2013, pp. 95–98.

[44] Zhenyu Chen et al. “Inferring social contextual behavior from bluetooth traces”. In: *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. 2013, pp. 267–270.

[45] Jiangchuan Zheng and Lionel M Ni. “An unsupervised learning approach to social circles detection in ego bluetooth proximity network”. In: *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. 2013, pp. 721–724.

[46] Kazushige Ouchi and Miwako Doi. “Indoor-outdoor activity recognition by a smartphone”. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 2012, pp. 600–601.

[47] Shin’ichi Konomi. “Colocation networks: exploring the use of social andgeographical patterns in context-aware services”. In: *Proceedings of the 13th international conference on Ubiquitous computing*. 2011, pp. 565–566.

[48] Avik Ghose, Chirabrata Bhaumik, and Tapas Chakravarty. “Blueeye: A system for proximity detection using bluetooth on mobile phones”. In: *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. 2013, pp. 1135–1142.

[49] Niklas Palaghias et al. “Accurate detection of real-world social interactions with smartphones”. In: *2015 IEEE International Conference on Communications (ICC)*. IEEE. 2015, pp. 579–585.

[50] Su Zhuoran, Zhang Pengyu, and Kaveh Pahlavan. “CSI-Based Motion Detection with Multiple Access Points”. In: *Adv. Sci. Technol. Eng. Syst. J.* invited paper, under preparation. 2020.

[51] Ruijun Fu et al. “Doppler spread analysis of human motions for body area network applications”. In: *2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE. 2011, pp. 2209–2213.

[52] Kaveh Pahlavan et al. “Indoor geolocation in the absence of direct path”. In: *IEEE Wireless Communications* 13.6 (2006), pp. 50–58.

[53] Kaveh Pahlavan et al. “RF cloud for cyberspace intelligence”. In: *IEEE Access* 8 (2020), pp. 89976–89987.

[54] Kaveh Pahlavan and Allen H Levesque. *Wireless information networks*. Vol. 93. John Wiley & Sons, 2005.

[55] S. Viswanathan and S. Srinivasan. “Improved path loss prediction model for short range indoor positioning using bluetooth low energy”. In: *2015 IEEE SENSORS*. 2015, pp. 1–4. DOI: 10.1109/ICSENS.2015.7370397.

[56] Zehua Dong et al. “A model-based RF hand motion detection system for shadowing scenarios”. In: *IEEE Access* 8 (2020), pp. 115662–115672.

[57] Qifan Pu et al. “Whole-home gesture recognition using wireless signals”. In: *Proceedings of the 19th annual international conference on Mobile computing & networking*. 2013, pp. 27–38.

[58] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[59] Marti A. Hearst et al. “Support vector machines”. In: *IEEE Intelligent Systems and their applications* 13.4 (1998), pp. 18–28.

[60] Sinan Gezici and H Vincent Poor. “Position estimation via ultra-wide-band signals”. In: *Proceedings of the IEEE* 97.2 (2009), pp. 386–403.

[61] Zhuoran Su, Kaveh Pahlavan, and Emmanuel Agu. “Performance Evaluation of COVID-19 Proximity Detection Using Bluetooth LE Signal”. In: *IEEE Access* 9 (2021), pp. 38891–38906.

[62] Oleksandr Semenov et al. “COVID-19 Social Distance Proximity Estimation using Machine Learning Analyses of Smartphone Sensor Data”. In: *IEEE Sensors Journal* (2022), pp. 1–1. DOI: 10.1109/JSEN.2022.3162605.

[63] Ahmad Hatami and Kaveh Pahlavan. “Performance comparison of RSS and TOA indoor geolocation based on UWB measurement of channel characteristics”. In: *2006 IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE. 2006, pp. 1–6.

[64] *DW1000, 3.5 - 6.5 GHz Ultra-Wideband (UWB) Transceiver IC with 1 Antenna Port*. Decawave, 2022. URL: https://www.decawave.com/product/dw1000-radio-ic/.

[65] Satyam Dwivedi et al. “Positioning in 5G networks”. In: *IEEE Communications Magazine* 59.11 (2021), pp. 38–44.

[66] Armin Dammann, Ronald Raulefs, and Siwei Zhang. “On prospects of positioning in 5G”. In: *2015 IEEE International Conference on Communication Workshop (ICCW)*. IEEE. 2015, pp. 1207–1213.

[67] Ryan Keating et al. “Overview of positioning in 5G new radio”. In: *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*. IEEE. 2019, pp. 320–324.

[68] Fuxi Wen et al. “5G positioning and mapping with diffuse multipath”. In: *IEEE Transactions on Wireless Communications* 20.2 (2020), pp. 1164–1174.

[69] *BEST-NEST Invitational Online Workshop on “New Paradigms in Intelligent Spectrum Management and Regulations, Future Directions, Technologies, Standards, and Applications*. Webinar from WPI. Worcester, MA, 2020.

[70] Ali Abedi et al. “Introduction to Special Issue on New Paradigms in Intelligent Spectrum Management”. In: *International Journal of Wireless Information Networks* 29.3 (2022), pp. 203–205.

[71] Yafei Hou et al. “Modeling and Predictability Analysis on Channel Spectrum Status Over Heavy Wireless LAN Traffic Environment”. In: *IEEE Access* 9 (2021), pp. 85795–85812.

[72] Bassel Al Homssi et al. “Free spectrum for IoT: How much can it take?” In: *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE. 2018, pp. 1–6.

[73] Farrukh Aziz Bhatti et al. “Shared spectrum monitoring using deep learning”. In: *IEEE Transactions on Cognitive Communications and Networking* 7.4 (2021), pp. 1171–1185.

[74] Keysight Technologies. *Keysight Technologies ESA-E Series Spectrum Analyzer*. 2023. URL: https://www.keysight.com/us/en/product/E4407B/%7Besae%7D-spectrum-analyzer-9-khz-to-265-ghz.html (visited on 01/16/2023).

[75] u-blox. *UBX-G7020 u-blox 7 GPS chips*. 2023. URL: https://content.u-blox.com/sites/default/files/products/documents/UBX-G7020_ProductSummary_%28UBX-13003349%29.pdf (visited on 01/16/2023).

[76] PyVISA Authors. *PyVISA Documentation*. 2022. URL: https://pyvisa.readthedocs.io/en/latest/.

[77] Dawei Chen et al. “Mining spectrum usage data: a large-scale spectrum measurement study”. In: *Proceedings of the 15th annual international conference on Mobile computing and networking*. 2009, pp. 13–24.

[78] Kaveh Pahlavan and Prashant Krishnamurthy. *Principles of wireless access and localization*. John Wiley & Sons, 2013.

[79] Alexandre Dolgui and Dmitry Ivanov. “5G in digital supply chain and operations management: fostering flexibility, end-to-end connectivity and real-time visibility through internet-of-everything”. In: *International Journal of Production Research* 60.2 (2022), pp. 442–451.

[80] Pedro Bustamante et al. “Unassigned Spectrum: An Institutional Analysis of Radio Spectrum Management”. In: *Available at SSRN 4528676* (2023).

[81] 5G America. *Mid-Band Spectrum Update: a 5G Americas White Paper*. 2023. URL: https://www.5gamericas.org/wp-content/uploads/2023/03/Mid-Band-Spectrum-Update-2023-Id.pdf (visited on 07/03/2023).

[82] Daniel Romero and Seung-Jun Kim. "Radio map estimation: A data-driven approach to spectrum cartography". In: *IEEE Signal Processing Magazine* 39.6 (2022), pp. 53–72.

[83] Sagar Shrestha, Xiao Fu, and Mingyi Hong. "Deep spectrum cartography: Completing radio map tensors using learned neural models". In: *IEEE Transactions on Signal Processing* 70 (2022), pp. 1170–1184.

[84] Li Li et al. "UAV Trajectory Optimization for Spectrum Cartography: A PPO Approach". In: *IEEE Communications Letters* (2023).

[85] Daniel Romero et al. "Aerial spectrum surveying: Radio map estimation with autonomous UAVs". In: *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2020, pp. 1–6.

[86] Subash Timilsina, Sagar Shrestha, and Xiao Fu. "Deep Spectrum Cartography Using Quantized Measurements". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[87] Henning Ids et al. "Spectrum cartography using adaptive radial basis functions: Experimental validation". In: *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE. 2017, pp. 1–4.

[88] Marko Höyhtyä et al. "Spectrum occupancy measurements: A survey and use of interference maps". In: *IEEE Communications Surveys & Tutorials* 18.4 (2016), pp. 2386–2414.

[89] Md Habibul Islam et al. "Spectrum survey in Singapore: Occupancy measurements and analyses". In: *2008 3rd International conference on cognitive radio oriented wireless networks and communications (CrownCom 2008)*. IEEE. 2008, pp. 1–7.

[90] Matthias Wellens, Jin Wu, and Petri Mahonen. "Evaluation of spectrum occupancy in indoor and outdoor scenario in the context of cognitive radio". In: *2007 2nd International Conference on Cognitive Radio Oriented Wireless Networks and Communications*. IEEE. 2007, pp. 420–427.

[91] Jiachen Sun et al. "Long-term spectrum state prediction: An image inference perspective". In: *IEEE Access* 6 (2018), pp. 43489–43498.

[92] Roel Schiphorst and Cornelis H Slump. "Evaluation of spectrum occupancy in Amsterdam using mobile monitoring vehicles". In: *2010 IEEE 71st Vehicular Technology Conference*. IEEE. 2010, pp. 1–5.

[93] Kaveh Pahlavan, Prashant Krishnamurthy, and Zhuoran Su. *Evolution and Impact of Wi-Fi Technology and Applications in Emerging Smart Word and IoT: A Historical Perspective*. Keynote speech for IEEE Spectrum Webinar. 2023. URL: https://event.on24.com/wcc/r/4094397/C59D7263C1361089DAB3FD501759159F (visited on 02/22/2023).

[94] Yoshihisa Okumura. "Field strength and its variability in VHF and UHF land-mobile radio service". In: *Review of the Electrical communication Laboratory* 16.9 (1968).

[95] Ferit Ozan Akgul and Kaveh Pahlavan. "Location awareness for everyday smart computing". In: *2009 International Conference on Telecommunications*. IEEE. 2009, pp. 2–7.

[96] K Pahlavan et al. "Taking positioning indoors Wi-Fi localization and GNSS". In: *Inside GNSS* 5.3 (2010), pp. 40–47.

[97] Kaveh Pahlavan. *Indoor Geolocation Science and Technology: At the Emergence of Smart World and IoT*. CRC Press, 2022.

[98] Edward James Morgan et al. *Location beacon database*. US Patent 7,433,694. Oct. 2008.

[99] Gareth James et al. "Tree-based methods". In: *An Introduction to Statistical Learning: with Applications in Python*. Springer, 2023, pp. 331–366.