Achieving Robust and Cost-effective Head Nodding and Body Gesture Recognition in Real-world Scenarios with WiFi-CSI Analysis

マルワー, レダ モハメド バスタウェシ

出版情報：Kyushu University, 2023, 博士（学術）, 課程博士
バージョン：
権利関係：

Kyushu University

Graduate School of Information Science and Electrical Engineering
Department of Information Science and Technology

Achieving Robust and Cost-effective Head Nodding and Body Gesture Recognition in Real-world Scenarios with Wi-Fi CSI Analysis

By

Marwa Reda Mohamed Bastwesy

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Supervised by

Professor Yutaka Arakawa

March 2024

Marwa Reda Mohamed Bastwesy

Achieving Robust and Cost-effective Head Nodding and Body Gesture Recognition in Real-world Scenarios with Wi-Fi CSI Analysis

Supervisor: Professor Yutaka Arakawa

Advisory Committee: Professor Yutaka Arakawa, Professor (Associate) Tsunenori Mine and Professor (Associate) Shogo Fukushima

Kyushu University

Graduate School of Information Science and Electrical Engineering

Department of Information Science and Technology

In the name of Allah, the Beneficent, the Merciful

Abstract

Wireless gesture recognition (GR) sensing systems, which leverage Wi-Fi Channel State Information (CSI) signals, have emerged as foundational technologies with profound implications for cutting-edge applications. These systems offer distinct advantages over traditional methods, such as vision-based and wearable sensors, due to their wide accessibility, user-friendliness, and incorporation of robust privacy protection mechanisms. Nevertheless, the current Wi-Fi CSI tools present certain limitations, including restricted device support, hardware compatibility issues, complex setup procedures, constrained data processing capabilities, and a lack of official support and updates.

In response to the aforementioned challenges, our research endeavors to address them through two specific applications: enabling communication between quadriplegic individuals and others, and monitoring workers' moods using gesture recognition. To attain these goals, we explicitly employ the ESP32 microcontroller, selected solely for its hardware compatibility, compact dimensions, energy efficiency, and cost-effectiveness. These features render the ESP32 well-suited for scenarios characterized by limited power and memory resources. Our work introduces three innovative and cost-efficient systems that not only demonstrate feasibility for real-world deployment but also exhibit robustness across diverse environmental conditions.

We begin by introducing the Wi-Nod system as a key component of a novel communication system tailored for individuals with quadriplegia. Wi-Fi CSI is utilized to encode Morse symbols, with head down and right motions representing dot and dash, respectively, and head left motion representing the new symbol, space. To establish distinct signatures for each head motion, the system employs Short Time Fourier Transform (STFT). Furthermore, a learning model based on the inception module is implemented to improve classification accuracy and diversity user robustness.

Next, we propose HeMoFi4Q which is the extention of the Wi-Nod system. HeMoFi4Q introduces a new communication method based on the combination between different Wi-Nod classified blocks to make the 26-alphabet characters. It utilizes Wi-Fi CSI waveforms to passively track head motions and derive distinctive gesture signatures for each character. Employing a real wheelchair and an ESP32 microcontroller, our approach diverges from previous HAR CSI systems by addressing domain-independent challenges in multi-human environments. Drawing inspiration from few-shot learning algorithms, we enhance location robustness by integrating samples from unseen environments during the learning phase. Our focus on domain independence involves studying the impact of amplitude and phase features, leading to improved recognition accuracy with minimal samples.

In addition to the quadriplegia communication systems, we introduce a passive desk body gesture recognition system aiming to autonomously identify the mood of a worker. Here, this system incorporates a multiple input multiple output (MIMO) configuration, employing three ESP32 microcontrollers sharing a common channel to enhance the reliability and robustness of data transmission. Additionally, the calibrated phase variations in the wavelet domain are fed into a straightforward machine learning model. This system aims to develop an on-desk body gesture recognition system on a single chip.

We conducted comprehensive experiments to evaluate the performance of the aforementioned three systems individually. The results demonstrated significant advancements in environmental robustness, particularly in multi-human context environments that closely resemble real-world scenarios.

Acknowledgements

I would like to express my profound and heartfelt gratitude to my supervisor Professor Yutaka Arakawa for his constant help, support, trust, and guidance, without which this work would not have been possible. His relevant, pointed, and constructive advice helped shape and guide my research, ensuring it reached its full potential. His insights, expertise, support, and constructive criticism have been invaluable to me, ensuring it reached its full potential. I feel fortunate to have had such dedicated and knowledgeable mentors.

I am honored to have Professor (Associate) Tsunenori Mine and Professor (Associate) Shogo Fukushima as advisory committee members. Their comments, advice, and feedback helped define and, when necessary, expand the scope of my work and helped ensure that it remained on track.

I would like to express my sincere gratitude to Dr. Hyuckjin Choi for his unrelenting support, precious advice, and kind mentorship during the course of my research. I am profoundly grateful for the time and energy he spent helping me conceptualise, conduct, and publish my research.

I am profoundly grateful for the invaluable assistance and guidance provided by Dr. Reem Elkhouly and Dr. Sherif Hashima, as without their support, it would have been inconceivable for me to embark on my Ph.D. journey in Japan, a leading country in the field of artificial intelligence. Their selfless dedication, tireless efforts, and extensive discussions have enlightened me on the path to studying in Japan, enabling me to maximize my potential as a researcher. I am deeply appreciative of their unwavering belief in my abilities and their continued support, which has motivated me to strive for excellence throughout my academic pursuit.

I was lucky enough to be surrounded by fantastic colleagues and fellow researchers with whom I shared memorable times; in particular Dr. Ristu

Saptono, Dr. Landy Rajaonarivo, Mr. Muhammad Ayat Hidayat, Mr. Min-Yen Lu, Mr. Kiichiro Kai, and Ms. Sana Ozono.

I would like to express my heartfelt thanks to Ms. Yuko Fukuda, Ms. Yukiko Hiranaka, and Ms. Yoshimoto Takayo for their help and support with everything.

I would like to extend my sincere appreciation to my friends, Ms. Samah Shady, Ms. Mariya Farag, Dr. Eman Farag, and Ms. Menna Fateen, for their invaluable assistance during my stay in Japan. Their unwavering support, both in personal matters and in maintaining my motivation throughout my Ph.D. journey, has been instrumental. I am immensely grateful for their continuous patience and unwavering encouragement.

I am profoundly grateful to extend my heartfelt appreciation to my parents, Mr. Reda Bastwesy and Mrs. Najat Gad, as well as my husband, Dr. Tarek Said, for their unwavering support, encouragement, unconditional love, and affection during the entirety of my doctoral research. Their unwavering commitment, invaluable guidance, and profound wisdom have served as a perpetual motivation for me to strive diligently towards achieving success in both my academic and practical pursuits. I am acutely aware that without their prayers, my journey in Japan would have been considerably more challenging

Lastly, I would like to express my profound gratitude to my beloved daughters, Arwa and Sara, with the sincere hope that I have been able to achieve accomplishments that bring them a sense of pride

Contents

List of Figures

List of Tables

Acronyms

1 Introduction

1.1 Background

Sensing technologies serve as the fundamental building blocks for numerous fields in modern society. Applications such as smart homes [1, 2], human behavior analysis [3, 4], human identification [5, 6], healthcare systems [7, 8], self-driving cars [9, 10], and others heavily rely on diverse sensing technologies for their operation and functionality. These technologies serve as the cornerstone for progress and innovation in various domains, enhancing convenience, efficiency, and security in our lives. Notably, sensing approaches utilizing computer vision and wearable sensors have demonstrated significant potential, benefiting from advancements in image processing and the capabilities of wearable sensor technology. The advancements in image processing methodologies have facilitated the emergence of computer vision-based sensing methods, including infrared and depth image sensors [11, 12, 13]. These approaches have garnered significant interest owing to their acceptable recognition accuracy and superior detection capabilities. However, camera-based systems require good lighting conditions and raise privacy concerns and bad performance in non Line-of-Sight (non-LOS) scenarios. On the other hand, wearable sensor-based systems [14, 15, 16] offer a lightweight and cost-effective solution. Nevertheless, they can be problematic if the user forgets to wear them, particularly in healthcare applications.

However, the implementation of these techniques in real-world scenarios poses challenges such as privacy concerns, limited coverage, and user inconvenience. Therefore, in this thesis, our objective is to tackle the challenges of privacy concerns and user inconvenience that arise when implementing sensing technologies in real-world scenarios. To address these issues, we present novel approaches in the healthcare domain that leverage Wi-Fi signals from low-cost microcontrollers. The aim is to develop an advanced gesture

recognition system capable of passively detecting head and body gestures in diverse locations, particularly in multi-human context environments. This innovative approach holds promise for practical applications in healthcare and offers potential solutions to the aforementioned challenges. This chapter introduces the research by presenting the research problems, objectives, and significance.

1.2 Research Problem and Questions

In recent years, there has been a significant surge in the development of Wi-Fi CSI-based sensing systems for passive Human Activity Recognition (HAR) and gesture recognition. However, these systems face limitations that hinder their practicality in real-world scenarios, despite some studies showing promise in addressing domain shift adaptation.

One notable limitation revolves around the reliance on open-source CSI tools developed by Halperin et al. [17] and Atheros CSI Tool [18] in existing Wi-Fi CSI-based HAR and gesture recognition systems. While widely used, these tools possess certain drawbacks, including limited device support and hardware compatibility, complex setup processes, limited data processing capabilities, and a lack of official support and updates.

Moreover, previous HAR CSI systems have attempted to address the challenge of environmental robustness by employing deep learning techniques such as few-shot learning. However, these systems struggle to adapt to real-world environments due to their tailored nature, primarily designed for specific user settings. This limitation becomes even more apparent when faced with the complexities posed by multi-human contexts within the sensing area.

The research problem under investigation involves addressing the limitations of existing Wi-Fi CSI-based systems in device support, hardware compatibility issues, complex setup procedures, and handling dynamic objects within the sensing area. The current focus on single-user environments neglects the real-world scenarios where users are present amidst multiple individuals. The challenge lies in developing techniques that can effectively disregard

scattering signals from surrounding people while extracting meaningful patterns solely from the target user.

The primary objective of this thesis is to address these limitations to improve the robustness and reliability of Wi-Fi CSI signals in practical sensing approaches. The aim is to enable the system to effectively handle real-world environments and expand the generalization and applicability of Wi-Fi signals in detecting different activities from various users in diverse multi-human context environments.

To achieve this, the thesis will address the following research questions:

RQ1 How can the Wi-Fi CSI system extract target user patterns while effectively filtering out scattering signals from surrounding individuals in a multi-human context?

RQ2 How can a general Wi-Fi CSI system be developed to detect gestures despite variations in their execution by different users?

RQ3 What approaches can be employed to configure ESP32 as a Multiple-Input Multiple-Output (MIMO) system, thereby increasing subcarrier resolution and enhancing overall system performance?

RQ4 How can signal interference and collisions be mitigated to ensure reliable Wi-Fi CSI sensing using ESP32 in a MIMO configuration?

1.3 Research Objectives

The primary objective of this thesis is to address the above questions and tackle the challenges associated with the deployment of Wi-Fi CSI sensing systems in diverse real-world environments with multiple human contexts. The research aims to develop robust and generalizable Wi-Fi CSI systems capable of handling the inherent dynamics and variability of such environments by investigating and developing techniques to address the variability of working conditions in real-world environments, enabling the Wi-Fi CSI system to adapt and perform reliably across different locations and contexts. To achieve this objective, the thesis will focus on two specific gesture use

cases: communication with quadriplegia and workers' mood recognition, utilizing the low-cost ESP32 microcontroller. These use cases have been selected due to their potential impact on improving the quality of life for individuals with quadriplegia and enhancing workers' mental well-being. The ESP32 CSI toolkit is utilized in a Single-Input Single-Output (SISO) configuration for a quadriplegia communication system and in a MIMO configuration for tracking a worker's mood and emotions.

The thesis objectives can be summarized as follows:

O1 **Explore methods to handle the variability of target users within the multi-human context, ensuring accurate gesture recognition and communication for individuals with quadriplegia.** Introduce Wi-Nod system that employs Wi-Fi CSI to detect Morse symbols, including dot, dash, and a new symbol called space, through head nodding gestures using time-frequency features. Evaluate the robustness of the system concerning user and session diversity in a fixed location with surrounding individuals.

O2 **Investigate techniques for addressing location diversity within a multi-human context, while minimizing the reliance on extensive data labeling from unseen environments.** This objective involves the development and evaluation of HeMoFi4Q, a novel sign language system built upon the Wi-Nod system. HeMoFi4Q aims to facilitate communication with quadriplegia patients by recognizing all 26 letters of the alphabet using Morse code and head movements within diverse spatial contexts. The objective focuses on achieving robust and accurate recognition of gestures in real-world scenarios, without the need for large amounts of labeled data specific to each location.

O3 **Develop a MIMO configuration for ESP32 to enhance subcarrier resolution and enable precise tracking of on-desk gestures.** Develop a contactless method to recognize a worker's mood based on body gestures using Wi-Fi CSI signals while the worker is situated at their desk. This objective focuses on extracting phase variations in the wavelet domain to mitigate interference and collisions of Wi-Fi CSI signals. By accomplishing this objective, the aim is to improve the

performance and reliability of the Wi-Fi CSI sensing system for on-desk gesture recognition.

O4 Evaluate the robustness and generalizability of the developed Wi-Fi CSI system through comprehensive experimental studies and comparisons with existing approaches, taking into account key performance metrics such as gesture recognition accuracy and system reliability.

1.4 Research Significance

The importance of this study is to maximize the potential of Wi-Fi CSI across diverse applications, with a specific focus on head gestures and body gestures as primary use cases. The main objective is to enhance the robustness of Wi-Fi CSI in real-world settings by introducing innovative techniques and approaches. One of the main challenges we address is the impact of multipath propagation on Wi-Fi CSI. In real-world environments, signal reflections and multipath effects caused by obstacles and reflections can distort CSI measurements. To overcome this, algorithms and signal processing techniques are proposed to effectively mitigate the negative effects of multipath propagation. This ensures more accurate and reliable CSI analysis. Another important aspect to investigate is the issue of location diversity. Real-world scenarios often involve multiple locations with unique characteristics and layouts. This presents challenges in training models and achieving consistent performance across diverse locations. To address this, we develop novel methodologies that adapt Wi-Fi CSI models to different environments, enhancing their generalization capabilities. Furthermore, we explore the complexities that arise in multi-human context environments. In such scenarios, the presence of multiple individuals introduces interference and signal variability in Wi-Fi CSI measurements. To tackle these challenges, we introduce techniques that enable more effective utilization of Wi-Fi CSI in multi-human contexts.

Each chapter's significance within this thesis is outlined as follows:

1. **The significance of the Wi-Nod system lies in its ability to generalize the Wi-Fi CSI for different users in a real-life scenario.** It leverages the frequency domain to mitigate environmental noise and scatter from

multiple individuals by generating signatures based on micro-head movements for Wi-Nod symbols. Verifying user and session diversity in a multi-human context environment provides valuable insight into the system's performance and adaptability.

2. **The significance of HeMoFi4Q system lies in its approach to simulate real-world scenarios more accurately and comprehensively, facilitating a thorough exploration of the system's effectiveness in diverse environments.** By employing the amplitude time domain as base signals and applying the efficient channel attention (ECA) model, this system aims to extract the most unique features for each character, thereby improving symbol recognition performance.

3. **Set ESP32 microcontroller in a MIMO configuration and overcome signal interference and collisions.** We have built the MIMO configuration for ESP32, enabling an increase in subcarrier resolution for tracking on-desk gestures. By extracting phase variations in the wavelet domain, this system aims to mitigate interference and collisions of signals, leading to improved performance and reliability.

1.5 Thesis Organization

The organization of this thesis is shown in Fig. 1.1. The remaining sections of this dissertation are organized as follows:

- Chapter 2: Fundamentals and Background: In this chapter, the fundamental concepts related to the background of the dissertation are introduced. This provides the necessary groundwork for understanding the subsequent chapters.

- Chapter 3: Overview of Wi-Fi CSI for Gesture Recognition: This chapter provides a comprehensive overview of various Wi-Fi CSI techniques used for gesture recognition tasks. It explores the existing literature and discusses the different approaches and methodologies employed in this field.

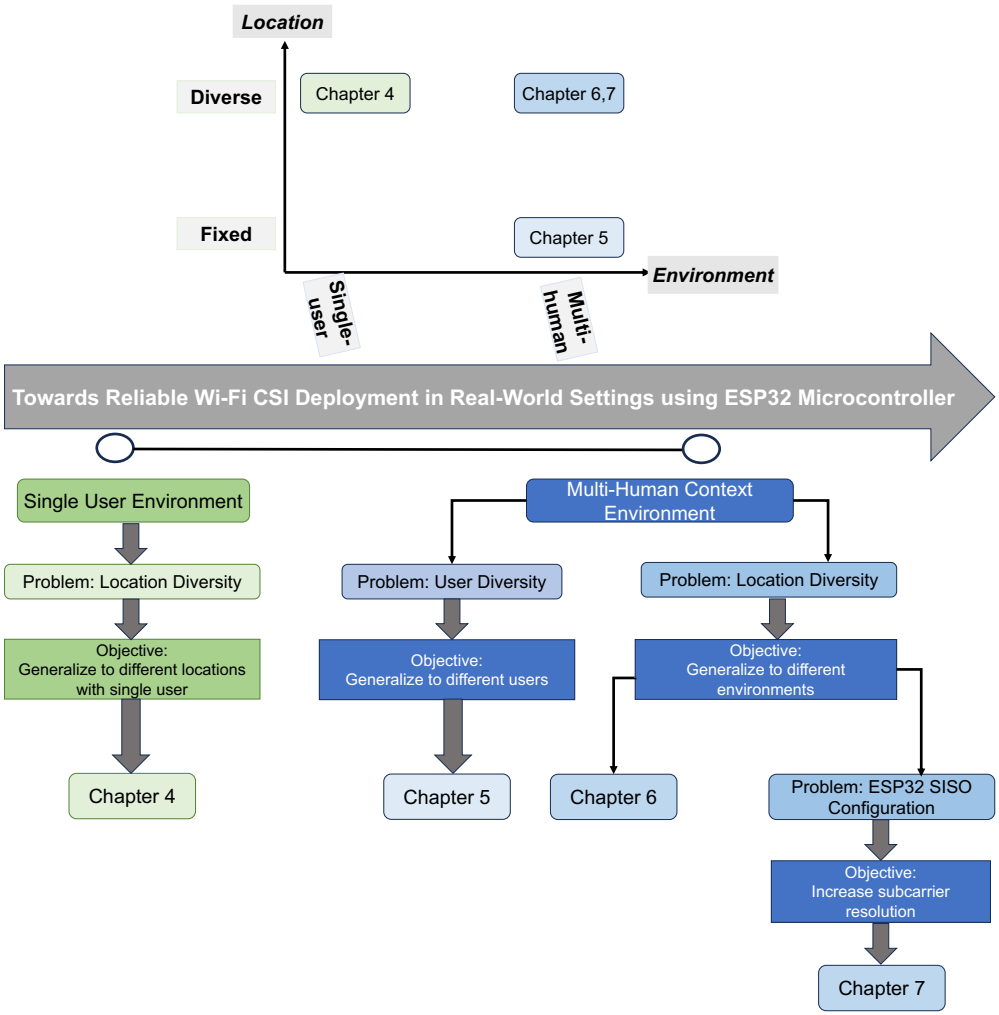


Figure 1.1: The overall challenges and objectives of the thesis are summarized. Chapter 4 provides a foundational understanding of existing Wi-Fi CSI systems and highlights the limitations associated with current CSI tools. In Chapter 5, the thesis introduces the Wi-Nod system, which addresses the issue of user diversity in the context of Wi-Fi CSI. Chapters 6 and 7 are dedicated to addressing the challenge of location diversity within a multi-context environment. Chapter 6 presents a novel system that effectively integrates a limited amount of data from previously unseen environments. Chapter 7 focuses on the implementation of the ESP32 microcontroller in a MIMO configuration.

- Chapter 4: Leveraging public Wi-Fi CSI sign language gesture dataset to improve location diversity with single user environment. In this chapter, we focus on the contributions and challenges of a significant use case that aims to enhance the quality of life for deaf people using a publicly available dataset of Wi-Fi CSI sign language gestures.
- Chapter 5: Wi-Nod System for Head Nodding Recognition: Chapter 4 delves into the Wi-Nod system, which focuses on recognizing head nodding gestures using Wi-Fi CSI. The system is designed to provide communicative support for quadriplegics. This chapter explores techniques and methodologies to accommodate the varying characteristics and behaviors of different users within the system.
- Chapter 6: HeMoFi4Q: Morse Communication Based on Wi-Fi and Head Motion for Quadriplegia with Environmental Robustness: a new sign language based on the previous Wi-Nod system is introduced, enabling communication with quadriplegia patients. The chapter specifically investigates the recognition of all 26 letters of the alphabet using Morse code and head movements within diverse spatial contexts. Additionally, this chapter aims to mitigate the scarcity of labeled data and the distribution shift problem, enhancing the generalizability of the Wi-Fi CSI system.
- Chapter 7: Tracking On-Desk Gestures Based on Wi-Fi CSI on Low-Cost Microcontroller: addresses the contactless tracking of on-desk gestures for recognizing worker's moods. The research focuses on the implementation of the ESP32 microcontroller in a MIMO configuration and explores how Wi-Fi CSI can be utilized to track gestures performed on a desk surface and infer the mood of the worker.
- Chapter 8: Conclusion and Future Directions: concludes the contributions of the thesis and outlines the future directions for practical Wi-Fi CSI-based sensing systems. Specifically, it focuses on the challenges that need to be addressed in real-world manufacturing environments. The chapter provides a summary of the key findings and suggests potential areas for future research.

2 Fundamental Concepts

2.1 Radio Frequency based Sensing Techniques

In recent years, radio signal-based approaches have garnered significant interest from researchers across various disciplines, particularly in the field of occupancy detection. These approaches leverage the characteristics of RF signals to enable a range of applications, including but not limited to Gesture Recognition (GR), HAR, and human counting. Several radio-based technologies have been explored for these purposes, such as Bluetooth Low Energy (BLE) transmissions, frequency shift radars, and Wi-Fi signals from access points.

Radio signals offer unique advantages for sensing and detection due to their ability to propagate through the environment and interact with objects and individuals in their path. By analyzing the properties of received RF signals, valuable information can be extracted to infer occupancy, monitor human activities, detect gestures or count the number of individuals in a given space.

2.1.1 Bluetooth Low Energy based systems

Bluetooth Low Energy (BLE) [19, 20] transmissions have gained popularity as a radio-based technology for human behavior recognition and healthcare applications. BLE-enabled devices, such as smartphones or wearable devices, emit periodic signals that can be detected and utilized to determine the presence of individuals within a specific area. These signals exhibit distinct characteristics that enable accurate occupancy detection. However, BLE has

a relatively limited range compared to Wi-Fi signals. While BLE can typically reach up to 100 meters in an open space, its range can be significantly reduced in environments with obstacles or interference. This range limitation can restrict the coverage area for BLE-based sensing applications.

2.1.2 Radio Frequency Identification based systems

Radio Frequency Identification (RFID) [21] is an active radio frequency sensing technique that is widely employed for object identification and tracking purposes. The technology involves the use of tags, also known as transponders, and readers. These tags consist of a microchip and an antenna, enabling them to both receive and transmit radio frequency signals. On the other hand, readers emit radio frequency signals and establish communication with the tags. When a tag enters the range of a reader, it captures the emitted signal and utilizes it to power itself. Subsequently, the tag responds by transmitting its unique identification information back to the reader. This process allows for the identification and tracking of objects equipped with RFID tags. Wang et al. [22] introduced RF-IDraw system, which employs off-the-shelf RFID readers and tags for capturing in-air hand gestures to interact with a device. The RF-IDraw system utilizes the concept of a virtual touch screen, where hand gestures are interpreted as if they were touching a physical surface. The system achieved a character recognition rate of 97.3% and a word recognition accuracy of 88%. However, RFID systems typically exhibit relatively shorter range capabilities, typically spanning up to a few meters. The achievable range is contingent upon the specific tags and readers employed, as the performance characteristics can vary. Factors such as the power output of the reader and the sensitivity of the tag's receiver impact the effective range of an RFID system.

2.1.3 Radar based systems

In recent years, there has been a notable expansion in the application of radar for recognition tasks. Radar-based HAR approaches [23] offer advantages over vision-based methods due to radar's insensitivity to lighting conditions

and weather disturbances. As individuals move, the speeds and Doppler frequencies of different body parts vary over time, providing distinctive signatures that can be captured by radar. Radar-based sensing system operates on the principles of electromagnetic reflection and transmission. These systems employ radio frequency signals transmitted and received by transceiver antennas to detect and analyze the interactions between the signals and objects present in the surrounding environment.

mmWave Radar

Recently, millimeter Wave (mmWave)-based radar systems have found applications in health monitoring and HAR applications due to their large frequency range, from 30 GHZ to 300 GHZ, and small size. By leveraging the unique characteristics of mmWave signals, such as their ability to penetrate certain materials and provide detailed range and velocity information, these systems enable non-intrusive monitoring of vital signs and movement patterns. Among the mmWave radar systems, Frequency-Modulated Continuous-Wave Radar (FMCW) radar [24] is the most commonly employed method in the existing literature. FMCW radar utilizes variations in the time taken for the carrier frequency to shift, known as Time of Flight (ToF), to measure distances. It accomplishes this by mapping the differences in time to the shifts of the carrier frequency. This approach facilitates the measurement of ToF, which represents the time it takes for a wireless signal to travel from the transmitter to the human body being reflected and back to the receiver.

Sun et al. [25] employed mmWave radar technology in their research on fall detection. They leveraged the capabilities of mmWave radar, combined with the Long Short-Term Memory (LSTM) model, to preserve the temporal features necessary for accurate fall detection. In their study, Sun et al. recognized the significance of temporal information in detecting falls, as it provides crucial context and helps distinguish falls from other activities or movements. To address this, they integrated the LSTM model with mmWave radar data, allowing for the capture of sequential and time-dependent patterns associated with falls. mBeats [26] is a robot equipped with a mmWave radar system designed for obtaining periodic heart rate measurements across various user poses. The authors aimed to address the challenge of obtaining

accurate and continuous heart rate measurements, particularly when users are in different poses or positions.

Ultra Wide Band based systems

Ultra Wide Band (UWB) technology in radio frequency sensing refers to a form of communication that utilizes a significantly larger effective bandwidth, surpassing 500 MHz. This wide bandwidth allows for the transmission of substantial amounts of data over short distances. UWB radar, in particular, leverages a series of quick, short pulses that occupy the entire available bandwidth. One advantage of UWB radar is its insensitivity to the multi-path effect, which is caused by signal reflections and interference from different paths. The high-time resolution of UWB radar mitigates the impact of multi-path effects, resulting in more accurate and reliable sensing. This characteristic makes UWB radar a robust choice for various applications. In [27], the authors utilized UWB radar signals from a sensor in combination with Wi-Fi Channel State Information (CSI) measurements obtained through Universal Software Radio Peripheral (USRP) devices to develop a system for lip reading under mask-wearing scenarios. By exploiting the radar signals and integrating them with CSI measurements and leveraging Deep Learning (DL) models, the system achieved lip reading recognition accuracy of more than 80%.

In spite of the favorable aspects of radar technology, its broader implementation faces challenges due to the comparatively higher cost associated with it. This cost barrier arises from the requirement for customized hardware, which tends to be expensive. Consequently, when assessing the feasibility and practicality of deploying radar technology, particularly in scenarios involving large-scale applications or cost-sensitive environments, the cost factor must be thoroughly taken into account.

2.1.4 Wi-Fi based sensing systems

Generally speaking, the techniques discussed above entail additional burdens in terms of complex hardware installation and diverse maintenance require-

ments. These limitations necessitate the development of a cost-effective and non-intrusive solution capable of capturing human body movements during daily activities.

In recent years, there has been a surge of interest in leveraging Wi-Fi-based techniques for human activity sensing. This growing trend can be attributed to the widespread adoption of Wi-Fi technology in various home and office environments. With the proliferation of smart devices such as smartphones, smart TVs, smart thermostats, and home security systems, wireless interconnectivity through Wi-Fi has become pervasive.

Wi-Fi signals are emitted by Commercial-Of-The-Shelf (COTS) access points and typically possess a substantial coverage range, extending up to tens of meters within indoor environments. These broadcast signals interact with the surrounding environment, including objects and human bodies. By monitoring the changes in the received Wi-Fi signals, it becomes possible to capture and interpret human body movements.

Two signal descriptors associated with Wi-Fi signals have emerged as key metrics for quantifying the variations in the received signal. These descriptors are known as Received Signal Strenght Indicator (RSSI) and CSI. In the subsequent sections, we will delve into a comprehensive explanation of these descriptors, elucidating their characteristics and significance in the context of Wi-Fi signal analysis.

RSSI based methods

RSSI serves as a commonly available metric in a wide range of COTS Wi-Fi devices. When a target object exists within the transmission range of a Wi-Fi transmitter and receiver, it introduces fluctuations in the received signal power due to reflections. These variations in signal power are captured and quantified through RSSI values. RSSI essentially quantifies the path loss within the received Wi-Fi signals, relative to a specific distance. The Log Distance Path Loss Model (LDPL) model [28], as depicted by Equation (2.1), can be employed to estimate this relationship.

$$P(d) = P(d_0) + 10\gamma \log(\frac{d}{d_0}) + X_\delta \tag{2.1}$$

where $P(d)$ represents the RSSI measurement, which serves as an indicator of the path loss at a given distance d and is expressed in Decibel (dB), $P(d_0)$ corresponds to the path loss at the reference distance d_0, γ represents the path loss exponent, which characterizes the attenuation of the signal with increasing distance. Additionally, X_δ denotes a zero-mean normal noise component caused by flat fading, which introduces random variations in the received signal strength. RSSI, as a coarse-grained metric, provides a single path loss value per packet, making it suitable for various applications such as indoor localization [29] and crowd estimation [30]. Abdelnasser et al. proposed a WiGest [31], which leverages the fluctuations in signal strength values, specifically RSSI values, induced by hand movements. The primary objective of the WiGest system is to track the gestures made by a user's hand in the vicinity of their mobile device, eliminating the need for physically holding it, and translating these variations into specific actions. The WiGest system achieves a recognition accuracy of 87.5% when utilizing a single access point, and this accuracy improves to 96% when employing three access points within the sensing environment.

However, the efficacy of RSSI-based systems diminishes notably in intricate sensing areas, thus rendering it an unreliable metric.

CSI based methods

The utilization of RSSI-based approaches for reliable human activity sensing is often limited due to their low granularity. In contrast, CSI serves as a fine-grained metric capable of capturing more intricate details within the sensing area. Fig. 2.1 provides an overview of the multipath effect arising from both static and dynamic objects. The figure demonstrates that static objects such as walls, floors, and furniture give rise to reflections, while dynamic objects, such as human movements, cause scattering. These multipath phenomena collectively contribute to the measurement of CSI waveforms. In contrast to the RSSI metric, CSI comprises a complex set of values encompassing both amplitude and phase information for multiple Orthogonal Frequency Division Modulation (OFDM) subcarriers. The wireless channel exhibits unique multipath fading effects on each subcarrier, which is determined by its slightly varying center frequency. When considered collectively, these

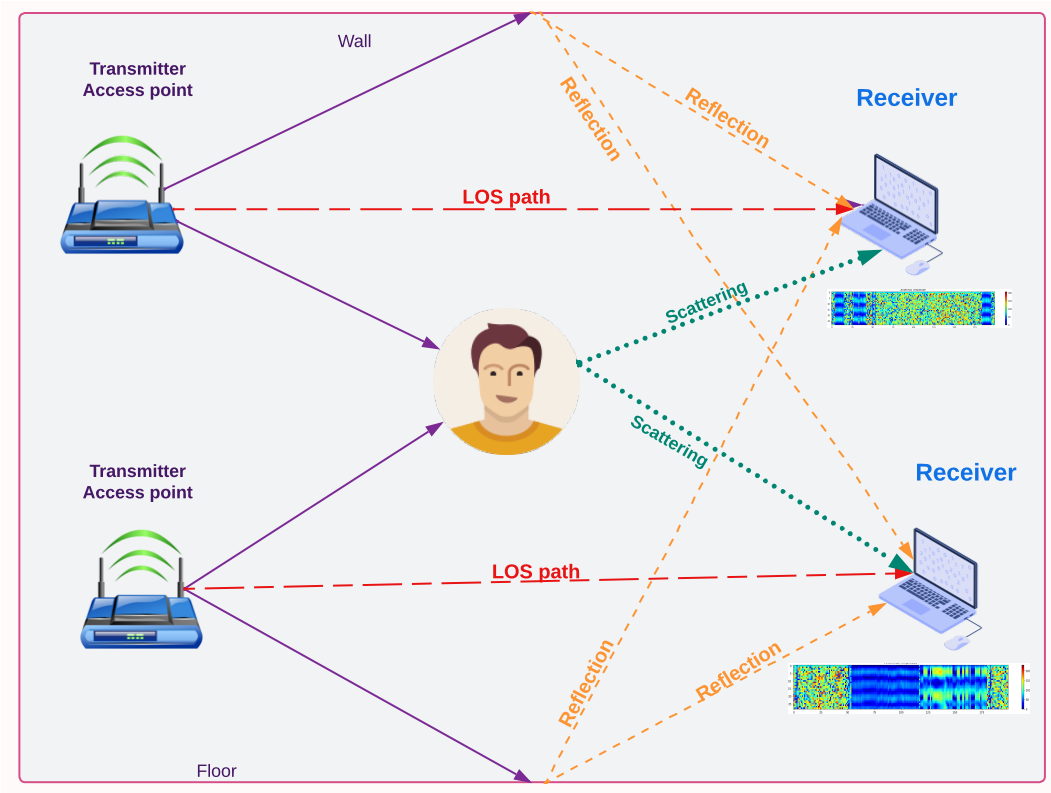


Figure 2.1: Visualization of the multipath propagation effect in RF signals

subcarriers offer a comprehensive representation of the characteristics of the wireless channel.

2.2 Fundamental Concepts of CSI

In recent years, there has been a notable increase in the utilization CSI measurements in various applications related to HAR. The existing literature has demonstrated successful implementations of CSI measurements in diverse HAR domains, including device-free indoor localization [32], smoking detection [33], action recognition [34], gesture recognition [35, 36], and crowd counting [37]. In this section, we provide an introduction to the background knowledge of CSI, exploring its fundamental concepts and principles.

2.2.1 CSI Base Signals

As previously mentioned, CSI serves as a metric that characterizes the properties of the wireless communication channel. It encompasses the variations in signal reflection and scattering encountered along the transmission path between the Transmitter (Tx)- Receiver (Rx). CSI is derived from Channel Frequency Response (CFR) and extracted from each subcarrier within the OFDM system which mathematically expressed in the frequency domain as shown in Equation (2.2).

$$Y(f,t) = H(f,t) \times X(f,t) + N \tag{2.2}$$

where $H(f,t)$ denotes the complex matrix of CSI for different subcarriers with frequency f at time t. $Y(f,t)$ and $X(f,t)$ represent the received and transmitted signals, respectively. N is a noise vector.

Typically, the H matrix records the the amplitude attenuation, $||H_{\mathrm{i}}(f)||$, and phase response, $\angle \hat{H_{\mathrm{i}}}(f)$, of each i_{th} subcarrier frequency at time t.

$$\hat{H_i}(f) = \|\hat{H_i}(f)\| * e^{j\angle \hat{H_i}(f)} \tag{2.3}$$

The amplitude of CSI provides information about the signal's power decay as it traverses through the environment, while the phase reflects the changes in the signal's phase due to different path lengths and reflections. By analyzing the variations in the amplitude, phase, or both, it becomes possible to detect and track human movements in a passive manner. When a human moves within the sensing area, the reflected signals from their body create variations in the received CSI amplitude and phase. These variations can be attributed to the changes in the path length and the scattering effects caused by the human's presence. By monitoring and analyzing these variations over time, it becomes feasible to discern different types of human movements, such as walking, running, or gestures, without requiring the individual to carry any specific devices or wear any sensors. Through a thorough analysis of the variations in CSI amplitude, phase, or a combination of both, referred to as CSI base signals, it becomes feasible to extract resilient and discriminative patterns that are uniquely associated with different forms of movement. These patterns encapsulate valuable information and can be effectively utilized as

features within diverse learning tasks, enabling the development of accurate models for movement recognition and other related applications.

2.2.2 CSI Noises

In the realm of passive sensing systems, CSI emerges as a more robust Wi-Fi metric when compared to RSSI. Nevertheless, the efficacy of CSI is not immune to the influence of various noise sources, including environmental factors, hardware synchronization issues, and interference signals. To better understand the impact of noise on CSI measurements, it is essential to categorize these perturbations into two distinct types: CSI amplitude noise and CSI phase noise.

CSI amplitude noise

The outliers in the raw CSI amplitude primarily arise due to the presence of environmental noises, as well as fluctuations resulting from transmission power changes and transmission rate adaptation. The environmental noises encompass a wide range of factors, including multipath propagation, interference from neighboring devices, and electromagnetic interference from various sources. These external influences contribute to the inherent instability in the amplitude of CSI values. Additionally, changes in transmission power and transmission rate adaption, which occur dynamically in wireless communication systems, further contribute to the observed large variations in CSI amplitude. One widely employed approach to mitigate this issue is the utilization of a low-pass filter. This method aims to selectively remove high-frequency noise originating from the environment while preserving the low-frequency components relevant to gestures or movements of interest. By employing a low-pass filter, the undesired high-frequency variations can be attenuated, enabling the extraction and analysis of the desired low-frequency components associated with the specific activities being monitored.

CSI phase noise

It is well known that the measured phase of CSI exhibits a significant amount of randomness due to Carrier Frequency Offset (CFO), Sampling Frequency Offset (SFO), and Symbol Timing Offset (STO). These offsets introduce phase shifts in the raw CSI measurements, leading to the observed randomness in the phase values.

1. Carrier Frequency Offset (CFO): CFO occurs due to slight frequency variations among the subcarriers in the transmitter-receiver link. The lack of synchronization between the clocks of the sender and receiver results in a uniform phase offset being introduced across all subcarriers.

2. Sampling Frequency Offset (SFO) and Sampling Time Offset (STO): SFO and STO originating from the Analog-to-Digital Converter (ADC) in the receiver, introduces distinct phase offsets to individual subcarriers within the OFDM system. This dependence on the subcarrier index results in varying phase shifts across different subcarriers. It is important to note that these offsets manifest in both the frequency and time domains, and that the SFO and STO exhibit a linear relationship with each subcarrier.

The measured CSI phase can be expressed as in Equation 2.4 [38]

$$\angle \hat{H_i}(f) = \angle H_i(f) + 2\pi \frac{m_i}{N} \times \delta_t + \gamma + Z \tag{2.4}$$

where $\angle \hat{H_i}(f)$ and $\angle H_i(f)$ represent the raw and actual CSI phase, respectively. δ_t denotes the time lag due to SFO and STO and $m_i$is the subcarrier index of the ith subcarrier. The value of$N$ is set to 64, representing the discrete Fourier Transform. γ is the unknown phase offset resulting from CFO. Z denotes the presence of measurement noise.

3 General Approach

3.1 System Overview

The general framework for HAR utilizing Wi-Fi CSI encompasses four fundamental modules: CSI data collection, base signal extraction, signal preprocessing, and a learning model as shown in Fig. 3.1. In this chapter, we will delve into each stage, providing a comprehensive overview. Additionally, we will discuss various gesture recognition systems and the challenges they present, which we aim to address in this thesis.

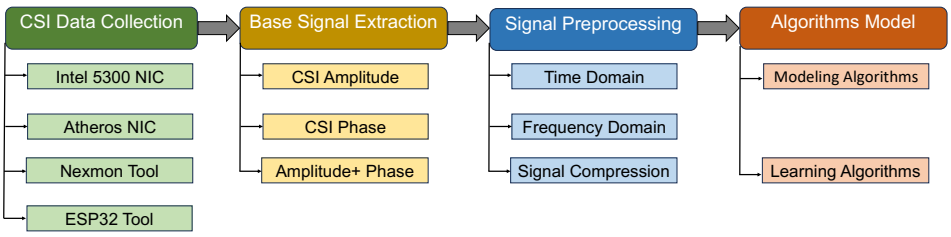


Figure 3.1: General approach overview.

3.2 CSI Data Collection

Wi-Fi CSI serves as a valuable source of data that captures information about the impact of surrounding objects and motions on the multi-path propagation of wireless signals. In the process of collecting CSI waveforms, the Long Training Symbol (LTS) serves as a crucial component of the communication process. The transmitter transmits LTS, which contains predefined information for each subcarrier, enabling the receiver to estimate CSI by comparing the original LTS with the received LTS [39]. Within OFDM technology, subcarriers play a crucial role in transmitting data over the wireless channel.

There are three types of subcarriers: null subcarriers, pilot subcarriers, and data subcarriers. Null subcarriers, also known as zero subcarriers, are unused subcarriers intentionally inserted to act as a guard against interference from adjacent channels. Their primary purpose is to mitigate the impact of interference and ensure reliable communication within the allocated frequency band. Pilot subcarriers, on the other hand, are not utilized for conveying modulated data. Instead, they serve as reference signals for channel measurements and synchronization between the transmitter and receiver. These pilot subcarriers employ a predetermined data sequence and introduce additional overhead to the channel due to their dedicated purpose. The specific Physical Layer (PHY) standard and the allocated bandwidth determine the total number of subcarriers that can be utilized for transmission. The remaining subcarriers, apart from the null and pilot subcarriers, are referred to as data subcarriers. In the context of 802.11ac, these data subcarriers exploit the same modulation format as specified by the standard. They are responsible for carrying the encoded information and transmitting it over the wireless channel, enabling the effective transmission of data in an OFDM-based system.

Various tools are employed to record CSI waveforms in the sensing area, each offering distinct advantages and limitations. In this section, we provide a comprehensive overview of the most commonly used tools for capturing CSI waveforms, shedding light on their respective strengths and weaknesses.

3.2.1 Intel 5300 NIC

While Wi-Fi technology has incorporated CSI since the IEEE 802.11n standard, it is important to note that not all commercially available Wi-Fi cards provide access to CSI data. Among the tools commonly employed for CSI measurements, the 802.11n CSI Tool [40] has emerged as the most widely utilized [35]. This tool utilizes Intel 5300 Wi-Fi cards to report compressed CSIs from 802.11n-compatible Wi-Fi networks. It offers C scripts and MATLAB source code that facilitate CSI measurements and subsequent processing. In terms of the specific capabilities of the Intel WiFi Link 5300 Network Interface Card (NIC), it exports CSI information for only 30 out of the total 56 subcarriers for each antenna, assuming a 20MHz bandwidth. The applicability of the 802.11n CSI Tool, which is widely employed for Wi-Fi CSI based

approaches, is limited to older Intel 5300 NIC. However, the availability and acquisition of these NICs can be challenging due to their outdated nature.

3.2.2 Atheros NIC

The Atheros CSI Tool [18] is an open source CSI tool. One of its key features is the ability to extract comprehensive PHY wireless communication information from Atheros Wi-Fi NIC. It offers the capability to capture uncompressed CSI utilizing Qualcomm Atheros WiFi cards. When considering a 20MHz, 40MHz WiFi channel, this tool provides access to 52 and 114 CSI subcarriers, respectively.

Despite the presence of a substantial community focused on Atheros NIC in the context of Wi-Fi CSI sensing [41] due to its high CSI resolution, certain issues have been encountered that affect the accuracy and reliability of CSI measurements using these NIC. One such issue arises when employing a configuration with one transmitter and two receiver antennas. In this scenario, the system mistakenly reports the existence of two transmitter antennas, leading to the generation of noise on an antenna that, in reality, does not exist. Consequently, the CSI measurements obtained for this non-existent antenna are misleading and can adversely impact subsequent analysis and interpretation. Another issue manifests when utilizing a setup comprising three transmitter antennas and two receiver antennas. In this case, the system fails to accurately estimate the CSI data, resulting in a consistent sinusoidal shape in the obtained measurements. This deviation from the expected CSI behavior poses challenges in accurately capturing and representing the true characteristics of the wireless channel, hindering the reliability of subsequent analyses and conclusions.

3.2.3 Nexmon Tool

The Nexmon CSI Tool [42] has revolutionized the extraction of CSI from various devices such as Raspberry Pi 3B+ and 4B, Google Nexus 5, and select routers. One notable advantage of the Nexmon tool is its ability to support multiple transmit-receive antenna configurations, enabling up to 4x4 MIMO.

An important feature of the Nexmon CSI Tool is its customizable CSI collection filters, which allow researchers to extract relevant CSI specifically from chosen transmitters. Unlike other tools, it does not necessitate the suppression of complete CSI data, offering greater flexibility and precision in the extraction process. Furthermore, Nexmon provides a configuration option to assign a distinct interface solely for monitoring frames on the Raspberry Pi, once configured on the host. This facilitates targeted monitoring and analysis of specific frames, enhancing the precision and efficiency of the research process [43]. In terms of technical capabilities, the Nexmon CSI Tool supports bandwidth utilization of up to 80 MHz, enabling the extraction of CSI from 242 subcarriers out of a total of 256. This heightened CSI resolution per packet contributes to enhanced fidelity and granularity in capturing wireless channel characteristics.

Despite its continuous refinement, it is noteworthy that the adoption of the Nexmon CSI Tool remains limited, primarily due to the restricted range of supported devices and the scarcity of available information. These factors pose challenges to its widespread usage within the research community.

3.2.4 ESP32 Tool

The current state-of-the-art Wi-Fi Channel State Information (CSI)-based systems employed in HAR, head detection, and gesture recognition rely on open-source CSI tools developed by Halperin et al.[17], Atheros[18], and the Nexmon CSI Tool [42]. Despite their widespread usage, these tools are associated with certain limitations, including restricted device support and hardware compatibility, intricate setup procedures, limited data processing capabilities, and a lack of official support and updates. Addressing the challenges of Wi-Fi CSI sensing systems necessitates the development of a small, low-power, memory-efficient, cost-effective, and compatible CSI tool, capable of handling noisy CSI values in multi-human contexts and extracting informative base signal variations related to head motions.

In this regard, the ESP32 unit [44] emerges as a microcontroller with a single system-on-a-chip architecture, integrating multiple components, such as the processing unit, memory, and communication interfaces, onto a single

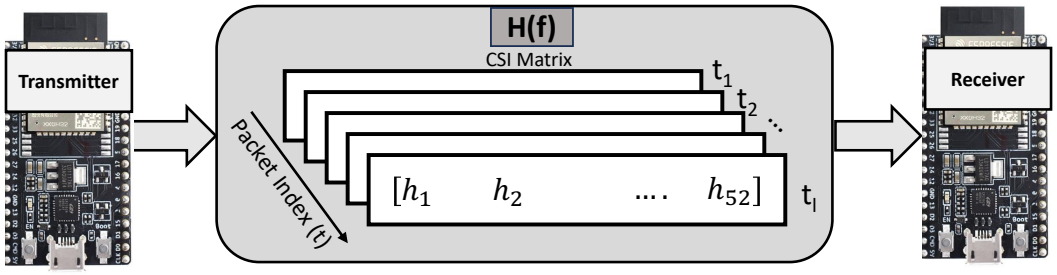


Figure 3.2: CSI vector for each t packet in ESP32 system

chip. Notably, it exhibits characteristics of low power consumption and cost-effectiveness, making it a viable option for the proposed research.

Utilizing OFDM for wireless signal transmission, the ESP32 module employs 64 narrowband subcarriers, with 12 of them designated as null subcarriers and the remaining 52 serving as data subcarriers.

The CSI data collected by the ESP32 node is represented as a complex matrix with dimensions of $l \times m$, where l represents the number of packets and m represents the number of subcarriers as illustrated in Figure 3.2.

Our proposed systems in this thesis rely on the ESP32 toolkit, which serves as the main tool for CSI recording. Notably, the ESP32 toolkit possesses advantageous qualities, including low power consumption, user-friendly operation, and portability, making it a standalone CSI solution. Furthermore, the ESP32 toolkit eliminates the need for specific hardware compatibility or device support as it functions as both a transmitter and receiver. It supports the IEEE802.11n protocol and achieves a data rate of up to 150 Mbps at 2.4 GHz.

Table 3.1 summarizes the main differences between the existing CSI tools.

Table 3.1: Comparison between different CSI tools

CSI Tool	Devices/ NIC	Data subcarriers	MIMO	Device Support
802.11n CSI Tool [40]	Intel 5300 Wi-Fi cards	30	✓	×
Atheros CSI Tool [18]	Qualcomm Atheros WiFi cards	52 or 114	✓	×
Nexmon CSI Tool [42]	Raspberry Pi 3B+ and 4B, Google Nexus 5	242	✓	×
ESP32 Tool [44]	ESP32 microcontroller	52	×	✓

3.3 Base Signal Extraction

The extraction and analysis of variations in CSI amplitude, phase, or a combination of both are crucial components of sensing systems based on Wi-Fi CSI approaches. These variations provide valuable insights into the dynamics of wireless signals and enable the system to accurately detect and recognize various activities.

3.3.1 Amplitude

The amplitude of CSI in RF sensing methods represents the quantification of signal power attenuation caused by multi-path fading. In a sensing area, any motion within the human affects the propagation of wireless signals, leading to variations in the received signal's amplitude. This unique relationship between amplitude change and motion characteristics enables the detection and quantification of movements using amplitude measurements. Numerous studies [35, 37] have demonstrated the effectiveness of amplitude for activity recognition, highlighting its sensitivity across a wide range of movements.

3.3.2 Phase

The phase measurements capture the relative distance and direction of signal propagation. By examining changes in phase, the study of [45] can depict signal variations and infer corresponding motion patterns. It is important to note that phase is periodic compared to amplitude and is susceptible to influences from device clock and carrier frequency. Therefore, the calibration of phase measurements is necessary to mitigate noise and ensure accurate extraction of motion and distance information.

3.3.3 Amplitude and Phase

Leveraging both amplitude and phase measurements can enhance the sensitivity and accuracy of activity recognition [36]. While amplitude and

phase individually exhibit sensitivity to small movements in the physical environment, they do not directly provide spatial information, such as the spatial location of multiple movements or individuals in a three-dimensional physical space. Therefore, the combination of amplitude and phase measurements can improve the overall performance of activity recognition systems by incorporating spatial information into the analysis.

3.4 Signal Preprocessing

In general, raw CSI waveforms is subject to various influencing factors. One significant factor is the presence of a multi-path channel, commonly encountered in indoor environments, where signal reflections from multiple reflectors generate differently delayed versions of the signal that reach the receiver alongside the direct-path signal. This multi-path phenomenon introduces phase shifts and can result in constructive or destructive interference among the signal components, thereby affecting the accuracy of CSI estimation.

In addition to the multi-path channel, the accuracy of CSI estimation is influenced by various other factors, including processing operations performed by the receiver and transmitter, as well as hardware and software errors as mentioned in Chapter 2.2.2. These factors collectively contribute to the overall uncertainty and variability observed in the recorded CSI values.

Because of the above limitations, raw CSI values cannot be directly utilized as inputs to classification or detection models. They can significantly impair the performance of recognition systems. Therefore, appropriate techniques for noise mitigation and preprocessing of CSI data are necessary to improve the effectiveness of recognition systems in practical scenarios.

3.4.1 Time Domain

The most common filters applied to remove the amplitude noises and outliers in the time domain are sliding windows and Hampel filter.

Sliding window filter

Environmental noises introduce high-frequency disturbances, resulting in outliers within the CSI amplitude data. To address this issue and achieve a smoother representation of CSI amplitude, sliding window filters are commonly employed.

Sliding window filters operate by applying a window of a specific size to the raw amplitude data and applying a designated function within this window region. Two widely used filters for noise removal are the Moving Average and Median Filters.

The Moving Average Filter [46] replaces each data point with the average value of its neighboring data points within the sliding window. This filter effectively reduces high-frequency noises by smoothing out the variations caused by outliers. The choice of window size and the use of multiplying factors can further adjust the weights assigned to the neighboring data points. For instance, exponentially Weighted Moving Average (eWMA) assigns higher weights to recent data points, resulting in a greater emphasis on recent trends.

On the other hand, the Median Filter [47] replaces each data point with the median value of its neighboring data points within the sliding window. This filter is particularly effective in removing outliers, as the median value is less sensitive to extreme values compared to the mean value used in the Moving Average Filter.

Hampel filter

The Hampel filter is employed in the many CSI-based approaches [48, 49] for identifying outliers within the base signals.

For each value x in the base signals, a window is constructed, consisting of x and a specified number $k/2$ of neighboring points on each side. The median of this window is then computed. Subsequently, the standard deviation of x with respect to its window median is calculated using the Median Absolute Deviation (MAD) method.

To determine whether a value should be considered an outlier, a predefined threshold is established. If the difference between x and the median exceeds a specified number of MAD, the value is identified as an outlier. In such cases, the value is replaced by the median.

In essence, the Hampel Identifier declares discrete values as outliers if they lie outside the interval [μ - $\gamma\sigma$, μ+ $\gamma\sigma$], where μ represents the median, σ denotes the MAD, and γ is a parameter dependent on the application. The default value for γ is three, although it can be customized based on specific requirements.

3.4.2 Frequency Domain

The use of these signal transform techniques makes the time-frequency domain particularly important in the context of micro-movement recognition. This approach investigates the variations and dynamics of CSI across different frequencies, providing valuable insights into the underlying patterns and changes that occur over time.

Fast Fourier Transform (FFT)

The FFT technique is extensively employed to identify prominent frequencies within a given signal. By utilizing FFT, distinct dominant frequencies present in the signal are effectively extracted. Additionally, FFT can be combined with a Lowpass Filter (LPF) to mitigate high-frequency noise components resulting from the sensing environment, enhancing the quality of the signal.

In certain applications, such as human motion detection and breathing estimation, specific target signals within desired frequency ranges are of interest. This is achieved by employing Bandpass Filters (BPFs) alongside FFT. Applying BPFs isolates signals within particular frequency bands that are associated with different human activities.

FFT can be calculated as in Equation 3.1

$$x[k] = \sum_{n=1}^{N} x[n] \times e^{\frac{-j2\pi kn}{N}} \tag{3.1}$$

where k and N represent the frequency index and signal size, respectively.

Short Time Fourier Transform (STFT)

STFT provides a time-frequency representation of a signal, allowing for the analysis of frequency components at different time intervals. This is particularly useful in applications where the signal's frequency content changes over time, such as in dynamic environments or when analyzing non-stationary signals. Moreover, STFT excels in providing localized frequency information by employing a windowing function on segmented portions of the signal. Through this approach, changes in frequency content within specific time intervals can be accurately identified, offering a more comprehensive understanding of the signal's behavior and dynamics. Furthermore, STFT exhibits a high frequency resolution by utilizing shorter window sizes. This feature proves beneficial in detecting closely spaced frequency components within the signal. By leveraging a shorter window, STFT enables the identification and distinction of these fine-grained frequency components, enhancing the precision of the analysis. STFT can be mathematically expressed as in Equation 3.2

$$x(t,k) = \sum_{n=-\infty}^{\infty} x[n]w[n-t]e^{-jkn} \tag{3.2}$$

where t and k represent the time and frequency index, respectively. w is a window function. There are various window functions, such as the rectangular, Hamming, Hanning, and Blackman windows, which offer different characteristics and trade-offs, affecting parameters such as main lobe width, side lobe suppression, and spectral resolution. The choice of window function should align with the specific requirements and considerations of the analysis at hand.

Discrete Wavelet Transform (DWT)

STFT encounters a trade-off between time and frequency resolution. While STFT excels in identifying frequency components, it lacks the ability to precisely locate the timing of frequency changes. However, an alternative approach, namely the Wavelet Transform, offers advantages in terms of both

frequency and time resolution. With Wavelet Transform, low-frequency signals can be analyzed with high frequency resolution, while high-frequency signals benefit from good time resolution. This characteristic makes Wavelet Transform suitable for analyzing signals with varying frequency content over time. Furthermore, the output of the Discrete Wavelet Transform (DWT) can be directed to a wavelet filter, enabling noise removal and enhancing the quality of the signal. This filtering process aids in preserving the relevant information while reducing unwanted noise components. In addition, DWT exhibits robustness in different scenarios and surpasses the Doppler phase shift method in terms of mobility information preservation. This quality makes DWT a valuable tool [50], particularly when dealing with dynamic environments or situations where movement is involved. The DWT transforms the time-domain signal into the wavelet domain, decomposing it into wavelet detail and approximate coefficients. These coefficients are calculated through down-sampling convolutional equations, as expressed in Eq. 3.3 [51]:

$$\begin{aligned} Y_{m,low}[n] &= \Downarrow Q\left[\sum_{j=-\infty}^{\infty} X[j] \times g[n-j]\right] \\ Y_{m,high}[n] &= \Downarrow Q\left[\sum_{j=-\infty}^{\infty} X[j] \times h[n-j]\right] \end{aligned} \tag{3.3}$$

where $Y_{m,low}[n]$ and $Y_{m,high}[n]$ represent the approximation and detail coefficients, respectively, j denotes the frequency index, $X[j]$ is the input signal, $\Downarrow Q[\cdot]$ is a downsampling filter, $g[n]$ and $h[n]$ are a low-pass and high-pass filter, respectively [51].

3.4.3 Signal Compression

Signal compression plays a crucial role in reducing the dimensionality of the data and removing redundant and irrelevant information present in the raw CSI measurements across various domains. For CSI values, there are similarities between the adjacent subcarriers. Consequently, the primary objective of the compression stage is to extract relevant information related to movement while minimizing redundancy. By achieving this, the time

consumption and model complexity during the learning phase are reduced, thereby enhancing the efficiency of the subsequent classification task.

Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and Independent Component Analysis (ICA) are the common signal compression techniques. PCA depends on an orthogonal linear transformation approach, where the covariance matrix is computed to generate principal components. The principal components are then ranked based on their corresponding eigenvalues, with the top components containing the most significant information. By preserving these top components and discarding the rest, data compression can be achieved. However, one notable drawback of PCA is its reliance on linear projections, which may not align with the nature of CSI signals. As a result, significant information loss can occur, rendering the compressed data less representative of the original signal. SVD calculates the singular values to capture uncorrelated features with maximum variance from the original data. SVD components are orthogonal linear the same as PCA. While ICA changes the data from high dimensional to low space by finding the statistically independent features from the original data which are not ranked or linear.

3.5 Algorithms Model

The main objective of this stage is to find the function that maps the CSI signals to perform the classification task. In this section, we will provide a brief overview of the commonly used algorithms for Wi-Fi CSI sensing approaches to estimate the mapping function.

3.5.1 Modeling Algorithms

In modeling algorithms, the system models the preprocessed CSI signals by theoretical models based on physical theories. These models encompass concepts such as the Fresnel Zone Model, Angle of Arrival (AoA), Angle of Departure (AoD), ToF, Amplitude Attenuation, Phase Shift, and Doppler Spread. As discussed previously in Chapter 2.2, CSI serves as a highly sensitive metric

that characterizes the communication channel through the analysis of parameters such as amplitude attenuation and phase shift. These parameters are influenced by various factors, including the distance between the transmitter and receiver, as well as multipath effects including radio reflection, refraction, absorption, and scattering. The wireless signal amplitude attenuation by the LoS path can be expressed as follows:

$$\frac{p_r}{p_t} = D_t D_r (\frac{\lambda}{4\pi d})^2, d >> \lambda \tag{3.4}$$

where p_r and p_t are the received and transmitted signal amplitude. D_t and D_r represent the directivity of the transmitter and receiver antennas, while λ denotes the carrier wavelength. The distance between the transmitter and receiver is represented by d. In practical environments, the presence of static and dynamic objects within the sensing area introduces phase shifts, which arise from the time delays associated with each propagation path. Furthermore, the phase shift is influenced by Doppler effects when the transmitter is in motion towards the receiver. These factors contribute to increased randomness in the actual phase of the CSI.

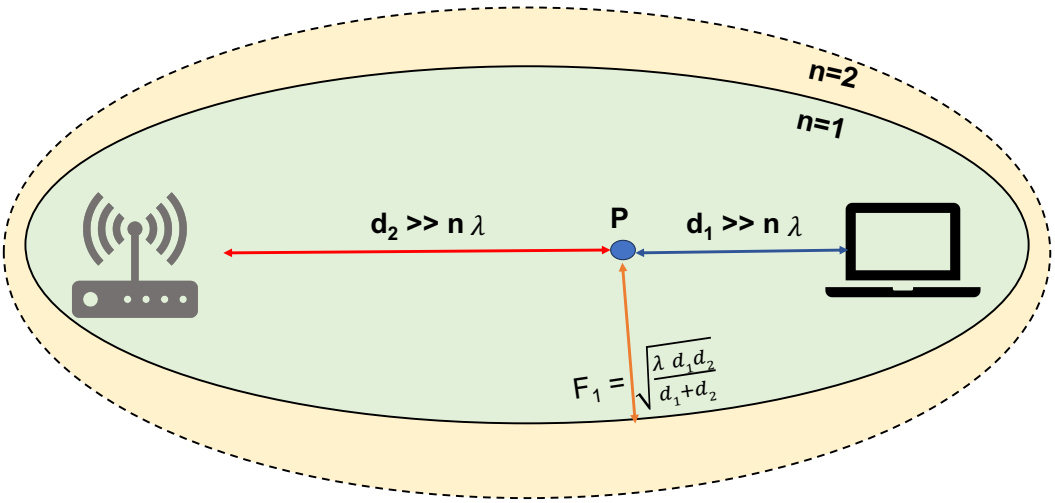


Figure 3.3: The illustration of Fresnel Zone Model [39].

The amplitudes and phases of CSI are influenced by the propagation of radio waves along multiple paths. The Fresnel Zone Model offers a framework for dividing the space surrounding the transmitter and receiver into concentric prolate ellipsoidal regions known as Fresnel Zones. The radius of the nth Fresnel Zone can be computed as illustrated in Figure 3.3. This model

elucidates how radio signals propagate and interact with objects within the Fresnel Zone regions, resulting in deflected signals that traverse multiple paths before reaching the receiver. The amplitude attenuation and phase shift incurred by these deflected signals, contingent upon the path length, contribute to constructive or destructive effects at the receiver [39].

Regarding CSI-based tracking and localization applications, AoA and ToF are two prominent models. These models characterize the amplitude attenuation and phase shift of multipath channels in terms of directions and distances. To estimate AoA and ToF [52, 53], phase shifts or time delays are derived from CSI measurements obtained through an antenna array. The Multiple Signal Classification (MUSIC) algorithm is widely employed for AoA estimation. It leverages the eigenvalue decomposition of the covariance matrix derived from CSI measurements [54]. By computing the orthogonal steering vectors to the eigenvectors, MUSIC calculates the AoA values.

3.5.2 Learning Algorithms

The effectiveness of a HAR system relies not only on the quality of the CSI waveforms but also on the choice of an appropriate learning model. The most common Machine Learning (ML) algorithms used for the CSI recognition and classification tasks are Naïve Bayes, Support Vector Machine (SVM), decisions tree, ensemble methods, linear regression, and K-Nearest Neighbor (KNN). However, it is important to note that the effectiveness of ML algorithms heavily relies on the quality of the hand-crafted features extracted from the CSI data. Feature extraction methods for CSI data can be divided into two main types: those based on statistical analysis and those based on signal compression techniques.

The Naive Bayes algorithm [55] is a probabilistic approach commonly used for classification tasks. It operates by calculating the conditional probabilities of various activities. Naive Bayes can extract continuous features from real-time applications. Furthermore, Naive Bayes demonstrates effectiveness in dealing with high-dimensional CSI data.

SVM [56] aims to find the best hyperplane that effectively separates samples of different classes. SVM investigates the CSI signatures corresponding to

each activity, maps them into points in a space, and classifies into a class based on which side of the gap they fall on. SVM employs a kernel function to capture complex relationships for accurate recognition.

RF is an effective classification method specifically designed to handle high-dimensional datasets. It achieves this by compressing the data into more meaningful and informative features. The fundamental principle behind the RF model lies in the creation and integration of decision trees. Each decision tree contributes a vote towards a particular category, and the final class is determined by the majority vote across all the trees in the forest. Recently, DL algorithms have gained significant prominence across various domains, including computer vision, Natural Language Processing (NLP), speech recognition, and text and music generation. DL can automatically learn optimal features from data, which has led to its widespread adoption. Common DL algorithms include Convolutional Neural Network (CNN), recurrent neural networks such as LSTM, generative models, and self-supervised learning approaches. However, these algorithms do face challenges related to memory complexity, and they typically require a large amount of labeled data to enhance model performance.

CNN is a deep learning model that utilizes supervised training. Its primary layers consist of the convolutional layers, pooling (subsampling) layers, activation function, and fully connected layers for classification purposes. The convolutional layer automatically extracts local spatial features. It utilizes a set of k kernels (filters) with a specific size, which are convolved with the input data to produce feature maps. The activation function, a nonlinear function, is applied to the generated feature maps. Common activation functions include Sigmoid, tanh, Rectified Linear Unit (ReLU), and leaky-ReLU. Two types of pooling layers are commonly used to reduce the size of feature maps: average pooling and max pooling. Lastly, the fully connected layer combines all the learned features from the previous layers. It enables the sharing of these features in the classification process.

Recurrent Neural Network (RNN) handles the time-series data since it extracts the temporal dependencies from the input. However, RNNS often encounter challenges with the vanishing and exploding of gradient descent problems, particularly when dealing with long input sequences. To tackle this issue, there are new RNN models such as LSTM and Gated Recurrent

unit (GRU) are evaluated to solve the problem of vanishing and exploding gradients. LSTM approach can handle long-term dependencies within data when sufficient data and computational resources are available. The main idea of LSTM models is the existence of a memory cell state that generates meaningful and relevant outputs. LSTM architecture consists of three gates: the forget gate, the input gate, and the output gate. The forget gate is responsible for determining which information from the past should be discarded. The input gate determines which values from the current input should be used to update the new state of the LSTM. The output gate determines the output of the LSTM model based on the input data and the memory stored within the cell. GRU extracts the temporal dependency from time series datasets. Compared to LSTM model, the GRU model comprises two essential gates: the update gate and the reset gate. The update gate is the combination of the forget and input gates which determines the extent to which previous information from past time steps should be propagated into the future. On the other hand, the reset gate is the combination of the cell state and hidden state which determines the extent to which past information from the previous state should be forgotten.

Initially employed in image processing applications, the attention model has become increasingly relevant in the field of radio frequency sensing. The underlying concept of the attention mechanism involves selectively focusing on specific regions of input during the recognition task, allowing for improved performance. By integrating the attention model [57] with other deep learning approaches, it becomes possible to assign different weights to different features based on their relative importance.

In contrast to supervised learning models, Generative Adversarial Network (GAN) seeks to generate synthetic data that closely appears real data through a game between its generator and discriminator models [58]. Recently, GAN is used for tackling the CSI environmental diversity problem [59, 60] since it can capture the distribution of the input data and generate more labeled samples from a well-trained generator model. For domain-adversarial training, GAN can capture the changes of the new domain to extract the domain-invariant patterns for the seen and unseen environments.

Variational autoencoder (VAE) is a generative method that maps the data based on its distribution to a multivariate latent space based on a stochastic

variational inference [61]. VAE has been employed for various CSI based approaches like CSI compression [62] and CSI-based localization [63].

Few Shot Learning (FSL) is an efficient learning approach that requires only a small amount of training samples per class. It applies contrastive or prototypical learning techniques to enable effective learning with minimal data [64]. In one-shot learning which achieves acceptable results, the network trains on one sample per category. Yang et al. [65] introduced a SiaNet which is a gesture recognition system by leveraging few-shot learning using Siamese network.

3.6 Evaluation Metrics

In this thesis, we adopted two metrics to evaluate the performance of various models for our proposed systems, namely, the accuracy (ACC) and macro-averaged F1-score (F1-score). Accuracy is defined as the total number of correct predictions divided by total number of predictions. The F1-score represents the harmonic mean of two measures (precision and recall). It is a range of numerical values from 0 to 1, where the worst and best values are 0 and 1, respectively. The performance metrics are calculated by following equations from Equation 3.5 and Equation 3.8:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{3.5}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.6}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.7}$$

$$F1 - score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{3.8}$$

where True Positive (*TP*) refers to the model's accurate identification for the positive class, whereas True Negative (*TN*) refers to the negative predicted and actual values. False Negative (*FN*) represents the cases when the actual is positive and the model classified them as negative while False Positive

(*FP*) represents the cases where the actual is negative and the predicted is positive.

3.7 CSI-based Sensing Approaches

In [66], the authors presented a driver's face localization system based on the variance of CSI amplitude and phase, with the aim of detecting distraction and fatigue activity. It consists of three modules: CSI preprocessing, feature extraction, and classification. The noise removal techniques employed are the Butterworth filter for amplitude and linear transformation for phase. The feature set includes mean, standard deviation, median absolute deviation, maximum peaks, 25th percentile, and 75th percentile, which are fed to the classification module. The system utilizes the KSVM classifier, which combines the advantages of SVM and KNN and achieves a recognition accuracy of over 91%.

The WiHead system [67] is a system that utilizes wireless signals to measure human head orientation in various directions, including yaw, roll, pitch, and their combinations, for obtaining feedback in online courses. It uses 56 subcarriers at 2.4 GHz with the Atheros CSI extractor tool to retrieve phase and amplitude, which are then filtered and calibrated to eliminate noise and unpredictability. The filtered data is then sent to a PCA method for dimensional reduction. Additionally, WiHead developed a CNN model that achieved recognition accuracy of 90% for three different head motion angles: pitch, roll, and yaw.

In [68], the authors utilized the CSI waveforms from ESP32 microcontroller to track the head motions. They aim to estimate the student's engagement in online courses based on their head movements. CSI amplitude is extracted and preprocessed via Hampel filter, discrete wavelet transforms and smoothed by Savitzky-Golay. After that, the filtered amplitude is fed to the XGBoost (XGB) model for the classification task which achieved 98% recognition accuracy.

Palipana et al. [69] introduced the FallDeFi system, which aims to detect falls by utilizing robust features that can withstand changes in the environment.

The system employs the amplitude of CSI as the fundamental signal. To mitigate the impact of environmental noise, a LPF is applied. Subsequently, a noise filtering technique based on DWT is employed. This filtering process effectively eliminates in-band noise while preserving important high-frequency components and minimizing signal distortions. Furthermore, PCA is utilized for stream uncorrelation and selection. This step aids in reducing redundancy and selecting the most informative components, which are then fed as input to a SVM for fall detection. Notably, the FallDeFi system demonstrated an 80% recognition accuracy even when the environment changes.

Yousefi et al. presented a human activity recognition system [70]. The authors recorded CSI waveforms, namely UT-HAR, using Intel 5300 CSI in one location for seven daily human activities with 5000 samples. For dimensionality reduction, the authors applied PCA to the normalized amplitude. After that, STFT is calculated to generate the spectrogram fed to a LSTM network to perform the classification task. Zheng et al. [71] introduced Widar which is a gesture recognition system. The main objective is to build a domain robust system. The authors collected 43,000 samples for 22 gesture categories from multiple locations employing the Intel 5300 NIC tool with three Tx and three Rx. To preserve the gesture-related features, the velocity of the collected CSI was calculated. Subsequently, a CNN-GRU model was employed as a classifier to compress the data, extract spatial features using convolutional layers, and capture temporal dependencies from the resulting feature maps. Widar3.0 achieved 92.7% and over 82% in-domain and cross-domain recognition accuracy.

Yang et al. [72] proposed AutoFi system which transfers knowledge from randomly collected CSI samples into human gait recognition. The AutoFi framework addressed the identified gaps by introducing a novel self-supervised learning approach. This approach utilized contrastive learning and mutual information to enhance the transferability of the framework. Additionally, a geometric structural loss is developed to further improve the framework's ability to adapt to different downstream tasks. AutoFi demonstrated successful cross-task transferability in the field of WiFi sensing. Notably, it achieved automatic WiFi sensing in new environments without the need for prior data collection. The AutoFi system was implemented and validated its robustness and effectiveness. Furthermore, simulations conducted using publicly avail-

able datasets, such as Widar [71] and UT-HAR [70], demonstrate that AutoFi outperforms existing domain adaptive systems.

WiGRUNT [73] is a gesture recognition system that leverages the attention mechanism to tackle a domain shift problem. The authors applied their proposed method on Widar [71] dataset. The CSI ratio technique is applied to remove the phase offset. The main idea behind CSI ratio is to calculate the ratio of the CSI measurements between two adjacent receiving antennas as calculated in Equation 3.9.

$$H_q(f,t) = \frac{H_i(f,t)}{H_{i+1}(f,t)} \tag{3.9}$$

, where $H_i(f,t)$ is the complex CSI matrix for the ith antenna. After that, the actual phase $P = \angle H_q(f,t)$ can be extracted from $H_q(f,t)$. To extract spatial-temporal features from the CSI phase tensor, WiGRUNT utilized the ResNet network as a backbone. Additionally, dual attention stages were incorporated to focus on extracting the most relevant features for gestures while disregarding features related to the sensing environment. Notably, WiGRUNT achieved recognition rates of over 93% and 83% with and without pretraining using the ImageNet dataset, respectively.

In [74], Wang et al. introduced CAUTION which is a user identification system that tackles the domain shift problem by leveraging FSL technique. The authors recorded the CSI measurements using the Atheros tool with one transmitter antenna and a three-antenna receiver from two different locations. Firstly, the authors extracted the amplitude and fed it to the convolutional network to compress the data and extract the spatial filter. The compressed features were fed to a prototypical network for few shot learning. It achieved over 87% accuracy in the worst scenario.

Gao et al. [75] introduced a gesture recognition system called WiGesture, which aimed to overcome location dependencies in gesture recognition. This system employed the extraction of location-independent Motion Navigation Primitives (MNPs) to capture changes in motion direction. The researchers evaluated the performance of WiGesture by conducting experiments involving ten different categories of gestures across four distinct locations. WiGesture

surpassed 90% recognition accuracy for all tested gesture categories and locations.

Table 3.2: Summary of related Wi-Fi CSI work.

Reference	Application	Base Signal	Preprocessing	Classifier	Location Independent	Multi-user Environment
[60]	HAR	Ampltitude	GAN	CNN	✓	×
[66]	Driver's face localization	Amp+Phase	Butterworth filter+ Linear Transformation	SVM + KNN	×	×
[67]	Head motion detection	Amp+ Phase	STFT+ PCA+ Linear Transformation	CNN	×	×
[68]	Head motion detection	Amp	Savizky-Golay	XGBoost	×	×
[76]	HAR	Amp	SVD	CNN + ProtoNet	✓	×
[74]	Gesture Recognition	Phase	CSI Ratio	ResNet +Attention	✓	×
[74]	User Identification	Amp	CNN	Prototypical	✓	×
[77]	HAR	Amp+ Phase	CM + PCA +FFT	CNN-LSTM + MatNet	✓	×
[78]	HAR	Amp+ Phase	Phase Sanitiziation	Inception model	✓	×
[79]	Head motion detection	Amp	STFT	Inception model	×	✓
[80]	Head motion detection	Amp	WMA	ECA model	✓	✓

4

Passive Wi-Fi CSI Sign Language Recognition

4.1 Introduction

This chapter presents an overview of sign language recognition systems utilizing Wi-Fi CSI. The primary aim of this chapter is to provide an introduction to relevant research endeavors that contribute to improving the quality of life for individuals with hearing impairments. By harnessing the inherent characteristics of Wi-Fi signals, such as privacy preservation, user-friendliness, and convenience, these systems offer promising avenues for empowering the deaf community.

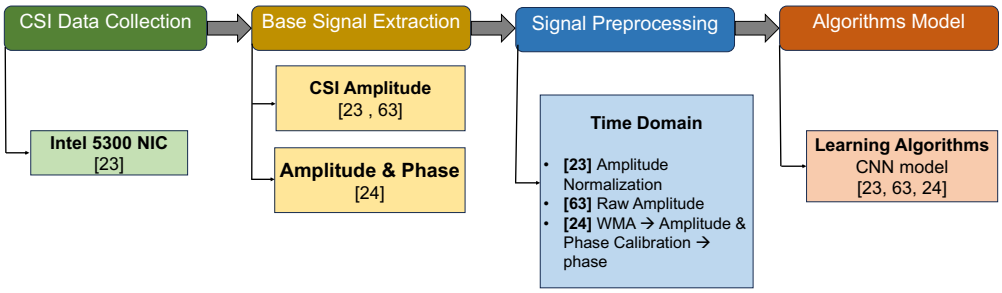


Figure 4.1: General architecture for SignFi systems

4.1.1 Background

In recent years, the field of human gesture recognition has garnered significant interest due to its diverse applications such as smart home systems, healthcare, virtual reality, and sign language recognition. Conventional sensing techniques include camera-based approaches [81] and wearable

sensors [82], each accompanied by inherent limitations. Camera-based methods achieve satisfactory accuracy by analyzing video frames and images; however, they raise concerns regarding privacy due to the potential leakage of facial information. Conversely, wearable sensors offer advantages such as lightweight design, cost-effectiveness, and ease of gesture monitoring and detection. However, they suffer from user inconvenience, and the failure to wear the sensor renders them ineffective. To address these shortcomings, there is a growing interest in Wi-Fi-based solutions, which have emerged as a promising alternative in the field of gesture recognition. Wi-Fi-based approaches effectively overcome the aforementioned limitations associated with camera-based and wearable sensor methods, thus attracting considerable attention from researchers in this field.

This chapter aims to present the contributions and challenges encountered in my previous research study [36], which utilized the SignFi dataset [35]. The SignFi dataset is a publicly available collection of daily 276 sign language gestures, encompassing a substantial number of samples per gesture. Fig. 4.1 provides a comprehensive overview of the general architecture employed in several existing SignFi dataset systems. The primary contributions of the aforementioned study [36] are as follows:

1. Investigating the impact of different variations in the base signal on the performance of sign language recognition.

2. developing the traditional CNN architecture and comparing its performance with other deep learning algorithms.

However, it is important to acknowledge the limitations of this study, which are taken into account and addressed in the subsequent chapters of this thesis. These limitations include:

1. The utilization of a public dataset collected using the Intel 5300 NIC, which possesses certain limitations as detailed in Chapter 3.2.1.

2. The SignFi dataset was obtained in a single-user environment, which does not fully reflect real-life scenarios where deaf individuals are often surrounded by others.

3. The deep learning model employed in the study exhibits complexity, lengthy training and testing times, and requires substantial memory

resources for its parameters, making it impractical for real-world deployment.

These limitations are thoroughly examined and addressed in subsequent chapters to enhance the applicability and practicality of the proposed approach for enhancing the quality of people.

4.2 Sign language based system

The pipeline for all CSI based systems is as described in Chapter 3.1. In this section, we will present my previous study [36] and compare it with other studies that use SignFi dataset in their work.

4.2.1 CSI Data Collection

Ma et al.[35] conducted a comprehensive data collection process to gather a substantial amount of CSI traces related to daily sign language gestures. The dataset was gathered by employing an access point (AP) with three external antennas as a transmitter, and a receiver equipped with an Intel 5300 NIC [40] with one internal antenna in two distinct environments: a laboratory setting and a home scenario. These environments differ in terms of room sizes, the distances between the AP, receiver, and transmitter's antenna orientations. The utilization of the 802.11n CSI tool [40] enabled the capture of 30 sub-carriers for each antenna pair, resulting in complex CSI measurements with dimensions of $200 \times 30 \times 3$, where each dimension corresponds to the number of packets, sub-carriers, and antenna pairs, respectively. Table 4.1 summarizes the properties of SignFi dataset.

Table 4.1: SignFi Dataset

Dataset	Environment	Number of Signs	Number of Users	Number of Samples
D1	Home	276	1	2760
D2	Laboratory	276	1	5520
D3	Laboratory	150	5	7500

4.2.2 Signal Preprocessing

Ma et al. [35] employ the variations in Channel State Information (CSI) amplitude as fundamental signals to represent sign gestures. Subsequently, a normalization procedure is applied to the CSI amplitude in order to prepare it as input for the classification model. The resulting dataset possesses dimensions of $200 \times 30 \times 3$, which effectively captures intricate details and nuances inherent in fine-grained actions associated with sign gestures.

Bastwesy et al.[36] employ a comprehensive approach in which they extract both the amplitude and phase variations as fundamental signals for the classification task. To mitigate the influence of environmental noise, they apply a Weighted Moving Average (WMA) filter to the raw amplitude. Additionally, in order to address the inherent randomness introduced by the hardware manufacturer, as discussed in Chapter2.2.2, a linear transformation is applied to the raw phase. Furthermore, the smoothed amplitude and true phase are concatenated on the subcarrier dimension to make the classification input size equal $200 \times 60 \times 3$.

Wei et al. [83] present a novel deep learning model that is capable of automatically extracting distinctive patterns from raw CSI amplitude data, without the need for preprocessing techniques to handle outliers resulting from environmental noise.

4.2.3 Learning Model

Ma et al. [35] propose the SignFi deep learning algorithm, which leverages a 9-layer CNN as the classification model. CNNs possess the ability to automatically learn parameters and features, making them well-suited for solving complex problems. Furthermore, CNNs exhibit high computational efficiency during the inference stage, even when confronted with a large number of classes. SignFi adopts three 3×3 kernels with a stride of 1 in both the vertical and horizontal directions. To maintain the output size of the convolutional layer and ensure equal utilization of all inputs, SignFi employs padding of 1 in both the vertical and horizontal directions. This padding involves adding a column/row of zeros along the edges of the original input.

Subsequently, a batch normalization layer, ReLU activation function, and average pooling layer are applied. The resulting feature maps are concatenated at the flatten layer, facilitating the flow of information to a fully connected layer comprising 276 neurons. This layer is followed by a softmax layer, which performs the classification task by assigning probabilities to each class label.

Bastwesy et al. [36] develope the SignFi CNN algorithm by enhancing the convolutional block. The classification model in their study comprises two CNN blocks, each consisting of sequential layers. The main objective of these CNN blocks is to autonomously extract relevant features associated with each gesture. The first layer is the Convolution Layer, which employs 32 filters of size 5×5 and a stride of 1. Padding is applied by adding zeros around the input edges with a padding value of 1, enabling the convolution operation to encompass all inputs. The output of this layer consists of a set of feature maps that serve as input for the subsequent layer. The Batch Normalization Layer is utilized to normalize the inputs, thereby accelerating the training process and enhancing network performance by mitigating the impact of overfitting followed by ReLU Layer serves as a nonlinear activation function, setting any input value below zero to zero. The Average Pooling Layer plays a role in reducing the number of connections and parameters in the network. In this framework, average pooling is applied with a window size of 3×3, returning the average value within this window, with a stride of 3. The Dropout Layer is incorporated to prevent overfitting during the training stage. It improves the overall performance of the network by randomly eliminating inputs based on a defined probability.

The learning model ended with two fully connected layers to aggregate the learned features from preceding layers and contribute to the classification process. The first fully connected layer includes 1000 neurons, while the second layer consists of 276 neurons, corresponding to the number of classes. The first layer utilizes the ReLU activation function and incorporates a dropout rate of 0.5 to mitigate overfitting. The second fully connected layer employs the softmax function, which assigns each CSI input to one of the 276 sign classes based on the predicted probabilities. This softmax classification layer facilitates the final classification of the input signals into distinct sign gestures.

Wei et al. [83] propose a deep learning model of substantial scale, consisting of seven convolutional layers followed by three parallel average pooling layers. The resulting feature maps from these layers are subsequently concatenated using a flatten layer, allowing them to be fed into a fully connected layer with 1000 neurons. To mitigate overfitting, a dropout layer with a keep rate of 0.8 is applied after the aforementioned fully connected layer. To further enhance the model's performance, an additional fully connected layer with 1000 neurons followed by dropout layer with keep rate equal to 0.8 is introduced. During training and inference, the softmax layer is utilized as the final layer, facilitating the classification process.

4.3 Performance Evaluation

The proposed model, as presented in [36], is subjected to a comparative analysis against the models introduced in prior works, namely [35, 83, 70, 84]. The evaluation metric for comparison is recognition accuracy. The experimental data is processed and analyzed using Google Colab, specifically the professional version with access to 2 Terabytes of storage. The computational infrastructure provided by Google Colab includes a server equipped with 26 Gigabytes of RAM and a P100 GPU. The implementation of the proposed model is realized in Python 3.6, utilizing the Keras deep learning library, which is a Python-based framework [85]. This section of the thesis is divided into two parts: a comprehensive description of the dataset employed and a thorough analysis of the obtained results.

4.3.1 Experiment Setup

The SignFi dataset comprises three publicly available datasets, obtained through experimental measurements conducted in both laboratory and home scenarios. These datasets include raw CSI measurements, capturing variations in room sizes, distances between the transmitter and receiver, and antenna orientations. The CSI measurements within these datasets are sampled at a rate of 200 Hz, with the duration of a single gesture instance

ranging between 0.5 seconds and 2.5 seconds. Notably, instances containing gestures have been carefully segmented, ensuring that each instance exclusively represents a single gesture.

Table 4.1 provides a comprehensive summary of the three SignFi datasets. In D1, a total of 276 gestures are performed by a single user within a home scenario, yielding 2,760 instances for data collection. D2 and D3 encompass CSI instances collected simultaneously from the receiver and transmitter, respectively, situated in a laboratory setting. A total of 5,520 CSI instances are recorded within this context. Similarly, D3 also encompasses CSI instances collected within the laboratory scenario, including 7,500 instances of gestures performed by five different users.

4.3.2 Results

The performance evaluation of the system [36] is conducted using a 5-fold cross-validation approach. This evaluation encompasses the home dataset, lab dataset, the combined dataset consisting of 8,280 instances from both home and lab environments, as well as the dataset comprising gestures performed by five users for system validation. Additionally, a self-test is conducted for each individual user, employing a 5-fold cross-validation methodology.

To facilitate the cross-validation process, the entire dataset is randomly divided into five folds. Subsequently, one fold is designated for testing purposes, while the remaining folds are utilized for training the system. This process is repeated five times, resulting in five distinct runs. Finally, the average accuracy across all five runs is computed as a measure of the system's performance. The evaluation of the proposed framework [36] in comparison to other deep learning techniques, namely LSTM and Attention-based Bidirectional LSTM (ABLSTM).

ABLSTM combines the Bidirectional LSTM (BLSTM) model with an attention mechanism. BLSTM processes sequential data in both forward and backward directions by employing forward and backward LSTM layers. This enables the automatic extraction of informative features. The attention model plays

a crucial role in assigning differential weights to various features, enhancing the overall performance by emphasizing more important features [26].

LSTM and ABLSTM are particularly well-suited for time series data, including Wi-Fi CSI data, which exhibit temporal dependencies. Previous studies [70, 84] have demonstrated the impressive performance of LSTM and ABLSTM in Wi-Fi CSI-based human activity recognition.

In [36] comparative analysis, the LSTM model [70] and ABLSTM [84] are employed on the SignFi datasets. For a fair comparison, raw CSI amplitude and phase as the input feature vector utilized for LSTM, with a single layer comprising 200 LSTM units. The LSTM approach achieves average accuracies of 74.06%, 49.82%, 76.92%, 56.9%, and 63.5% for the home, lab, home+lab, 5-users, and self-test scenarios, respectively. In the ABLSTM model [84], The raw CSI phase and amplitude are fed into the attention-based bidirectional long short-term memory model. The ABLSTM achieves recognition accuracies of 95.44%, 96.2%, 94.94%, 73.83%, and 70% in the home, lab, home+lab, 5-users, and self-test scenarios, respectively.

Based on the results, the proposed system [36] exhibits average recognition accuracies of 99.674%, 99.855%, 99.73%, 93.84%, and 99% for the home, lab, home+lab, 5-users, and self-test scenarios, respectively. These findings are summarized in Table 4.2. Notably, the performance of all approaches is generally higher in the lab environment compared to the home environment, owing to the increased complexity of the latter. The superior performance of the proposed model [36] can be attributed to its ability to automatically extract features, enabling accurate recognition of diverse sign gestures in complex environments.

Table 4.2: Overall performance for all systems based on SignFi dataset

Method	Home	Lab	Home & Lab	5−Users	Self test
[35]	98.91	98.01	94.81	86.66	98
[83]	**99.89**	**99.98**	–	–	**99.65**
[70]	74.06	49.82	76.92	56.9	63.5
[84]	95.44	96.2	94.94	73.83	70
[36]	99.67	99.85	**99.73**	**93.84**	99

[36] summarized the time consumption for different frameworks in Table 4.3. The SignFi system requires 8.28ms for training and 0.62ms for testing. On the other hand, the LSTM model necessitates 7.2ms for training and 3ms for

testing. The ABLSTM model exhibits a relatively long training time of 27.2ms compared to other approaches, with a testing time of 10ms. In contrast, the proposed CNN model [36] demonstrates the shortest training time of 1ms among all approaches, while its testing time is 0.66ms, which is comparable to the training time of SignFi.

Based on these findings, it can be inferred that the proposed model [36] is suitable for real-time Wi-Fi CSI-based sign gesture recognition. It achieves satisfactory accuracy, and its testing time is minimal, indicating its potential for real-time applications.

Table 4.3: Time Consumption for all systems based on SignFi dataset

Method	[35]	[83]	[70]	[84]	[36]
Training	8.28	–	7.2	27.2	**1**
Testing	**0.62**	–	3	10	0.66

4.4 Summary

The primary contribution of the proposed methodology [36] lies in the development of a robust deep learning model capable of achieving satisfactory recognition accuracy across various environments, while also demonstrating resilience in the face of user diversity. Remarkably, the system achieves recognition accuracy exceeding 99% when subjected to environmental variations and surpasses 93% when confronted with user variations. However, it should be noted that the study [36] relies on a publicly available dataset [35], collected using the Intel 5300 NIC, which is accompanied by inherent limitations discussed in the previous chapter.

Furthermore, the data collection process employed in the study [36] was conducted within a single-user environment, where the user performed the gestures in isolation. However, this setup does not align with real-world scenarios, as communication between individuals with hearing impairments and others is a crucial aspect. In subsequent chapters, we aim to address these challenges by introducing novel methodologies and collecting our own dataset using the ESP32, a more advanced tool for capturing CSI. Additionally, we present two use cases that aim to improve the quality of life for individuals,

namely communication assistance for quadriplegic individuals and mental health support for workers.

Moreover, we delve into the challenges associated with deploying Wi-Fi CSI in real-world settings, specifically focusing on robustness in the face of user diversity, location variations, and session variations. Wi-Fi signals can be influenced by the sensing area, and different users may perform the same gestures with varying styles and speeds. These challenges are thoroughly investigated in the subsequent chapters of this thesis.

5

Wi-Nod: Head Nodding Recognition by Wi-Fi CSI Toward Communicative Support for Quadriplegics

5.1 Introduction

This chapter introduces and validates a novel contactless sensing system called Wi-Nod, which utilizes ESP32 nodes as a CSI toolkit. Notably, this work stands out as the first to collect Wi-Fi signals in a multi-human context environment. Unlike previous studies that focused on single-user scenarios, our approach considers the presence of both a target patient (quadriplegic individual) and a caregiver, who acts as a scatter providing additional multi-path propagation in the sensing area. This realistic setup enhances the reliability and applicability of the system. Fig. 5.1 presents an overview of the architectural framework proposed for the Wi-Nod system.

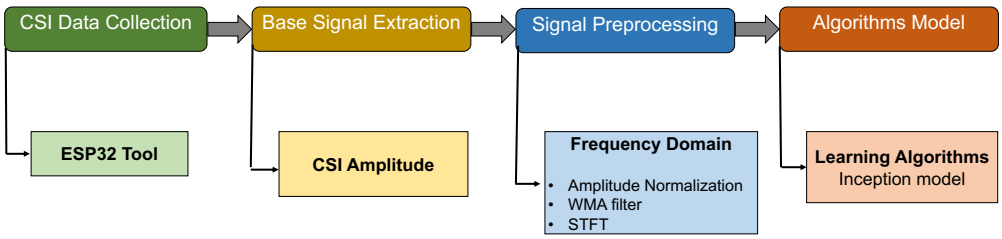


Figure 5.1: General architecture for Wi-Nod system

5.1.1 Background

The advent of Internet of Things (IoT) technologies has sparked significant interest in the realm of HAR and GR applications. These technologies empower passive sensing capabilities on diverse smart IoT devices, including Wi-Fi devices, smartphones, and smart speakers [86, 87, 88].

Such advancements have opened up new possibilities for leveraging Wi-Fi CSI as a means to recognize head nodding gestures and provide communicative support specifically tailored for individuals with quadriplegia. Quadriplegia, resulting from spinal cord injuries, severely impacts an individual's mobility and ability to engage in conventional forms of communication. According to a study [89] conducted by the American Spinal Injury Association in 2016, an estimated 1.3 to 2.6 million individuals experience varying degrees of spinal cord injuries annually, highlighting the pressing need for effective communication solutions for quadriplegics. Therefore, head nodding gestures, despite having a limited range of motion, serve as a vital means of expression for these individuals, allowing them to indicate affirmative responses, answer binary questions, and convey basic communication cues.

Quadriplegic individuals, even in the later stages of spinal injury, retain slight mobility in their head and eyeballs, offering a potential avenue for communication and interaction. Although vision-based eyeball-tracking systems, exemplified by Stephen Hawking's case, have demonstrated effectiveness, their complexity and cost hinder widespread adoption.

To address these challenges, there is a need for accessible and affordable communication solutions for quadriplegics. Wi-Fi CSI emerges as a promising approach. By leveraging the unique characteristics of wireless signals, Wi-Fi CSI captures subtle variations induced by human movements, including head nodding gestures. The ubiquity of Wi-Fi infrastructure enables a non-invasive and cost-effective means of recognizing and interpreting head nodding gestures, providing valuable communicative support.

Our idea is inspired by the WiMorse [90] that employed Intel 5300 NIC to collect the CSI waveforms produced by a finger. The authors created their own code that encoded the two Morse symbols based on the subtle finger movements. The authors built a mathematical model to detect the

characters and numbers using WiFi CSI measurements derived from these finger movements. WiMorse is a position-independent system that can be deployed in different environments. The system achieved an average accuracy of 95%.

By harnessing Wi-Fi CSI, we adopt a Morse code-inspired approach for representation. Head movements are used to convey Morse symbols, with downward and rightward head motions representing dots and dashes, respectively. Additionally, a third symbol, space, is introduced to separate words, indicated by a leftward head movement.

This research aims to pioneer the use of Wi-Fi CSI for Head Nodding Recognition, specifically tailored for communicative support in quadriplegic individuals. By designing and implementing the Wi-Nod system, we strive to empower quadriplegics to express themselves and engage in meaningful communication using intuitive head movements. Through extensive validation and experimentation in realistic multi-human contexts, we aim to establish the effectiveness and practicality of our approach.

5.1.2 Research Contributions and Questions

The main contributions of this study are as follows:

1. We collect Wi-Fi signals in a multi-human context environment, incorporating both quadriplegic patients and caregivers, unlike previous studies limited to single-user scenarios.
2. We extract the informative context related to the head motions in the presence of multiple objects around the patient, accounting for the influence of temperature and humidity on wireless signals, and accommodating variations in motion speeds among different patients.
3. We evaluate the system's performance in diverse scenarios to illustrate the effectiveness and robustness of our proposed system.

In this study, we answer the following research questions:

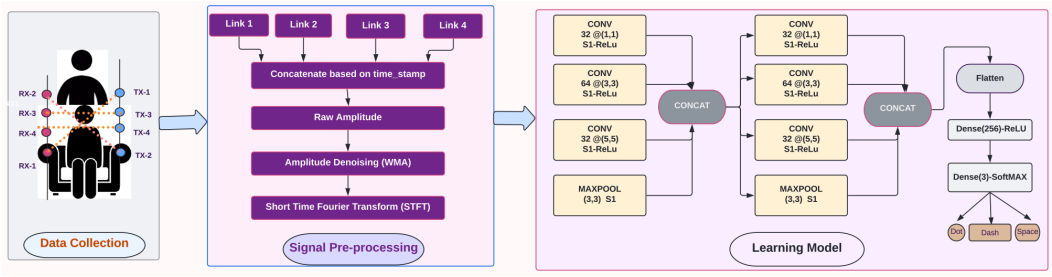


Figure 5.2: Wi-Nod System Framework

Q1 How can the system accommodate variations in motion speeds among different quadriplegic patients to achieve accurate and consistent recognition of head nodding gestures?

Q2 What is the impact of considering user and session diversity in a multi-human context on the performance of the Wi-Nod system?

5.2 Methodology

5.2.1 System Overview

Our proposed Wi-Nod system consists of three key modules aimed at capturing and analyzing patterns of channel variation associated with head nodding gestures. Fig 6.2 provides an overview of the general architecture of the Wi-Nod system, which comprises the data collection module, signal preprocessing module, and learning model. Each of these modules plays a crucial role in the overall functioning of the system.

5.2.2 Data Collection

Data was collected from two participants in two distinct time sessions: one in the morning and another in the evening. This approach aimed to capture variations in head nodding gestures that might occur throughout the day due to factors such as fatigue, alertness, or environmental conditions. By considering multiple time sessions and participants, the dataset encompasses

a diverse range of the head nodding patterns. The data collection process involved the utilization of eight ESP32 nodes, which were divided into two groups: transmitters and receivers. Each group performed a specific role in capturing the Channel State Information (CSI) streams between the nodes. The transmitters emitted Wi-Fi signals, while the receivers recorded the received signals and measured the resulting CSI. The configuration of the ESP32 nodes enabled the capture of CSI streams, which contain valuable information about the wireless channel and its variations caused by head nodding gestures. By analyzing these CSI streams, it becomes possible to extract meaningful patterns and features that can facilitate a head nodding recognition.

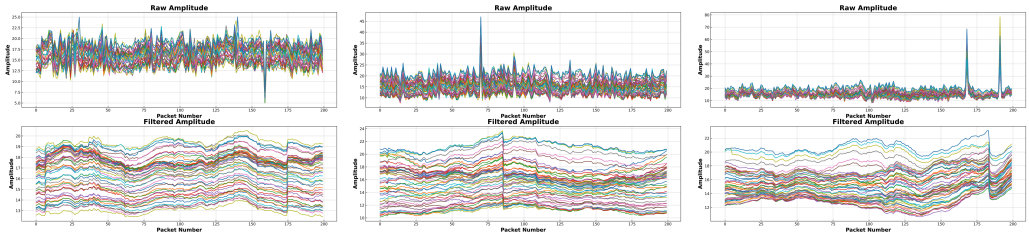
5.2.3 Signal Preprocessing

signal preprocessing plays a vital role in preparing the raw CSI measurements for effective classification. This module aims to address the inherent noise present in the CSI waveform and enhance the performance of the subsequent learning model. The signal preprocessing module consists of several stages, each serving a specific purpose, as described below:

1. **Data Segmentation**: The signal segmentation stage involves splitting the CSI measurements of each link based on their time stamp. This step enables the fusion of signals from each link, creating a unique pattern for each user's head motion. The segmented signals can then be mapped to the corresponding Morse code. This process helps capture the temporal aspects of head nodding gestures.

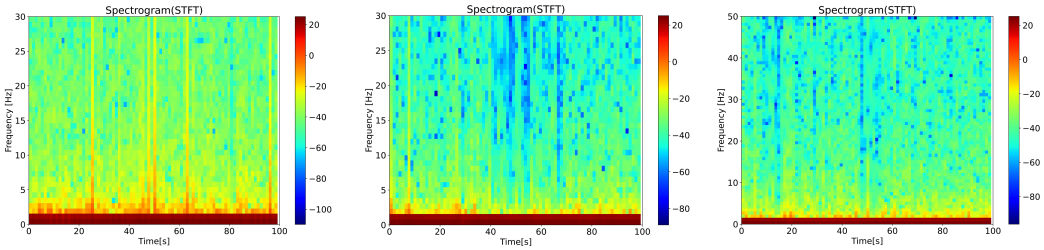
2. **Interpolation**: To overcome the issue of packet loss in the CSI measurements, padding interpolation is applied. This technique preserves the distribution of the received packets, ensuring that important information is not lost due to missing data. Interpolation helps maintain the integrity and continuity of the CSI waveform.

3. **Amplitude Extraction**: In this work, the amplitude is extracted as the base signal for analysis. The amplitude exhibits variations that are correlated with head motion and is considered more reliable and less



(a) Dot (b) Dash (c) Space

Figure 5.3: Raw and Filtered Amplitudes of Three Symbols across All Subcarriers in 1st Link



(a) Dot (b) Dash (c) Space

Figure 5.4: Spectrograms of 13th Subcarrier in 3rd Link for Three Symbols

random than the CSI phase. By parsing the CSI files, the amplitude is extracted and utilized as the primary signal for further processing.

4. **Amplitude Noise Removal**: The noise removal stage aims to smooth the raw amplitude and eliminate outliers caused by environmental changes. To achieve this, Weighted Moving Average (WMA) is applied. The WMA effectively reduces noise and enhances the clarity of the amplitude signal, ensuring that meaningful patterns associated with head nodding gestures are preserved. In general, the filtered amplitude can be calculated as:

$$A'_t = \frac{1}{m + (m-1)... + 1}.[m.A_t + (m-1).A_t - 1 + (m-2).A_t - 2 + ... + A_t - m + 1]$$

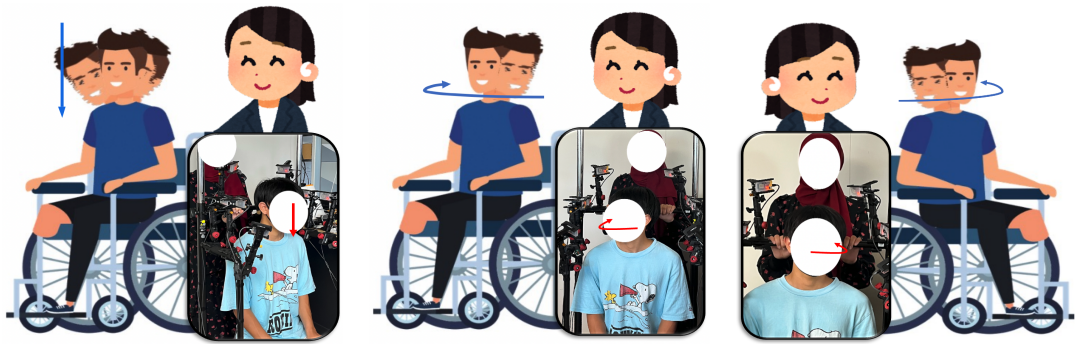
A'_t is the weighted average amplitude within a window size m for time t. Fig. 5.3 illustrates the results of the weighted moving average for each symbol sample, and the color curves represent the amplitude of each subcarrier within the first link, as it is observed that the amplitude is smoother and outliers are eliminated.

5. **Spectrogram Extraction**: Head motions and human movements introduce complex variations in the CSI amplitude. Users may perform the same head motion at different speeds, which can be captured through frequency variations in the spectrogram. To extract spectrograms, a sliding window technique is applied to the filtered amplitude, segmenting the signal into equal-sized segments. The Fast Fourier Transform (FFT) is then performed on each segment, converting the signal from the time domain to the frequency domain. This process produces spectrograms through Short-Time Fourier Transform (STFT), which provide a visual representation of the frequency information associated with head nodding gestures. Fig. 5.4 illustrates the spectrogram of the subcarrier with index 13th for each link, showcasing the frequency variations corresponding to different head nodding symbols.

5.2.4 Learning Model

By employing this feature extraction and classifier stage, utilizing the inception model, the system achieves improved accuracy and performance in recognizing and classifying head nodding gestures. The automatic feature extraction process, with low computational complexity, enhances the system's ability to understand and interpret the nuances of head movements, facilitating effective communicative support for quadriplegic individuals. The inception model, utilized in this work, is characterized by its wider architecture rather than deeper, enabling a faster learning process. This design choice allows for parallel manipulation of the classifier input, as depicted in Fig. 6.2. The spectrogram, representing the frequency variations of head nodding gestures, is inputted into the first inception module for meaningful feature extraction.

1. Feature Extraction: The feature extraction stage consists of two inception blocks, each comprising three parallel convolution layers with different numbers and sizes of kernels. After each convolution layer, the rectified linear unit (ReLU) activation function is applied. A maximum pooling (MaxPool) layer follows with a stride value of one, and the outputs of these layers are concatenated. The first convolution layer employs 32 kernels with a size of (1×1), the second convolution layer



(a) Dot (down) (b) Dash (right) (c) Space (left)

Figure 5.5: Head Motions Used in Experiment: Dot, Dash, and Space

has 64 kernels with a size of (3×3), and the third convolution layer utilizes 32 kernels with a size of (5×5). Finally, a MaxPool layer with a kernel size of (3×3) is applied.

2. Classifier: The output of the second inception layer is flattened to transform the multidimensional feature representation into a one-dimensional vector. This flattened representation is then passed to a fully connected layer, where all the extracted features are combined. The final classification layer is represented by a softmax layer with three classes, corresponding to the symbols dot, dash, and space, which represent different head nodding gestures.

5.3 Performance Evaluation

To evaluate our proposed system, we test the performance of Wi-Nod in a multi-human context environment to verify the robustness of the system.

5.3.1 Experiment Setup

In this phase, we deployed eight ESP32 nodes to construct our head nodding system. Among these nodes, four were utilized as transmitters connected to mini-PCs, while the remaining nodes served as receivers.

To evaluate head nodding gestures, we enlisted two participants, one male and one female, to perform these motions in a laboratory environment. Each participant executed head movements corresponding to three symbols: dot, dash, and space, represented by moving the head down, right, and left, respectively, as illustrated in Fig. 5.5. The participants were instructed to perform each gesture for a duration of four minutes. Consequently, we obtained four distinct datasets, each containing approximately 720 samples.

Notably, the data collection took place in a multi-human environment, simulating real-world scenarios faced by quadriplegia patients in wheelchairs. To replicate these circumstances, one person held a frame and moved behind the participant, while several other individuals were present in the surroundings. To assess the robustness of our system, data collection was conducted during two different time sessions: morning and evening. These sessions introduced variations in the environmental conditions around the frame. The morning session involved a smaller number of people (approximately three individuals) in the laboratory, while the evening session included a larger group (around 10 individuals). Python and the Keras platform were employed for the processing of the collected data, facilitating subsequent analysis and evaluation.

5.3.2 Results

We evaluated the performance of our proposed system using the cross validation approach, allocating 70% of the data for training and the remaining 30% for testing. The datasets were collected during different time sessions throughout the day, denoted as $F1$, $M1$, $F2$, and $M2$. Here, the numbers represent the session type, with 1 indicating the morning session and 2denoting the evening session. The symbols F and M indicate whether the dataset was gathered by a female or male, respectively, with a caregiver holding the frame during data collection. In our evaluation, we investigated the impact of different link configurations, namely L1_2, L3_4, and L_all. The L1_2 configuration represents the concatenation between link 1 and link 2, forming a diagonal configuration. The L3_4 configuration represents the concatenation between link 3 and link 4, creating a horizontal configuration. Lastly, the L_all configuration involves the concatenation of all links together.

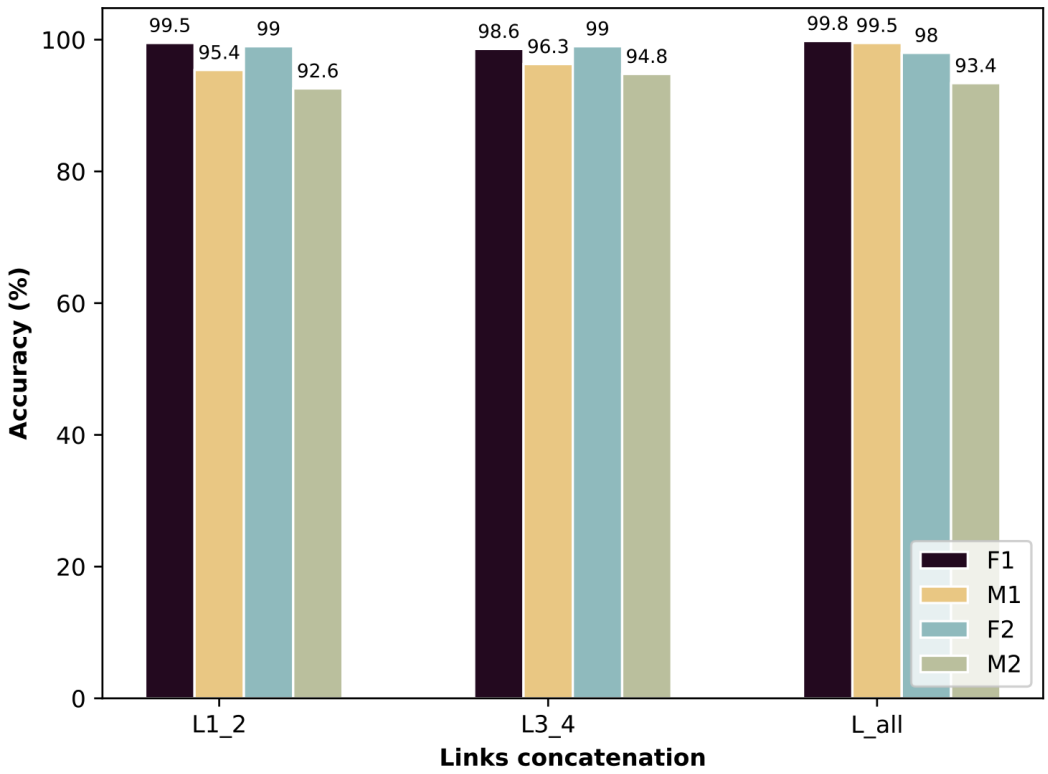


Figure 5.6: The Recognition Accuracy of Different Dataset

The concatenation process is performed based on the subcarrier index dimensions. To evaluate the system's performance, we analyzed the accuracy and confusion matrix metrics. These metrics provide insights into the system's classification accuracy and the patterns of misclassifications, respectively.

Table 5.1: Overall System Accuracy

Dataset	L1_2				L3_4				L_all			
	RAW_ AMP	WMA_ PCA_STFT	RAW_ STFT	WMA_ STFT	RAW_ AMP	WMA_ PCA_STFT	RAW_ STFT	WMA_ STFT	RAW_ AMP	WMA_ PCA_STFT	RAW_ STFT	WMA_ STFT
F1	97.7	83.4	98.16	99.5	94.93	85.71	96.3	98.6	98.62	84.79	99.1	**99.5**
M1	92.1	67.59	92.1	95.4	92.6	77.42	91.24	96.31	99.1	69.91	95.37	**99.0**
F2	98.15	80.09	98.6	99.07	92.13	78.2	92.13	**99.0**	98.7	84.26	97.7	98.1
M2	93.1	44.0	92.1	92.6	91.5	71.23	89.15	**94.8**	96.7	68.87	86.1	93.4
F1_M1	43.3	36.57	34.6	35.4	63.66	38.6	**68.2**	64.8	27.2	28.1	34.3	44.7
F1_F2	58.6	53.8	78.8	**79.1**	32.2	33.3	35.6	32.9	41.3	43.1	60.1	61.5

As shown in Fig. 5.6, the proposed system achieved over 95% recognition rate. Table. 5.1 reveals that the diagonal link configuration, $\mathbf{x} \in \mathbb{R}^{100\times104}$ where x is the classifier input, yields the highest accuracy when multiple individuals are present within the sensing area. This configuration enables the classifier to effectively extract relevant information from the spectrogram associated with

the patient's head nodding, while mitigating the impact of environmental noises and other individuals' movements. On the other hand, the all links configuration, $\mathbf{x} \in \mathbb{R}^{100\times208}$, exhibits superior performance in scenarios where the number of people surrounding the patient is relatively low. Moreover, it is obvious that the horizontal link configuration is more susceptible to variations in the environment and the presence of surrounding individuals, resulting in the lowest accuracy among the other link configurations.

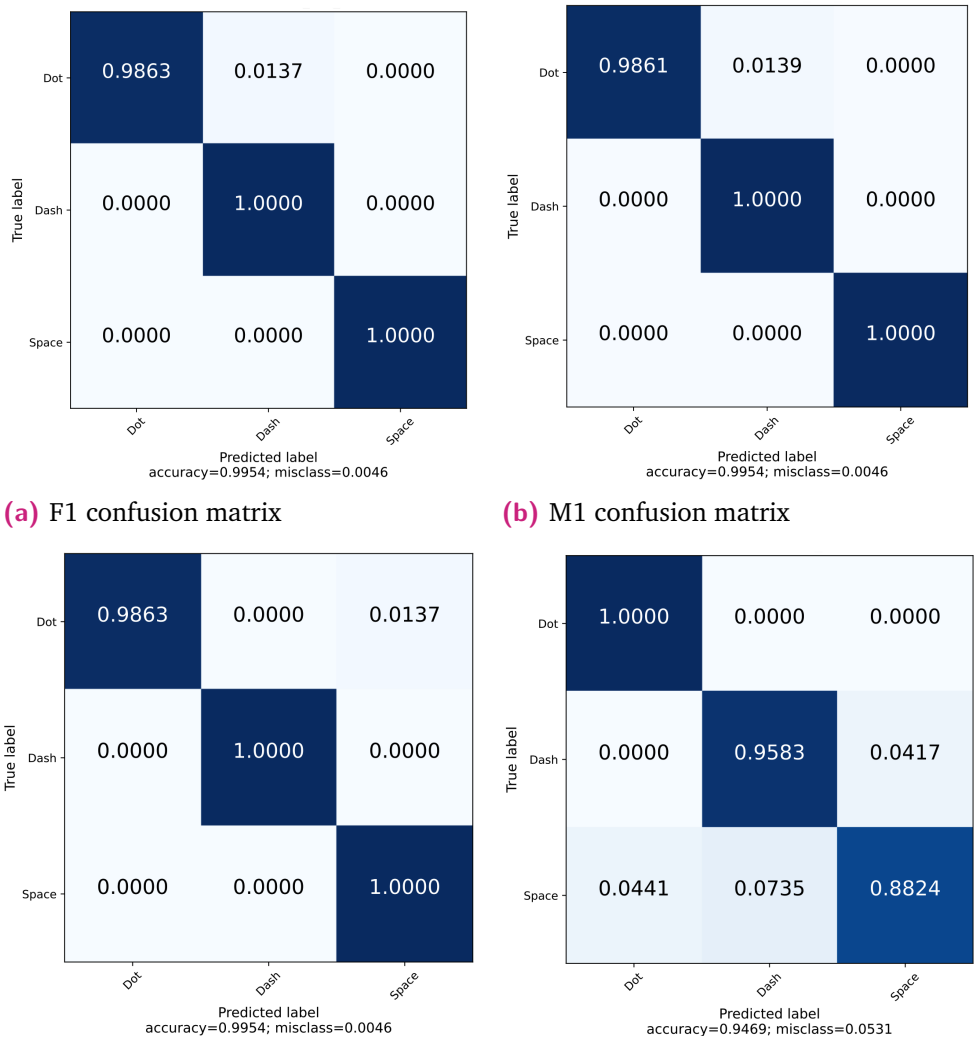
The confusion matrices in Fig 5.7 summarize the number of instances correctly and mistakenly classified by the learning model. In the first session, the model predicted 1.4% of the dot samples as dashed compared to the second session, where the misclassification rate is 11% between dashed and space in M2 dataset and about 5% in F2.

5.4 Discussion

5.4.1 User Diversity Robustness

To evaluate the system's robustness in handling user diversity, we conducted comprehensive evaluations using the F1 dataset as the training dataset and the M1 dataset for testing. In addition, we explored the impact of different base signals on the classifier performance to identify the most suitable signal for enhancing user diversity robustness. The results of these evaluations are summarized in Table 5.1, which presents the performance of various base signals utilized in the evaluation process. The evaluated signals include raw amplitude (RAW AMP), weighted moving average followed by PCA and STFT (WMA PCA STFT) for dimensionality reduction of the CSI waveforms, raw amplitude followed by STFT (RAW STFT), and weighted moving average followed by STFT (WMA STFT). These signals were analyzed to determine their effectiveness in achieving robustness in handling user diversity.

the 5th raw of Table 5.1 shows that the combination of the 3rd and 4th links achieved the highest accuracy when using the raw spectrogram as the base signal. This accuracy was slightly higher than the filtered spectrogram, as the movement of the user holding the frame affected the amplitudes of the



(a) F1 confusion matrix

(b) M1 confusion matrix

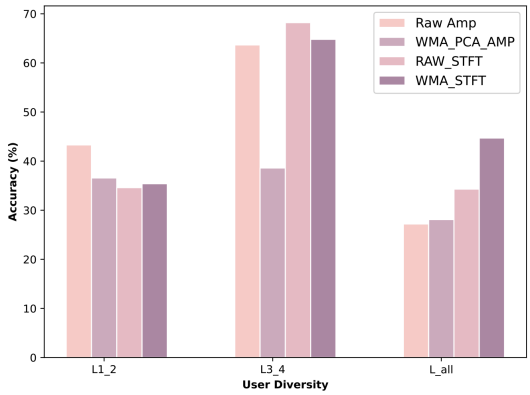
(c) F2 confusion matrix

(d) M2 confusion matrix

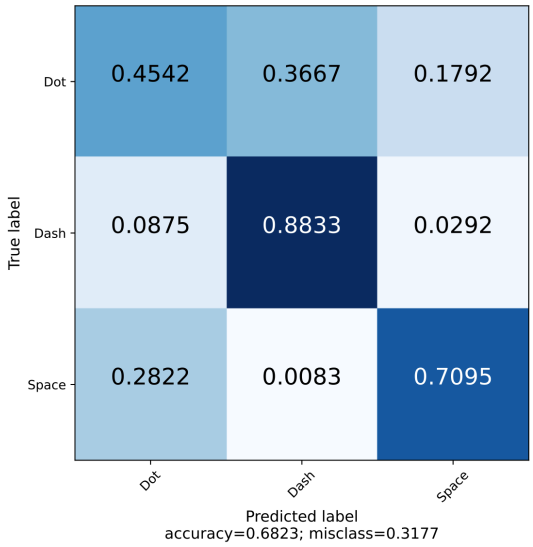
Figure 5.7: The Confusion Matrix of Different Users

1st and 2nd links. We also examined the impact of dimensionality reduction using PCA, which resulted in the lowest accuracy for links 3 and 4. This is because PCA eliminates correlated variables, potentially leading to the loss of informative data.

Moreover, the results highlight the significance of using the spectrogram as a base signal for robustness in handling user diversity. Different movements generate distinct frequencies in the frequency domain, and the spectrogram captures this information. Figure 7a illustrates the accuracy based on the leave-one-user-out validation approach. The confusion matrix for the integra-



(a) Acc. (User Diversity)



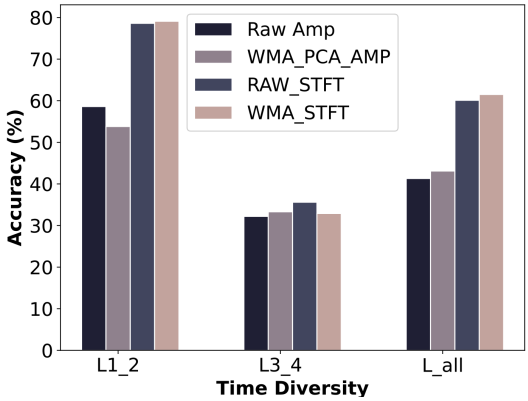
(b) CM (User Diversity)

Figure 5.8: The Accuracy and Confusion Matrix of User Diversity

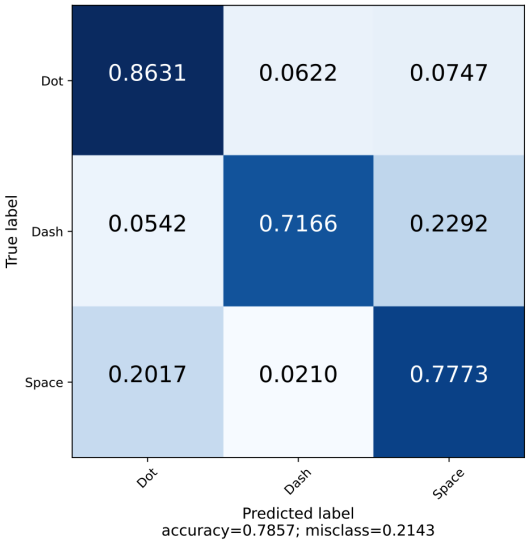
tion of 3rd and 4th links using the raw spectrogram is depicted in Fig. 5.8. It reveals that the highest rate of misclassification occurs between the dot and space samples, with over half of the dot samples being predicted as space, 30% as a dash symbol, and only 12% correctly classified.

5.4.2 Time Diversity Robustness

To investigate the robustness of time diversity, we conducted experiments by training the model using the dataset collected in the morning and subsequently testing it on the evening dataset from the same user. The last raw



(a) Acc. (Time Diversity)



(b) CM (Time Diversity)

Figure 5.9: The Accuracy and Confusion Matrix of Time Diversity

in Table 5.1 presents the results, indicating that the integration of the first and second links achieved the highest accuracy of 79.1% by employing the weighted moving average amplitude followed by STFT. This high accuracy can be attributed to the fact that the first and second links capture both the user's head motion and the movement of the person behind them. The inception model effectively extracts meaningful features from these movements using the spectrogram, which maps different speeds to distinct frequencies and translates them into unique patterns.

Fig. 5.9(a) illustrates the accuracy summary, highlighting the performance of the system in terms of time diversity. However, the confusion matrix depicted

in Fig. 5.9(b) reveals some inaccuracies in the model's predictions. Specifically, approximately 23% of the dash samples were incorrectly classified as space, and around 20% of the space samples were misclassified as a dot.

5.5 Summary

In this chapter, we present Wi-Nod, a novel head nodding recognition system that utilizes Wi-Fi CSI to facilitate communication between quadriplegia patients and others, potentially serving as a foundation for a Morse code system operated by head motions. The Wi-Nod system is described in detail, starting with the data collection phase using compact and cost-effective ESP32 nodes. This is followed by a data preprocessing module that incorporates data segmentation, concatenation, amplitude extraction, outliers removal filtering, and frequency domain transformation based on the STFT. The processed data is then fed into an inception model for the classification task. To evaluate the performance of the Wi-Nod system, we collected four distinct datasets, involving two different users and conducted during separate time sessions, all within a multi-human context environment. The evaluation results revealed that the system achieved an impressive head motion recognition accuracy exceeding 95%. This study highlights the potential of utilizing Wi-Fi CSI for head motion recognition and demonstrates the efficacy of the Wi-Nod system in accurately recognizing and classifying head nodding gestures. The findings contribute to the development of assistive communication technologies for quadriplegia patients, enabling improved interaction and communication capabilities.

6

HeMoFi4Q: Morse Communication Based on Wi-Fi and Head Motion for Quadriplegia With Environmental Robustness

6.1 Introduction

The primary objective of our proposed framework is to enhance our previous research efforts by integrating Morse symbols based on head motions to generate the complete set of 26 alphabet letters. In this chapter, we commence by introducing a novel sign language approach. Additionally, we outline the pipeline of our proposed system and elaborate on the strategies employed to address the challenge of environmental robustness. Fig. 6.1 provides an overview of the HeMoFi4Q architecture.

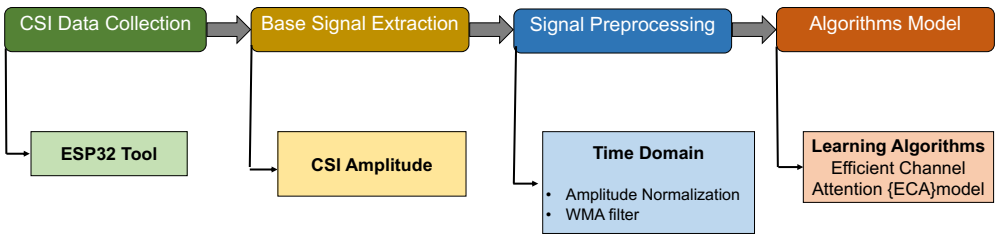


Figure 6.1: General architecture for HeMoFi4Q system

6.1.1 Background

Numerous studies have been undertaken to leverage Wi-Fi CSI for the identification and detection of human activities. These studies commonly involve either manual feature extraction from the CSI data coupled with machine learning models, or the adoption of deep learning techniques for automated feature extraction. Additionally, recent investigations have focused on developing methodologies to recognize human activities in diverse environmental settings through the utilization of Wi-Fi signals.

Bahadori et al. introduced a single-user environment human activity recognition system called ReWis [76]. The impact of employing multiple receivers and multiple antennas per receiver was explored to enhance the accuracy of the system. ReWis reduces dimensionality by leveraging time diversity through SVD and captures the correlation among subcarriers by estimating Pearson correlation coefficients. During the training phase, the CSI waveforms from the source environment are fed into a CNN model to extract representative features, along with five samples of each of the four activities from the unseen/target environment. This approach aims to bridge the gap between the seen and unseen environments. Additionally, the ProtoNet model is employed to investigate the similarity between the two different domains and address the domain independence issue.

Shi et al. introduced HAR [77] focusing on enhancing the quality of the Wi-Fi [**CSI**] waveforms through the utilization of Conjugate Multiplication (CM). The CM technique aims to mitigate phase randomness by identifying the CSI waveform with the highest quality and designating it as a reference. Specifically, the reference CSI vector is determined based on the maximum ratio of the mean of the CSI amplitude to the standard deviation across the subcarriers. Subsequently, CM is computed between the reference vector and all transmitter-receiver pairs. To reduce dimensionality, PCA is employed. To capture activity-related features, the authors apply the FFT to obtain the spectrogram of the filtered CSI. This spectrogram is then fed into a CNN-LSTM network for the embedding task. Additionally, the authors introduce the MatNet network, which aims to maximize the cosine similarity between the source environments to extract representative features. Notably, the proposed system achieves an average recognition accuracy exceeding 74%.

Francesca et al. introduced a single independent environment HAR system called SHARP [78], which utilizes Wi-Fi CSI. The authors propose a novel phase sanitization method, leveraging the strongest path, to eliminate phase offsets from the raw CSI signals. Subsequently, the extracted Doppler trace is obtained from the normalized amplitude and filtered phase. For the classification task, an Inception model is employed. The data collection for the SHARP system is conducted in three distinct environments using the Nexmon tool. Notably, even in the worst-case scenario, the average accuracy achieved by SHARP is approximately 95%.

However, despite the valuable contributions made by previous studies in addressing the challenge of environmental diversity, it is important to note that these approaches are predominantly designed for single-user environments. This limitation restricts the feasibility and practical deployment of passive Wi-Fi sensing in real-world settings. Herein, we have utilized Wi-Fi CSI waveforms to passively track head motions and extract gesture signatures for each character. The data collection process involved the utilization of a real wheelchair and an ESP32 microcontroller. Drawing inspiration from few-shot learning algorithms, we have addressed the location robustness issue in multi-human context environments by combining the seen environment with a few samples from the unseen environment. In order to achieve domain independence, we have conducted a thorough investigation of the impact of amplitude and phase features on improving recognition accuracy using the smallest number of samples from the target source. The classification results highlight the effectiveness of utilizing a diagonal links configuration and smoothing the time-domain CSI amplitude variations, which yield the best performance in diverse domain-independent scenarios. Fig. 6.2 summarizes the objective of our proposed head motions for quadriplegia.

6.1.2 Research Contributions and Questions

The main contributions of this study are as follows:

1. We propose a novel communication method that utilizes Morse code generated through head motions.

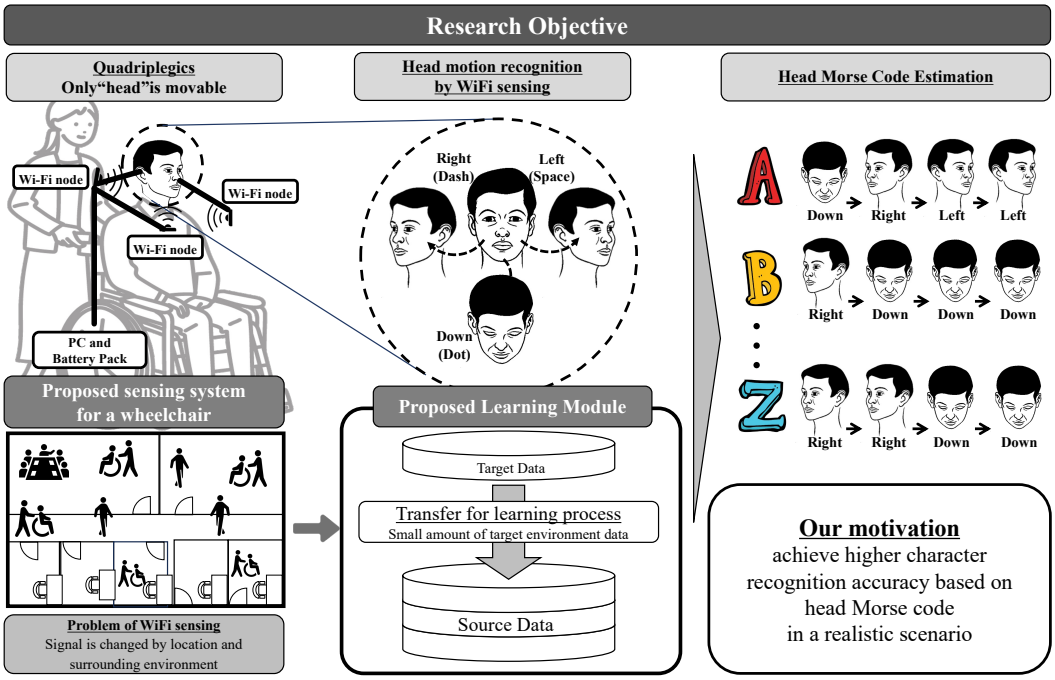


Figure 6.2: Conceptual diagram of the proposed Morse code based on head-motion framework. The system consists of three modules: data collection of CSI reading of the 26 English alphabets, the noise removal techniques, learning module, and the classification module

2. We investigate the impact of location diversity in a multi-context environment, where patients are surrounded by others. By considering this practical scenario, our research aligns closely with real-world situations and enhances the feasibility of the proposed communication method.

3. Inspired by few-shot learning algorithms, we introduce a technique to improve system performance. In the learning phase, we merge randomly selected small samples from the target environment, enabling the classifier to effectively extract the unique signature associated with each alphabet.

Our research addresses the following research questions:

Q1 What are the effects of location diversity in a multi-context environment on the practicality and performance of the proposed communication method?

Q2 How does the proposed communication method perform in different environments and locations? Does the performance of the distribution-adapted model follow the rank of distribution identification results?

Q3 What are the comparative effects of different base signals, link configurations, and state-of-the-art classifiers on system performance?

6.2 Methodology

This research introduces a novel head motion system consisting of three key modules: data collection, data preprocessing, and feature extraction and classification. Fig. 6.3 illustrates the overall architecture of the system. Notably, this study marks the first time that a comprehensive Wi-Fi CSI dataset has been collected using a cost-effective and low-power ESP32 microcontroller. This device holds promise for CSI sensing in the context of the IoT due to its standalone capabilities. Following data collection, signal processing techniques were applied, including parsing, CSI segmentation, interpolation, and filtering. These steps aimed to enhance the quality and usability of the collected data. Lastly, an extract, classify, and analyze Efficient Channel

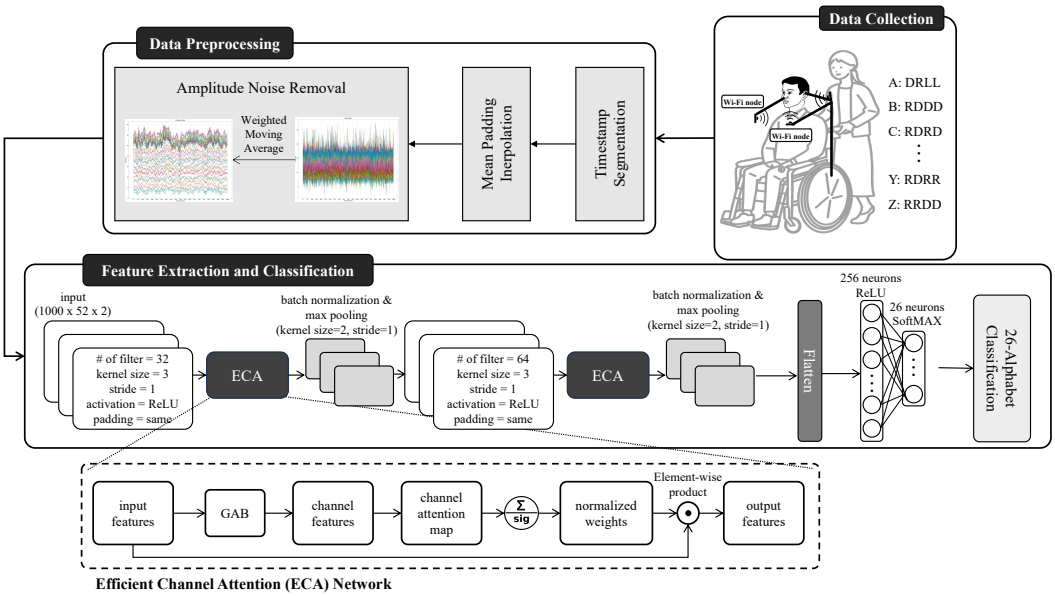


Figure 6.3: Overview architecture of HeMoFi4Q system

Attention (ECA) model was constructed to extract CSI patterns corresponding to specific head motion characteristics and perform the classification task.

6.2.1 Data Collection

To gather the CSI waveforms, three pairs of ESP32 units were affixed to a real wheelchair. Data collection occurred in two distinct environments: a single-user environment with only the patient present and a multi-human context environment in three different locations. Each environment was recorded in different locations. The Morse head motion for each character is depicted in Table 6.1, presenting the corresponding movements denoted by D, R, and L representing down, right, and left motions respectively. These motions serve as the foundational blocks for our sign language system.

Table 6.1: HeMoFi4Q Code: D-down motion, R-right motion, L-left motion

Char	Code	Char	Code	Char	Code	Char	Code	Char	Code
A	D-R-L-L	G	R-R-D-L	M	R-R-L-L	S	D-D-D-L	Y	R-D-R-R
B	R-D-D-D	H	D-D-D-D	N	R-D-L-L	T	R-L-L-L	Z	R-R-D-D
C	R-D-R-D	I	D-D-L-L	O	R-R-R-L	U	D-D-R-L		
D	R-D-D-L	J	D-R-R-R	P	D-R-R-D	V	D-D-D-R		
E	D-L-L-L	K	R-D-R-L	Q	R-R-D-R	W	D-R-R-L		
F	R-R-D-L	L	D-R-D-D	R	D-R-D-L	X	R-D-D-R		

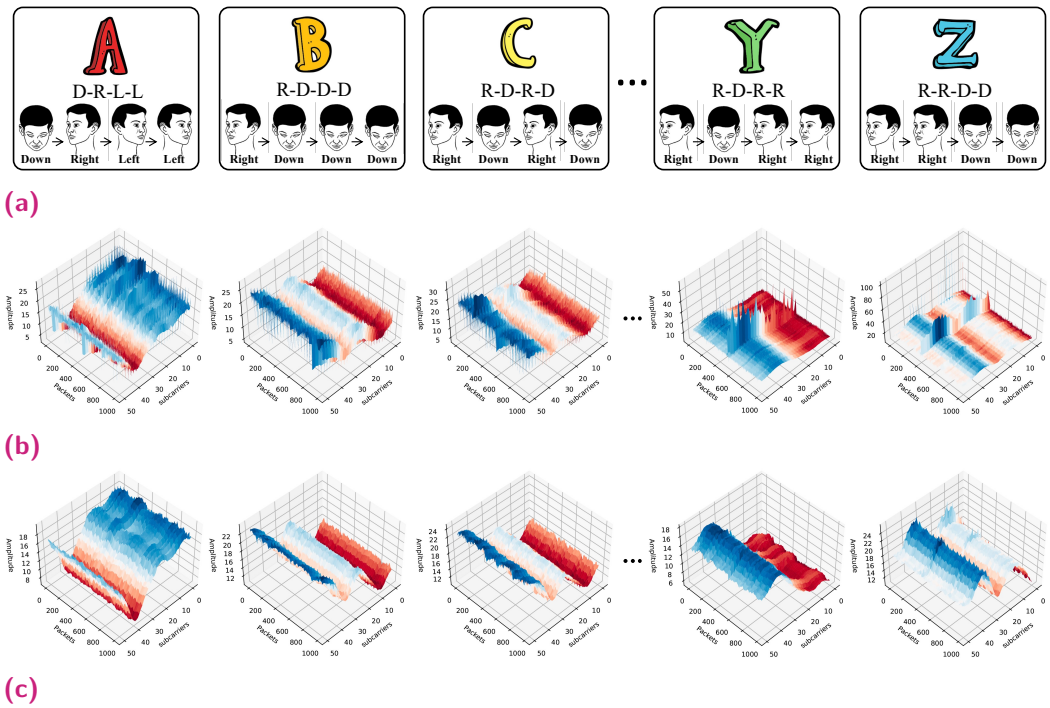


Figure 6.4: First three and last two alphabets motions corresponding to the raw and filtered CSI amplitude. **a.** Visual representation of A, B, C, . . . , Y, and Z characters. **b.** Raw CSI amplitude for each character across the first link. **c.** Filtered CSI amplitude after applying weighted moving average.

6.2.2 Signal Preprocessing

The overall process consisted of three distinct stages. Firstly, the data was divided into segments based on timestamps to facilitate further analysis. Secondly, mean values were incorporated to address the impact of packet loss in Wi-Fi CSI measurements. This step aimed to mitigate any potential distortions caused by missing data. Finally, efforts were made to eliminate noise present in the signal amplitude, enhancing the accuracy and reliability of the measurements. Additional detailed information regarding the segmentation, padding, and filtering procedures can be found in the subsequent descriptions.

Timestamp segmentation

The objective of signal segmentation is to partition the Wi-Fi Channel State Information (CSI) measurements of each link based on their respective timestamps. This crucial process aims to combine the signals from each link, enabling the creation of unique patterns that correspond to the head movements of individual users. Consequently, these distinct patterns can be mapped to their corresponding signatures, facilitating the identification and differentiation of various head movements made by each user. This process is essential for accurate and reliable communication through the system.

Mean padding

The technique of mean padding is commonly employed to handle the problem of packet loss in Wi-Fi CSI measurements, while ensuring the preservation of the remaining packet distribution. Instead of discarding the missing packets, mean padding involves estimating their values by considering the neighboring data points within the time series. By utilizing the average value of the adjacent data points to fill the gaps, mean padding assists in maintaining the distribution of the packets and the overall continuity of the time series. This approach effectively mitigates the impact of packet loss and facilitates the analysis of the CSI measurements.

Amplitude noise removal

This section outlines the noise removal technique used to smooth the CSI amplitude. It is worth mentioning that we investigated three different features to evaluate the system's performance. These features include the filtered amplitude, time-frequency feature, and the combination of the filtered amplitude and the calibrated phase. However, the variations of the filtered time-domain CSI amplitudes outperform other features in the recognition accuracy. They reveal distinct signatures for different alphabets.

Since it is not reliable to use raw amplitude directly as a feature selection due to the signal interference and environmental changes noise. Therefore, we

adopt a weighted moving average filter (WMA) [91] to remove the outliers and smooth the CSI amplitude waveforms as shown in Eq.6.1

$$\hat{A}_{\text{t,i}} = \frac{m \times A_{\text{t,i}} + (m-1) \times A_{\text{t-1,i}} + ... + 1 \times A_{\text{t-m-1,i}}}{m + (m-1) + ... + 1} \tag{6.1}$$

Where $\hat{A}_{\text{t,i}}$ and $\mathbf{A}_{\text{t,i}}$ are the filtered and raw amplitude corresponding to the subcarrier i at time t, and m that is empirically set to 30 in this paper. The illustration in Fig. 6.4(a) showcases the visual head motions of A, B, C, Y, and Z alphabets, while corresponding raw and filtered CSI amplitudes are shown in Fig. 6.4(b) and Fig. 6.4(c), respectively. Further details of the experiment setup are described in the following sections.

6.2.3 Feature Extraction and Classifier Phase

The main issue with the current DCNN is that more layers added to it will improve the performance, but will make the model more complex. Additionally, the current attention technique non-identically weights the extracted feature depending on how crucial it is to the classification task. Therefore, information is lost as a result of the unnecessary and ineffective dependencies across several channels.

To get over these restrictions, the ECA module [92] is presented to aggregate the local inter-channel information. During the learning process, ECA assigns the weights of each channel based on its significance in the Channel Attention Map (CAM). Then, to emphasize the distinct patterns and suppress the uncorrelated data, the input features of each channel are multiplied by the corresponding CAM weights.

The following is a description of how Fig. 6.3 depicts the ECA network's process. First of all, the Global Average Pooling (GAP) layer [93] aggregates the input features f_a and can be expressed as in Equation 6.2.

$$g(X) = \frac{1}{WH} \sum_{i=1,j=1}^{W,H} X_i j \tag{6.2}$$

where $\mathbf{x} \in \mathbb{R}^{W \times H \times C}$ is the convolution neural network output. W, H, and C denote width, height, and channel dimension (i.e., number of filters).

Unlike the fully connected layer, GAP is used to decrease the complexity of the modal by decreasing the number of parameters. Furthermore, it solves the overfitting problem caused by the fully connected layer. It is worth mentioning that there are no parameters or weights to be optimized for GAP.

after that, the fast one-dimensional convolution layer *1d* with *k* kernel size f_{1d}^{k}, $f^{1d} \in \mathbb{R}^{C}$, is applied to each input channel to generate the weights of CAM f_b which can be mathematically calculated as in Equation 6.3.

$$f_b = \sigma(W \times (f_{1d}^{k})) \tag{6.3}$$

, where $W \in \mathbb{R}^{C \times C}$ is the weight matrix and C represents the channel dimension. The *k* establishes how many neighbors take part in the attention prediction via each channel. In the proposed system, *k* value is equal to 3. After that, the sigmoid function,σ, normalizes the attention weights $f_b^{'}$. Finally, the significant input features are generated by multiplying the input features with the channel weights $f_b^{'}$. In ECA network, the adaptive selection of the kernel size is an exponential function that depends on the number of channels *C*.

This model is a CNN with an ECA module. The input to the model is a tensor of shape (1000,52,2). The model has two convolutional layers, each followed by an ECA module, batch normalization, and max-pooling layers. The output of the second max-pooling layer is flattened and fed into a dense layer with 256 units, followed by a dropout layer with a dropout rate of 0.25. The final output layer contains 26 units representing the number of characters with softmax activation.

The ECA module is a feature recalibration mechanism that enhances the performance of CNNs by recalibrating the feature maps. It applies a convolutional operation with a small kernel size to squeeze the feature maps into a single channel and then applies a sigmoid activation function to obtain attention maps. The attention maps are multiplied with the original feature

maps to generate the scaled feature maps, which are then passed to the next layer.

6.3 Performance Evaluation

6.3.1 Experiment Setup

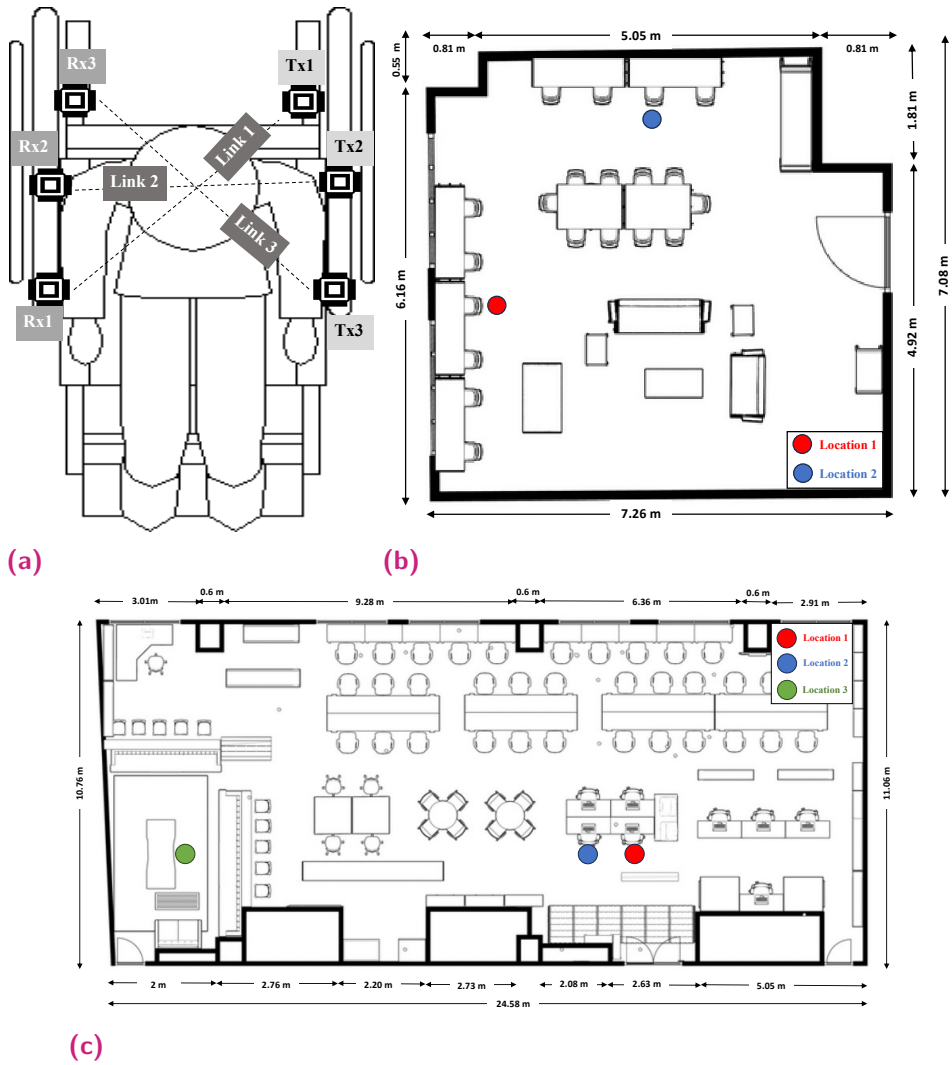


Figure 6.5: Data collection setup based on ESP32 microcontroller. **a.** Top view of wheelchair setup used in both environments. **b.** Single-user environment layout. **c.** Multi_human context environment layouts.

In our experimental setup, we considered two distinct environments to investigate the capabilities of ESP32 modules as both Tx and Rx. The research employed a total of six ESP32 microcontrollers operating in the 2.4 GHz frequency band and utilizing the IEEE 802.11n protocol for CSI data collection. Fig. 6.5(a) illustrates the configuration of the setup, which involved three ESP32 modules functioning as transmitters. These transmitters were connected to a mini-PC running the Ubuntu 16.04 operating system. The remaining three ESP32 modules served as receivers. To establish the link between transmitters and receivers, specific distances were maintained. For link 1, the distance between the transmitter and its corresponding receiver was 84.3 cm. Similarly, for link 2, the distance was 52.3 cm, and for link 3, it was 84.3 cm

In both environments, *Env1* and *Env2*, the user's head movements followed a predefined pattern outlined in Table 6.1. Each symbol, along with the end motion, lasted for two seconds, resulting in a total of 10 seconds per character. Data collection was organized into separate files, with each character's data spanning a duration of 10 minutes. The transmission of Wi-Fi frames from the three Tx to the Rx occurred concurrently at a rate of 100 Hz.

Env1 represented a single-environment scenario where only the user was present. Data was collected from two different locations within this environment, as depicted in Fig. 6.5(b). On the other hand, *Env2* represented a multi-human context environment, as illustrated in Fig. 6.5(c). Each character in the dataset had dimensions of $1000 \times 52 \times n$, where 1000 denoted the number of packets, 52 represented the number of subcarriers, and the value of n depended on the number of links utilized. The investigation focused on exploring the impact of different link configurations, which will be discussed later in 6.3.3.

For a single-link configuration, the value of n was set to 1. When utilizing two links, n was set to 2, and when employing all available links, n was set to 3. The dataset collected for this research was balanced, ensuring an equal number of samples were obtained from each location. A total of five locations were considered, with 100 instances of each character collected from each location. This resulted in a total of 2600 samples for each alphabet character at each location. Cumulatively, across all locations, a total of $13,000$ samples were gathered. The focus of the experiments was on 26 alphabet characters

Table 6.2: Training and Testing Size for HeMoFi4Q Performance Evaluation

Samples	Env1		Env2		Env1 → Env2		Env2 → Env1	
Training samples	Loc1 + 2% Loc2	2652	Loc1 + Loc2 + 2% Loc3	5252	All_locs_Env1 + 2% Loc3_Env2	5356	All_locs_Env2 +2% Env1	7904
Testing samples	98% Loc2	2548	98% Loc3	2548	98% Loc3_Env2	2548	98% Env1	5096

as the number of classes. This selection was based on the functionality of the HeMoFi4Q system, which extracts Morse code signatures from head motion and maps them to the corresponding characters.

6.3.2 Evaluation metrics of classification models

In order to assess the effectiveness of HeMoFi4Q in various environments, we conducted an evaluation using CSI data. The implementation of HeMoFi4Q was designed to be applicable in real-world scenarios. Taking inspiration from few-shot learning algorithms, we introduced a strategy to address the challenge of location diversity robustness. Specifically, a small portion (denoted as x) from the unseen/test environment was incorporated into the seen/train dataset. This fusion process is depicted in Fig. 6.2.

To evaluate the performance of the proposed system, we employed accuracy and F1-score metrics. These metrics served as quantitative measures to assess the effectiveness and robustness of HeMoFi4Q in different environments.

6.3.3 Results

The major objective of this study is to introduce a passive communication method between quadriplegics and others based on head motions detected by Wi-Fi signals and DL algorithms.

To evaluate the robustness of location diversity, the source dataset is combined with 2% of the target dataset for training the model, and the remaining 98% is then used for the testing phase. Table 6.2 provides information about the sample sizes used in different experimental settings. In particular, we compared the performance of ECA learning model with two state-of-art classifiers, namely CNN and ResNet across different locations of the same

Table 6.3: Location diversity comparative results of different base signals at Env1

Link Conf	ECA			CNN			ResNet		
	Amp	DWT_Amp	Amp+Phase	Amp	DWT_Amp	Amp+Phase	Amp	DWT_Amp	Amp+Phase
Link1	79.3	77.7	37.3	84.5	73.7	33.4	81	74.8	34.5
Link2	86.8	81	62	88	81.7	69	86.7	80	65
Link3	82.7	77.2	61.4	88.4	72.7	60.2	76.2	74.7	61
Link1_2	92.4	93.8	78.5	93.3	90.2	72.3	88.3	**91.7**	73.4
Link1_3	**94.6**	90.2	79.3	92.8	88.9	74.7	88.4	87.2	72
Link2_3	93.3	94	79.7	**93.7**	92.5	75.6	87.2	90.6	74 .3

Table 6.4: Location diversity comparative results of different base signals at Env2

Link Conf	ECA			CNN			ResNet		
	Amp	DWT_Amp	Amp+Phase	Amp	DWT_Amp	Amp+Phase	Amp	DWT_Amp	Amp+Phase
Link1	65.9	76	48.3	76.8	73.3	44	57.5	70.6	42.6
Link2	74.6	83.4	55.6	84	70.6	52	55.8	68.6	48.7
Link3	74.7	88.9	57.6	82.8	**87**	51.7	58.8	83.8	40.5
Link1_2	86.4	87.1	69.3	84.6	85.4	59.7	57.8	73.1	51.4
Link1_3	**89.3**	88.3	71	83.3	83	52	60.7	76.3	45.8
Link2_3	87.4	85.5	70.2	84.4	83.8	60.3	57.5	**85**	57.8

Table 6.5: Cross Domain results of different classifiers

	Env1 →Env2			Env2→Env1		
	ECA	CNN	Resnet	ECA	CNN	Resnet
Accuracy	**84.3**	10	37	**85.6**	85	62.7
F1-score	0.84	0.07	0.36	0.83	0.78	0.57
Training_time(sec)	19	14	13	13	10	7
Testing_time (sec)	4	3	2	2	1	1

environment as in Table 6.3 and Table 6.4. Additionally, we evaluated the performance in cross-domain environments, where the training dataset consisted of a combination of locations from one environment and 2% of the locations from a second environment. The testing phase was conducted on the remaining 98% of the second environment, as illustrated in Table 6.5.

The parameters for the two baseline models are as follows.

- CNN model: This is a simple convolutional neural network (CNN) model for image classification. The model consists of two CNN blocks, each composed of several sequential layers. The input shape is (1000, 52, 2), representing a 2D image with 1000 packets, 52 subcarriers, and 2 links used. The first block starts with a convolution layer with 32 filters of size 5x5 and a stride of 1, applying zero-padding with a padding value of 1. This is followed by a batch normalization layer for input normalization, a ReLU layer for introducing non-linearity, and an average pooling layer with a window size of 3×3 and a stride of

3. A dropout layer is employed to prevent overfitting by randomly dropping out units during training. The second block consists of two fully connected layers. The first fully connected layer has 1000 neurons with a ReLU activation function and a dropout rate of 0.5. The second fully connected layer has 26 neurons, corresponding to the number of classes, and utilizes the softmax function for classification. The model is trained using the Stochastic Gradient Descent with Momentum (SGDM) optimization algorithm, with a learning rate of 0.02 and momentum of 0.9.

- ResNet model: This is a deep neural network architecture commonly used for image classification tasks. The input shape of the model is (1000, 52, 2), where 1000 refers to the number of packets, 52 represents the subcarriers, and 2 denotes the links used. To maintain the spatial dimensions, the input is initially padded with zeros using a (3, 3) padding size. The model consists of two stages. In the first stage, a convolutional layer with 64 filters, a kernel size of (7, 7), and a stride of (2, 2) is applied to extract features. This is followed by a batch normalization layer for activation normalization, a ReLU activation layer to introduce nonlinearity, and a max pooling layer with a pool size of (3, 3) and a stride of (2, 2) for downsampling. The second stage includes a convolutional block with three identity blocks. Each identity block comprises three convolutional layers with filter sizes of [16, 32, 64], a kernel size of 3, and a stride of 1. The first identity block has a different shape due to the change in filter sizes. An average pooling layer with a pool size of (2, 2) is applied to further downsampling the data. The output is then flattened into a 1D vector and fed into a fully connected layer with 26 neurons, which corresponds to the number of classes in the classification task. The softmax activation function is used to generate predicted probabilities for classification. The model is trained using the Adam optimizer with a learning rate of 0.001, beta_1 of 0.9, beta_2 of 0.999, and epsilon of 1e-08.

ECA classifier outperforms other algorithms as shown in Fig. 6.6 and Fig. 6.7 by achieving highest recognition accuracy and f1-score of location diversity evaluation for *Env1* and *Env2*, training on *Env1* with 2% of *Env2* and testing on 98% of *Env2* (*S1*), and training on *Env2* with 2% of *Env1* and testing

on 98% of *Env1* (*S2*), respectively. CNN algorithm achieves the lowest performance when using the single-user environment for the learning phase because of overfitting. CNN and ResNet cannot capture the unique patterns for each character compared to ECA which uses an attention layer to highlight the signatures of the characters from the single environment and the small amount of the multi-human sensing environment.

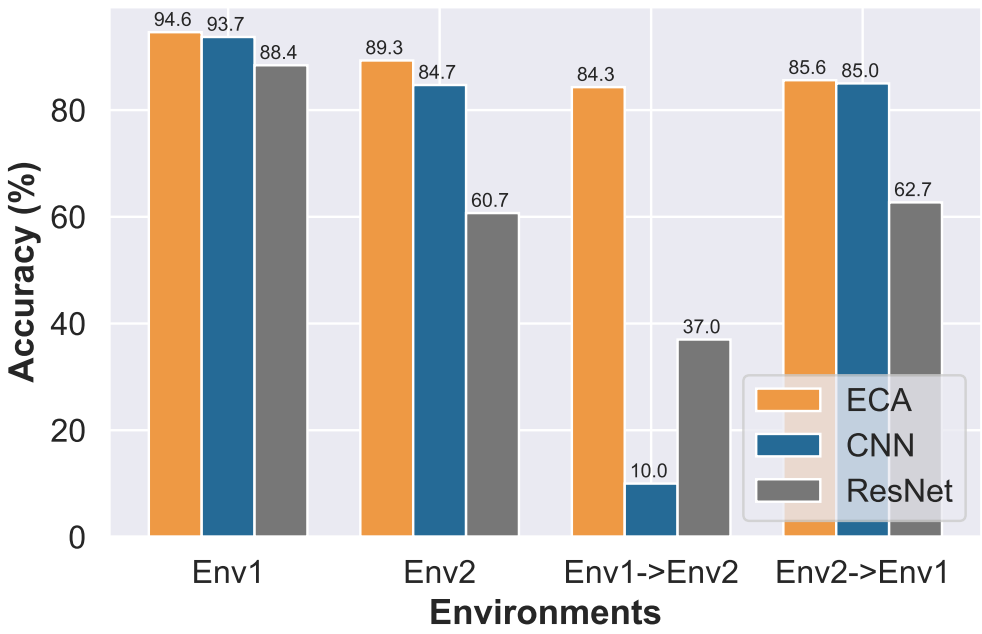


Figure 6.6: Overall accuracy of different learning models.

Single user environment data

The system was trained in a single-user environment where only the participant existed, *Env1*, using the entire dataset collected from location 1 and an additional 2% of data from location 2, which equates to two samples for each character collected in the second location. This approach was taken to improve the system's performance and location diversity robustness. Table 6.3 displays the evaluation outcomes of three distinct image classification algorithms, namely Efficient Channel Attention (ECA), Convolutional Neural Network (CNN), and Residual Network (ResNet). The results were

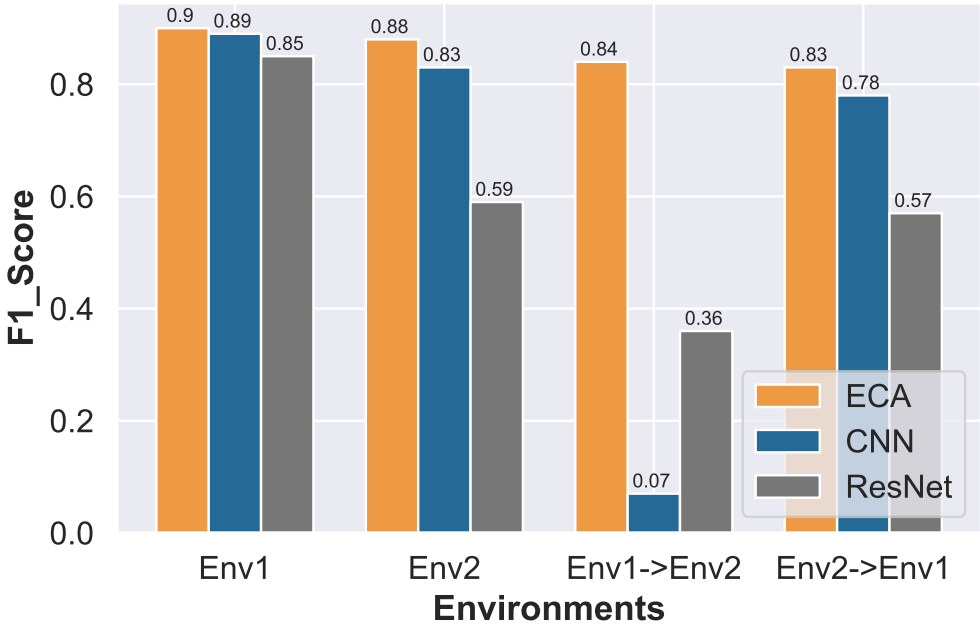


Figure 6.7: F1-score of different learning models.

obtained by assessing the accuracy of each classifier, and they are presented for comparison purposes.

It is observed from Table 6.3 that ECA slightly outperforms CNN on combined datasets of the diagonal links configuration (link1_3) for different base signals, while the ResNet classifier achieves the worst performance. Moreover, ECA achieves the highest recognition accuracy with 94% when using the variations of amplitude filtered by WMA filter from diagonal links datasets and transforming the filtered amplitude of the combined dataset between one of the diagonal links and the horizontal link to the wavelet domain by applying the discrete wavelet transform (DWT). The confusion matrix is introduced in Fig. 6.8. The confusion matrix shows that there are 13% and 9% misclassifications between C and D and B, respectively. Additionally, the model classified R character as T character with 16% rate and S as R character with 36% rate which is the highest misclassification rate. Finally, ECA classified W character with the same rate, 7%, as V and X.

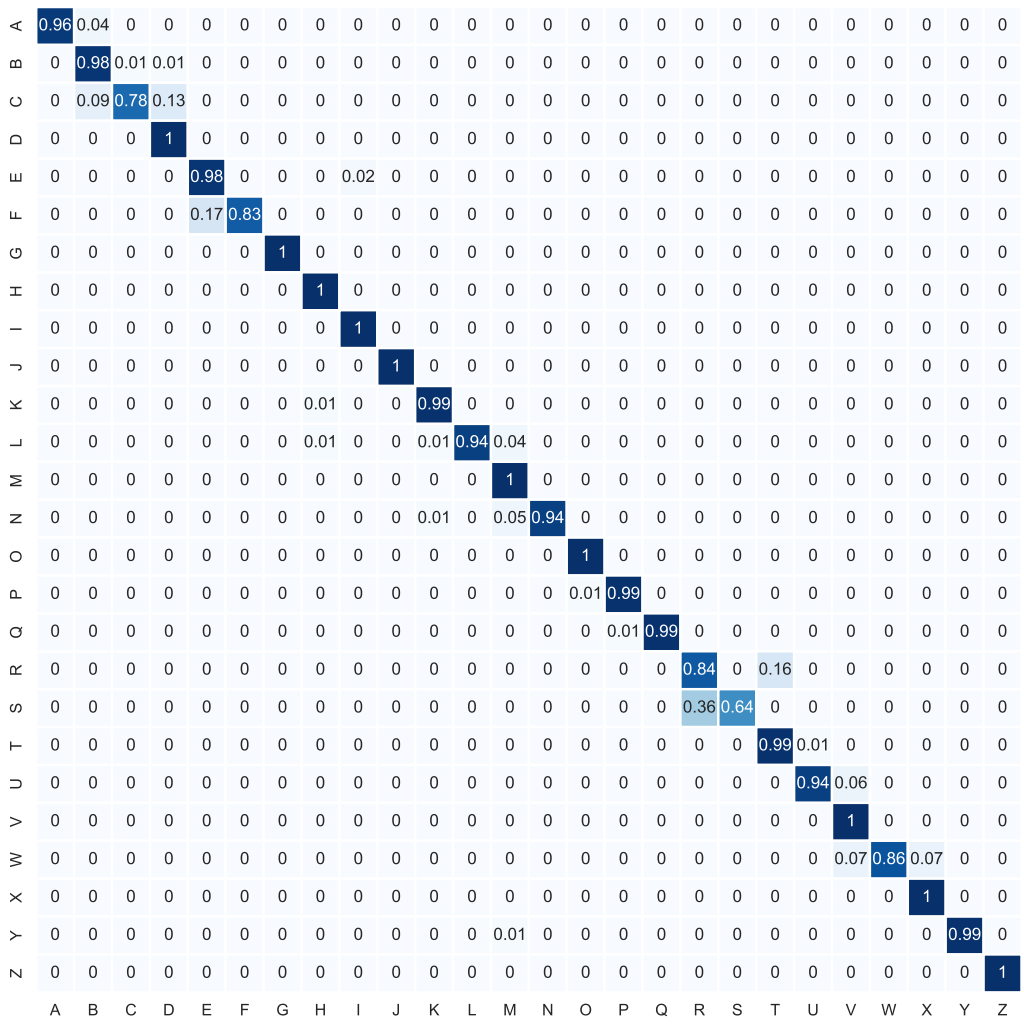


Figure 6.8: Confusion Matrix of Single User Environment

Multi-human context environment data

The classification accuracy of data collected from various locations in a multi-human context environment is presented in Table 6.4. Our study also investigated the optimal link configurations, different base signals, and three classifiers. The results demonstrated that the dataset of diagonal links outperformed other link configurations, achieving 89.3% accuracy based on the variations of filtered amplitude and using the ECA classifier. Furthermore, the ResNet classifier achieved its highest accuracy of 85% by utilizing the wavelet of the filtered amplitudes from one of the diagonal link data with the horizontal one. On the other hand, the combination of the variation of

the filtered amplitude and calibrated phase was verified as the worst base signal that could be fed to the system because the randomness of the phase increased with the number of people in the sensing environment, leading to a degradation of the system's performance.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	0.99	0	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0.01	0.48	0.48	0.02	0	0	0	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0.02	0.17	0.64	0.16	0	0	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0.01	0.97	0	0	0.03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0.98	0.01	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0.02	0.95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03	0
G	0	0	0	0	0	0	0.93	0.02	0	0	0	0	0	0	0	0	0.05	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0.96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.03	0
I	0	0	0	0	0	0	0	0	0.99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0
J	0	0	0	0	0	0	0.01	0	0	0.99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0.99	0	0	0	0	0	0	0	0.01	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0.93	0	0.07	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0.01	0.87	0.12	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0.22	0.78	0	0	0	0	0	0	0	0	0	0	0	0
O	0	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0.85	0.02	0.08	0	0	0.02	0	0	0.02	0	0	0
P	0	0	0	0	0	0	0	0.01	0.01	0	0	0	0	0	0.05	0.91	0.02	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.18	0.79	0	0.01	0.01	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.88	0.02	0	0	0	0.09	0.01	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0	0.94	0.01	0.02	0	0	0.02	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.95	0.01	0	0.01	0.03	0	0
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05	0	0.92	0	0.01	0.02	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.09	0	0.01	0.87	0	0.03	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05	0	0.05	0.79	0.11	0	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.04	0.01	0.02	0.01	0.92	0	0
Y	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.99	0
Z	0	0	0	0	0	0	0.01	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.98

Figure 6.9: Confusion Matrix of Multi-human Context Environment

Fig. 6.9 presents the confusion matrix. As it can be observed from this visual representation, the model classified the *B* character with about 48% accuracy as *C* character. Additionally, it misclassified the *C* character as 17% and 16% as *B* and *D* characters, respectively. Moreover, there is a misclassification rate between *M* and *N* with 12% and 22%, respectively.

Cross domain results

This study aimed to examine the robustness of the proposed system to cross-domain or environment diversity. This involved training the model on specific locations within one environment and merging it with a randomly selected 2% from another environment, thereby investigating the effectiveness and robustness of the system. Two worst-case scenarios were considered. The first scenario (S1) is training on the single-user dataset with small samples from the multi-human one and testing on the multi-human environment (Env1→Env2). The second one (S2) uses a multi-user environment dataset for learning with small samples from the single-user environment and for the interference stage using the unseen samples of the single-user environment (Env2→Env1).

The classification results of different classifiers and confusion matrix are given in Table 6.5 and Fig. 6.10. Table 6.5 shows the overall performance of different classifiers. ResNet gives the lowest accuracy in both scenarios while ECA model gives the best performance among the other classifiers in terms of accuracy and F1-score metrics. However, ECA takes a little bit more time consumption in the learning and inference stages than others.

To evaluate the performance of *S1*, the training dataset consisted of the single-user environment (*Env1*) dataset merged with 2% of the third location dataset from the multi-human context environment (*Env2*), while the testing dataset comprised the remaining 98% of the data. The ECA algorithm outperformed the other algorithms, achieving 84.3% accuracy and 0.84 F1-Score with slightly more time consumption than CNN. On the other hand, ResNet achieved the shortest time consumption compared to CNN and ECA models but yielded an unacceptable classification accuracy as 37%. Interestingly, CNN achieved poor accuracy due to overfitting, where the model could not distinguish the signatures for each alphabet of the multi-human environment from the small target samples used in the learning phase. the model achieves 70% and higher for each character except the *C and T* characters, it achieves 53% and 55% , respectively. ECA wrongly classified the *C* character as *B* and *F* characters with accuracy of 18% and 24%, respectively. For *T* character, it misclassified it as *S* character with 18% accuracy. The classification results for *S2* give better accuracy than S1 results

because the classifiers are able to extract the most significant features for each character. ECA achieves the best accuracy in 85.6% which is slightly higher than CNN model performance and 0.83 F1-score. From the confusion matrix in Fig. 6.10, ECA achieves 72% and above for most of the characters. In particular, *W* character gives the worst accuracy 43% since the model wrongly classified it as *X* and *V* with 32% and 14% accuracy, respectively. Moreover, *B* character accuracy is slightly higher than *W* one 54.3% since there is a wrong classification as *C* and *A* 29% and 14%, respectively. The model classified *C* character as *B* with 33% accuracy. Furthermore, There is also misclassification between *S* as *U* and *R* with accuracy rates of 15% and 14%, respectively.

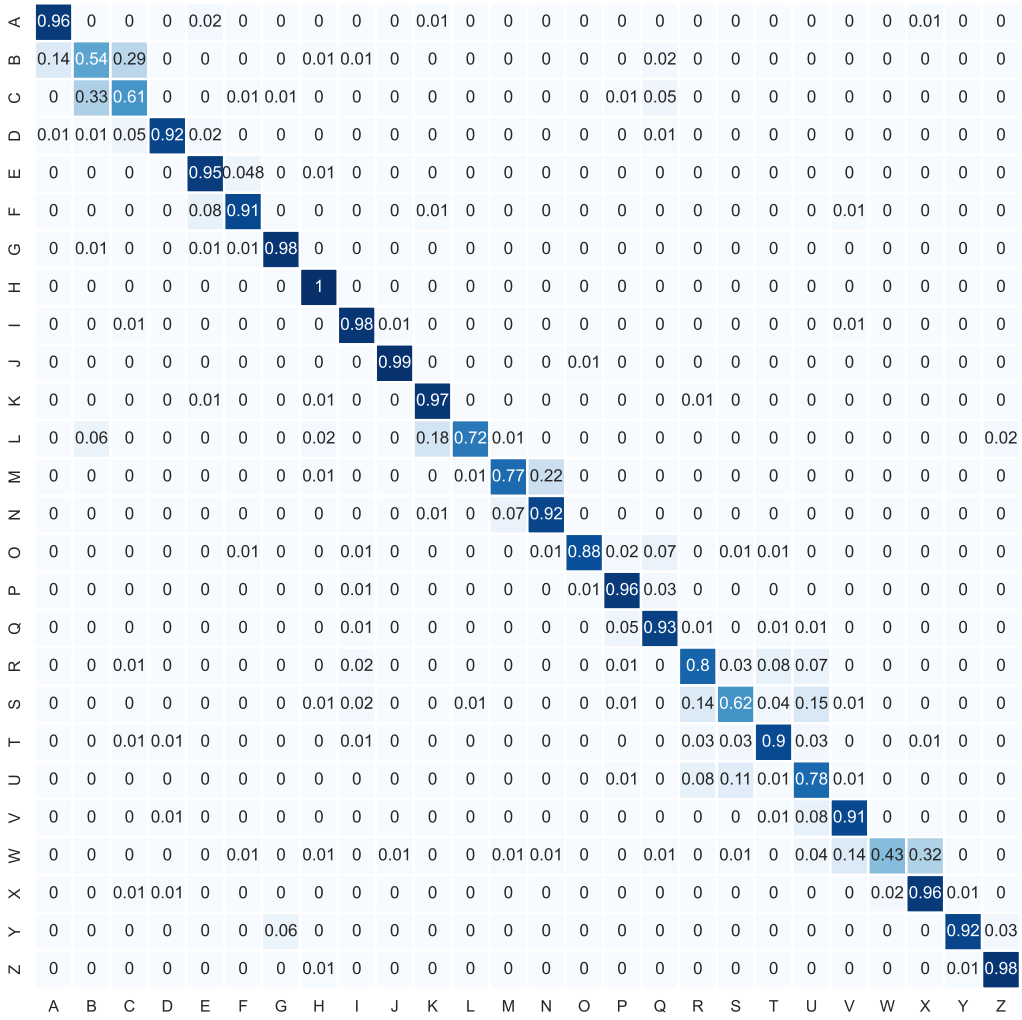


Figure 6.10: Confusion Matrix of Env2→ Env1

6.4 Discussion

In this study, a Morse code based on Wi-Fi CSI head motion detection is presented using ESP32 microcontroller. we investigated the impact of different link configurations, base signals, and state-of-art deep learning classifiers on the location diversity performance as shown in Table 6.3 and Table 6.4. It is worth mentioning that these results by using 2% amount From the target or unseen location. from these tables, it is obvious that the combination of the diagonal links' data outperforms other link configurations by extracting the variation of the CSI amplitude and filtering it using the weighted moving average algorithm outperforms other link configurations by achieving 94% and 89% recognition accuracy, for single and multi-human context environment, respectively.

6.4.1 Impact of different target amount

Our study aimed to explore the effects of altering the target amount incorporated into the training data on the enhancement of model accuracy during instances of limited data availability. Specifically, we conducted an experiment involving the testing of different target amount values within the training data and the subsequent evaluation of their impact on model performance in location diversity. Our findings contributed novel insights into the potential benefits of adjusting the target amount in the training data to improve the environmental robustness of the Wi-Fi CSI system.

Fig. 6.11 depicts the accuracy of the ECA algorithm, under various link configurations, for different amounts from the unseen location (Loc2) merged with the training dataset, which is the data gathered from Loc1, (Target Amount%) during the learning phase. According to the information presented in the figure, it is evident that the utilization of the diagonal links dataset results in the highest performance for the ECA algorithm. This configuration yielded the best results even when only two randomly selected samples were taken for each character collected in the target location. As shown in Fig. 6.12, the performance of each link alone yielded the worst recognition accuracies, ranging from 66% to 74.7% for the diagonal and horizontal links,

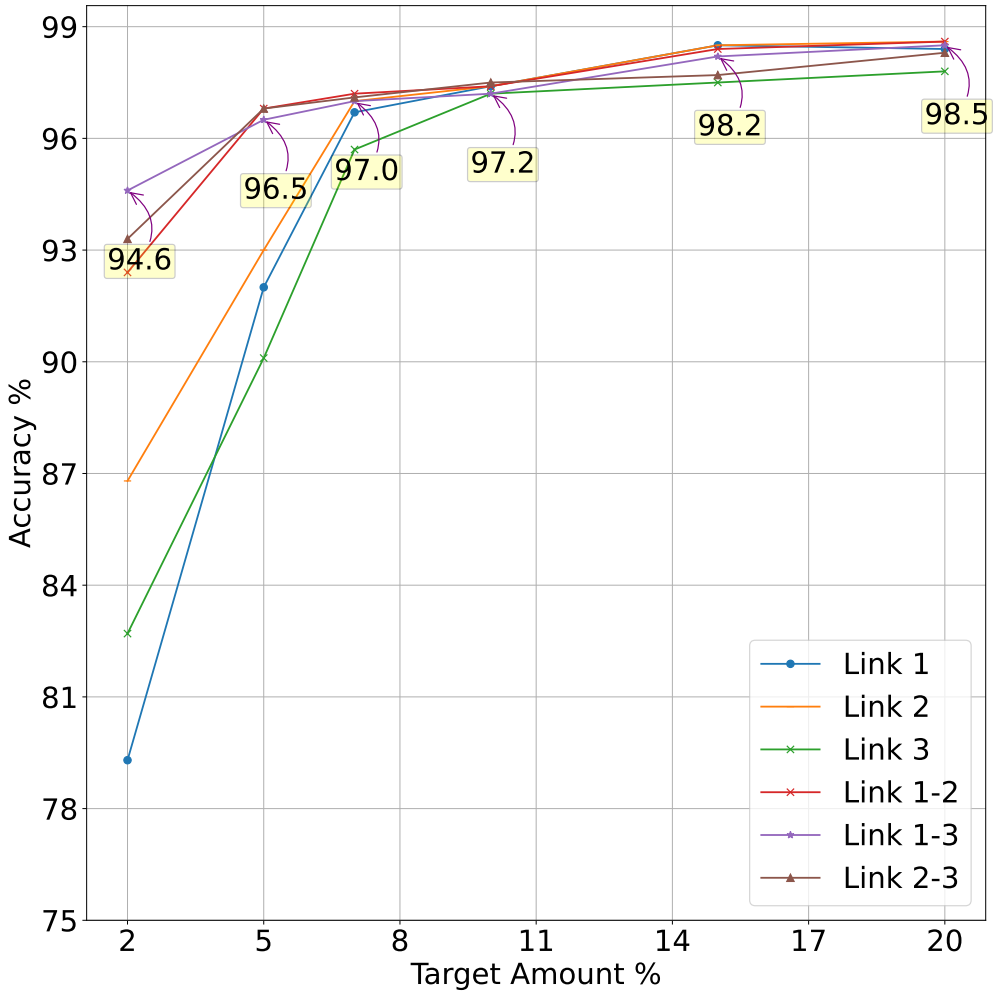


Figure 6.11: Classification accuracy using different amounts from the unseen location merged with the seen location in a single-user environment.

respectively, due to the influence of the target's surrounded movements by individuals. However, when the diagonal links were combined, the system achieved the highest performance of 89.3%. This can be attributed to the ECA classifier, which utilizes a channel attention layer to highlight the most significant features from the filtered amplitude captured by these links.

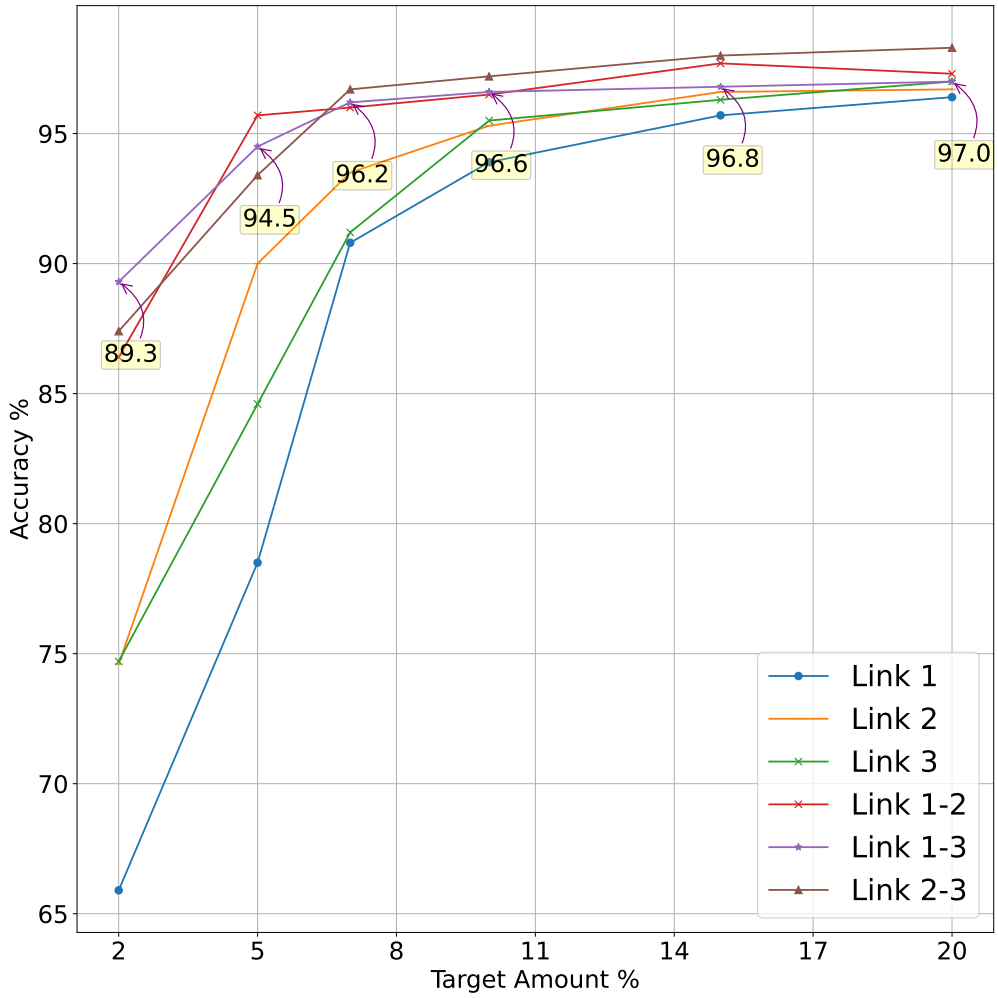


Figure 6.12: Classification accuracy for different amounts from the unseen location merged with the seen location in a multi-user environment.

6.4.2 Impact of different link configuration

Table 6.3 and Table 6.4 reveal a significant finding that the diagonal link configurations consistently achieve the highest accuracy for both single-user and multi-human context environments. Moreover, the variations of the filtered amplitude were found to be a robust and reliable signature for each character, regardless of the location diversity. In addition, the channel attention layer of the ECA algorithm was identified as a crucial component that enhances the classification task. It is noteworthy that the overall accuracy of the system in a multi-human context environment is lower compared

to a single-user environment. This is due to the fact that the presence of more individuals leads to increased interference and reflection, which ultimately degrades the system's performance. Overall, the study highlights the importance of an effective link configuration, signal processing technique, and utilizing advanced classification algorithms in achieving accurate and reliable character recognition based on passively tracking head motion via Wi-Fi CSI signals in various environments.

6.4.3 Impact of different base signals

We investigated the impact of the variations of different base signals on the classification performance as shown in Fig. 6.13. The impact of different base signals like, filtered amplitude using WMA filter, transforming the filtered amplitude to the wavelet domain by applying discrete wavelet transform technique (DWT_Amp), and combining both the filtered amplitude based on WMA and calibrated phase based on the linear transformation algorithm (Amp+Phase) on the performance are studied. As the results show in the previous tables, the combination of the CSI amplitude and phase variations degrades the performance as it achieves 79.7% and 70.2% for the first and second environments, respectively. The impact of the randomness of the Wi-Fi CSI is obviously shown in the multi-human environment because there are a lot of reflections and interference due to the existence of many dynamic subjects in the sensing environment.

6.5 Summary

This chapter presents HeMoFi4Q, a novel passive head motion detection system that utilizes Wi-Fi CSI analysis to facilitate non-verbal communication with quadriplegia patients through a newly developed sign language. To the best of our knowledge, this is the first attempt to create a sign language specifically designed for quadriplegia patients, integrating Morse code and head motions. HeMoFi4Q employs ESP32 microcontrollers to extract CSI amplitude variations, which are then processed using a weighted moving average filter and fed into the Efficient Channel Attention classifier. The

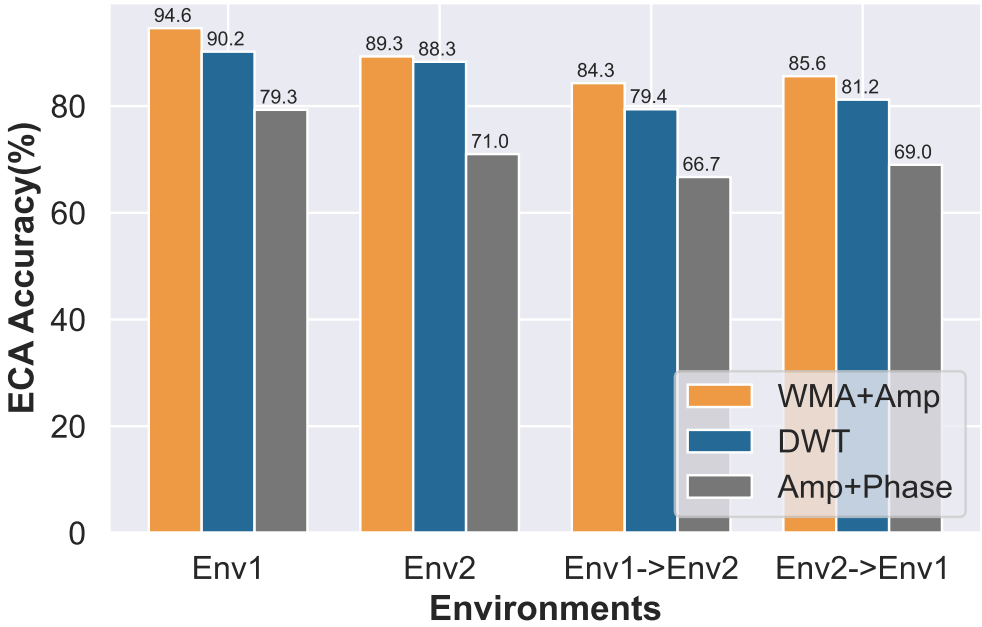


Figure 6.13: Accuracy of different base signals.

dataset used for training and evaluation was collected from two different environments, each containing multiple locations. To ensure the system's effectiveness and robustness in handling location diversity, we adopted a few-shot learning approach by randomly merging a small sample amount (2%) from the unseen environment with the training dataset during model learning. The performance of HeMoFi4Q was evaluated using various metrics, including accuracy, F1-score, and confusion matrix. The results demonstrated that HeMoFi4Q outperformed all baseline models, highlighting the significance of an effective link configuration, signal processing techniques, and advanced classification algorithms in achieving accurate and reliable character recognition through passive tracking of head motion using Wi-Fi CSI signals across diverse environments. Overall, this study emphasizes the importance of leveraging appropriate methodologies and technologies to enable effective communication for quadriplegia patients, showcasing the potential of radio frequency sensing methods in this domain.

7

Tracking On-Desk Gestures Based on Wi-Fi CSI on Low-Cost Microcontroller

7.1 Introduction

In recent years, several studies have highlighted that individuals who bottle up their emotions experience heightened levels of stress, anxiety, job dissatisfaction, emotional exhaustion, and reduced productivity. Furthermore, it is well-documented that emotional suppression can have negative consequences for both mental and physical well-being [94, 95]. Therefore, it is crucial to comprehend and interpret body gestures, a form of non-verbal communication encompassing facial expressions, gestures, and body movements, as it plays a significant role in fostering a healthier and more productive work environment. Fig. 7.1 provides the Wi-Fi CSI for on desk gesture tracking system.

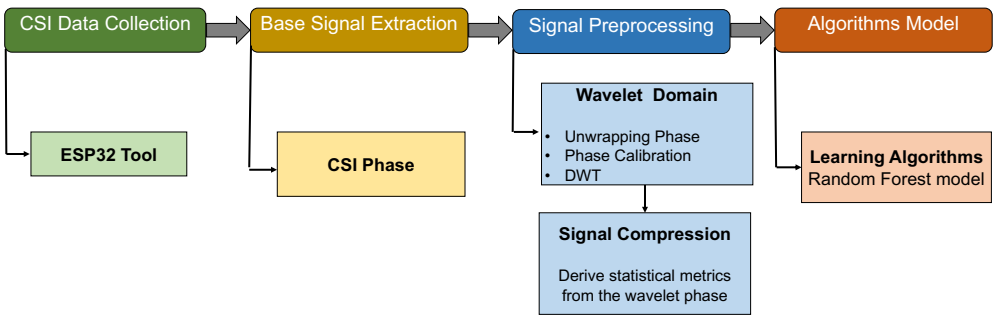


Figure 7.1: General architecture for on desk gesture tracking system

7.1.1 Background

Prior research has extensively investigated various physical cues as indicators of human emotions. These cues encompass a wide range of observable characteristics, including eye gaze direction, iris extension, postural features, and the dynamic movements exhibited by the human body. Some approaches [96, 97] utilize a Kinect sensor to discern and classify emotions from upper body movements and gestures. Specifically, the authors focused on extracting angles and displacements from these body joints to capture the sequential gestures associated with each emotion. However, the limited sensing range of the Kinect sensor and its sensitivity to environmental factors such as occlusions and lighting conditions restrict its ability to capture subtle and nuanced upper body movements, making it challenging to extract accurate features for emotion classification.

Meanwhile, in the context of video analysis, recent advancements have introduced innovative image-based methodologies that exploit the potential of deep learning algorithms. These techniques have been specifically designed to extract and discern keyframes of substantial importance from the continuous stream of frames [98, 99]. However, it is essential to acknowledge that camera-based sensing systems raise concerns regarding privacy, thereby amplifying individuals' anxiety and stress levels due to the perceived constant surveillance. Furthermore, the implementation of such systems may not be economically viable, as they necessitate high-end computational capabilities for efficient analysis and inference processes.

In order to address privacy concerns, utilization of sensors has been employed to gather a range of physiological signals, including brain electrophysiological signals, skin temperature, heart rate monitoring, and electrocardiogram data, for the purpose of human emotion recognition [100, 101]. However, it is crucial to acknowledge that this approach entails certain user inconveniences, as individuals are required to wear sensors to facilitate the monitoring of these signals. Moreover, it should be noted that most of these physiological signals are inherently subjective and exhibit inter-individual variation. Consequently, the reliability of system generalization tends to be compromised, posing challenges to achieving robust and consistent outcomes.

	Psychological state	Definition
G1	Defensive / disagree	Arms crossed in front of the chest
G2	Lost in thought	Hands on cheek
G3	Nervous / stress	Biting the nails
G4	Relaxed	Catapult posture
G5	Deep thought for a decision	Stroking the chin
G6	Impatient while waiting	Finger tapping
G7	Upset	Heads on hand

Figure 7.2: A visual representation of body gestures and their associated psychological state.

To overcome the aforementioned challenges, this chapter embarks upon a comprehensive exploration of the passive recognition of body gestures in relation to workers. Specifically, our study focuses on the identification of emotions by leveraging seven commonly observed body gestures, as depicted in Fig. 7.2. The ESP32 Wi-Fi CSI toolkit is employed as the chosen sensing technology for this purpose. Addressing the limitations associated with camera-based and wearable sensor systems, our approach capitalizes on the ESP32 toolkit as an independent and self-contained solution. This selection is driven by advantageous attributes offered by the toolkit, including low power consumption, user-friendly operation, and portability. Furthermore, inspired by the application of multiple-input multiple-output (MIMO) configurations in other CSI tools, we adopt the ESP32 toolkit in a MIMO configuration to enhance the resolution of the captured CSI data.

7.1.2 Research Contributions and Questions

This study provides the following contributions:

1. We introduce a novel implementation of multiple input multiple output (MIMO) using three ESP32 microcontrollers that operate within

the same channel to enhance the reliability and robustness of data transmission.

2. Our proposed approach for utilizing Wi-Fi CSI for gesture recognition involves two key steps. Firstly, we employ phase calibration techniques to ensure accurate and consistent measurements of the CSI. Secondly, we apply statistical analysis to quantify and analyze phase fluctuations in the phase wavelet domain.

3. We investigate the impact of different dimensional reduction techniques on the performance of the system. The results show that the proposed approach reduces the dimensionality of the data while preserving the accuracy and reliability of the gesture recognition system.

In this chapter, we conduct a series of experiments aimed at validating the proposed models. These experiments are designed to answer the following research questions: We conduct experiments to validate the proposed models to answer the following questions:

RQ1 To what extent does the CSI phase in the wavelet domain impact the performance of the model?

RQ2 Does the model evaluation with the statistical features have a significant impact on the model performance?

7.2 Methodology

This section elaborates on our proposed framework to implement the contactless MIMO ESP32 body gesture system. The proposed system aims to detect and differentiate various gestures in a multi-human context environment by extracting distinct signatures based on the phase information in the wavelet domain. Fig. 7.3 presents an overview of the system framework, which consists of four key modules: data collection and signal interpolation, noise removal, feature extraction, and gesture recognition utilizing machine learning techniques. In the subsequent subsections, we provide a comprehensive description of each step within our framework.

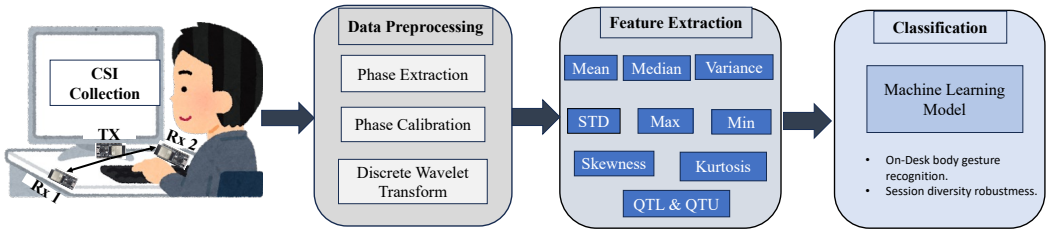


Figure 7.3: Flowchart illustrating the framework to implement on-desk gesture recognition system

7.2.1 Data Collection

The ESP32 nodes were affixed to a desk, enabling the acquisition of CSI measurements when the target individual performed various body gestures. Fig. 7.2 visually depicts the seven distinct on-desk gestures that were considered: "Arms crossed in front of the chest," "Hands on cheek," "Biting the nails," "Catapult posture," "Stroking the chin," "Finger tapping," and "Heads on hand." These gestures were deliberately selected based on their frequency of occurrence and their relevance to emotions and actions commonly observed in work environments. To account for potential variations and assess the system's robustness, data collection was conducted across three separate days, referred to as distinct sessions. By examining the impact of session diversity, we aimed to gain insights into the system's performance and its ability to handle variations that may arise in real-world scenarios.

This stage is divided into three steps as follows:

1. **ESP32 as MIMO system**: ESP32 is typically designed to be a single antenna system, meaning it has one antenna for both transmitting and receiving data. The ESP32 receiver was paired with an ESP32 transmitter, and they operated on the same channel using a single input single output (SISO) configuration. In contrast, the proposed system introduces a novel approach by utilizing the ESP32 as MIMO system as a 1×2, with one transmitter and two receivers. In this MIMO configuration, multiple receivers share the same channel with a single transmitter to improve spatial diversity and enhance the reception of signals. The data is collected by sending a ping between the transmitter and two receivers at certain time intervals.

2. **CSI segmentation**: In order to effectively segment CSI waveforms, two key components were taken into consideration:

   a) MAC address: Each ESP32 receiver is assigned a unique Media Access Control (MAC) address. This address serves as an identifier for the specific receiver, allowing the packets received by each receiver to be appropriately segmented. By associating CSI waveforms with their corresponding MAC addresses, the system is able to distinguish and process the data from individual receivers separately.

   b) Timestamp: To ensure accurate synchronization of the received signals from both receivers, the timestamps of the packets were utilized. By aligning the signals based on their received time, the system concatenated the CSI waveforms from both receivers. This process enabled the combination of the signals, providing a comprehensive representation of the received data from the multiple ESP32 receivers.

   Consequently, the gathered CSI is represented as $\mathbf{H}_{\mathrm{i}}(f) \in \mathbb{C}^{1\times2\times52}$. A total of 104 CSI waveforms are gathered for every received packet, resulting in a substantial enhancement in the level of detail and granularity of the collected data samples.

3. **Linear interpolation**: To tackle the packet loss problem, the linear interpolation method involves estimating the missing or delayed packets by inferring their values based on the surrounding packets. This technique allows us to fill in the gaps caused by packet loss or latency, ensuring a more uniform distribution of the CSI waveforms. By preserving the desired spacing between the packets, we can facilitate more accurate and consistent analysis of the received data with the same length.

7.2.2 Data Preprocessing

The CSI phase is a valuable source of information for capturing the nuances of the target's gesture. However, its practical utilization is often hindered by

the presence of environmental noise interference as mentioned in chapter2. In order to enhance the reliability and accuracy of the CSI phase, it is imperative to mitigate the impact of hardware and environmental fluctuations. This section focuses on achieving this objective through the minimization of CFO and SFO errors. By applying a phase calibration algorithm, the detrimental effects caused by CFO and SFO can be effectively eliminated, as will be discussed in detail later. Additionally, to extract the phase variations associated with body gestures, the discrete wavelet transform (DWT) algorithm is employed. This algorithm plays a crucial role in identifying and isolating the relevant phase variations induced by the targeted gestures. By leveraging the capabilities of the DWT algorithm, the system aims to extract and analyze the distinct features within the CSI phase, facilitating accurate and reliable gesture recognition.

1. **Phase calibration**: The measured CSI phase value θ_{i} of the ith subcarrier can be expressed as in Eq. 7.1 [102]:

$$\theta = \hat{\theta} + \frac{2\pi k_i \delta t}{M} + \beta + Z_f \tag{7.1}$$

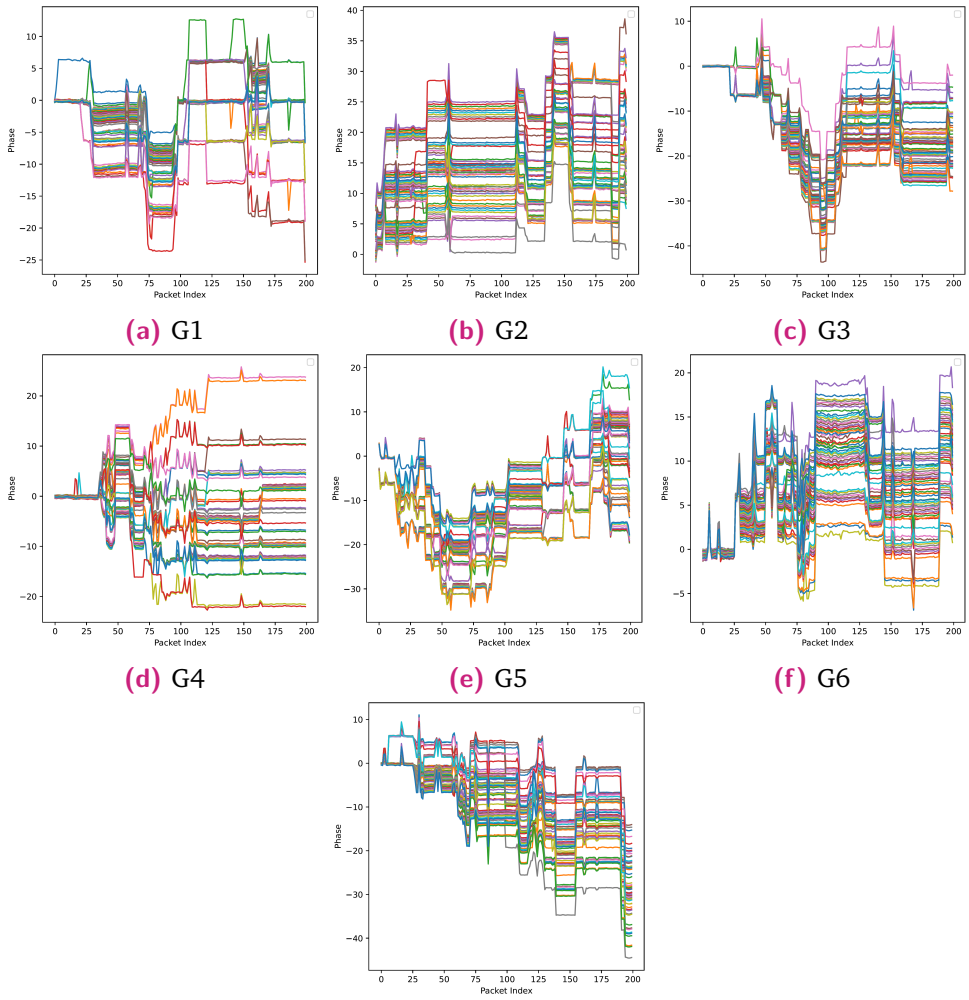
   where $\hat{\theta}$ is the true phase, δt is the time lag due to SFO, K_i is the subcarrier index of the ith subcarrier for *i=1 to 52*, β is the phase offset due to CFO, M is the fast Fourier transform (FFT) size and is set to 64 based on the IEEE 802.11n specification.

   To eliminate the effect of δt and β, a linear transformation is applied to the raw phase after unfolding it. The steps of phase calibration are as follows:

   a) **Unwrapping the raw phase:** The CSI phase measured by ESP32 node is within the range $[-\pi, \pi]$ while the true phase is in range $[0, 2\pi]$.

   b) **Applying linear transformation on unwrapped phase:** a and b are estimated as written in Eq. 7.2. They represent the slope of the phase and the phase offset across the frequency band, respectively.

$$a = \frac{\theta_{52} - \theta_1}{k_{52} - k_1}, b = \frac{1}{52}\sum_{i=1}^{52}\theta_i \tag{7.2}$$



(a) G1 (b) G2 (c) G3

(d) G4 (e) G5 (f) G6

(g) G7

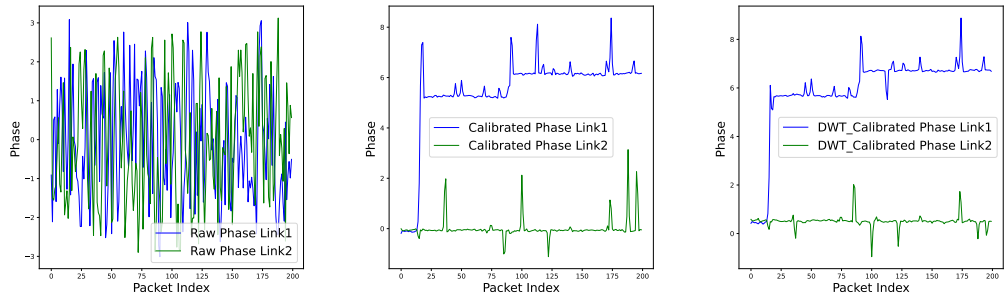
Figure 7.4: Calibrated Phase of Different Body Gestures across All Subcarriers

After that, the calibrated phase $\hat{\theta}$ can be given as in Eq. 7.3:

$$\hat{\theta} = \theta - ak_i - b \tag{7.3}$$

Fig. 7.4 illustrates the phase calibration signatures for the seven body gestures across all subcarriers.

2. **Discrete Wavelet Transform (DWT)**: By leveraging the advantages of DWT, such as its multiresolution analysis, localization properties, efficiency, and adaptability, the system aims to effectively remove in-band distortion and enhance the accuracy and reliability of the gesture recognition process.



(a) Raw Phase **(b)** Calibrated Phase **(c)** DWT Phase

Figure 7.5: Phase Noise Removal of Arms crossed in front of the chest gesture across 5th subcarrier across the two links

The calibrated phase obtained from the phase calibration process is subsequently fed into the DWT to address in-band distortion of the signal. The DWT transforms the time-domain signal into the wavelet domain, decomposing it into wavelet detail and approximate coefficients as explained in Chapter 3.4.2. It is important to emphasize the preservation of the approximation coefficients since they provide valuable insights into subtle body movements. These coefficients capture the fundamental details of the gesture, while the high-frequency components of the signal primarily capture ambient noise. Fig. 7.5 shows the raw, calibrated and applying DWT on the calibrated phase of the fifth subcarrier across both links between the transmitters and two receivers.

7.2.3 Feature Extraction

Feature extraction is a crucial step in converting complex CSI data into meaningful and representative features that can be effectively utilized in subsequent processing and integration with machine learning models. In the proposed approach, the focus is on extracting statistical features from the wavelet phase fluctuations observed in different environments, as these features provide a solid theoretical foundation for analysis.

To tackle the challenge of high dimensionality associated with subcarrier data per frame, the approach leverages standard statistical aggregation

functions. These functions, including mean, standard deviation, median, lower quartile, upper quartile, minimum, maximum, skewness, and kurtosis, are applied to extract essential statistical characteristics from the wavelet phase information. By compressing the wavelet phase into a single higher-level feature value, these functions effectively reduce the dimensionality of the data. Compared to other dimensionality reduction techniques such as Principal Component Analysis (PCA) and autoencoders, the utilization of statistical features provides several advantages. Firstly, it simplifies the representation of the data, reducing its dimensionality while preserving the essential statistical characteristics associated with the phase fluctuations. This compression facilitates more efficient computation and storage requirements, making it suitable for real-time and resource-constrained applications.

Furthermore, the extraction of statistical features enhances interpretability, which is particularly valuable in the context of explainable AI. By capturing statistical properties such as distribution, central tendency, dispersion, and shape of the phase fluctuations, these features provide insights into the underlying patterns and dynamics within the CSI data. This interpretability aspect enables researchers and practitioners to gain a deeper understanding of the relationships between the extracted features and the target gestures, facilitating model analysis, debugging, and the exploration of causal relationships.

7.2.4 Gesture Recognition Based on Machine Learning

The lack of interpretability and resource-intensive nature of deep learning (DL) models pose challenges in real-world applications where understanding the model's reasoning and deploying it on constrained devices are crucial. DL models are often perceived as black boxes that offer limited explanations for their results, making it difficult for users to comprehend the decision-making process behind the model's predictions. This interpretability issue can be problematic in domains such as business decision-making and medical diagnosis, where users require transparency in the model's outputs. Furthermore, DL models consume substantial time and memory resources during training and inference, which can be particularly challenging when deploying them

on low-resource microcontroller devices. These devices have limited computational capabilities, making it necessary to consider factors such as memory consumption, computation time, and energy usage when applying machine learning (ML) models to them.

In light of these challenges, our research focuses on utilizing ML models, particularly the Random Forest technique, to address the interpretability and resource efficiency requirements. Random Forest models offer several advantages over DL models in this context. Firstly, they provide greater interpretability, enabling users to understand the factors that contribute to a prediction. This interpretability is vital in real-world applications where the reasoning behind the model's output is essential. Additionally, Random Forest models are computationally efficient and require less memory compared to DL models. This computational advantage makes them well-suited for deployment on low-resource microcontroller devices, where limitations in time, memory, and energy consumption are critical factors.

7.2.5 Model Evaluation

We evaluate the model performance using the accuracy and confusion matrix, as in Chapter 3.6.

7.3 Performance Evaluation

In this section, we first introduce the experimental setup. Then, we demonstrate the evaluation results of our proposed on-desk gesture recognition system.

(a) Side View **(b)** Top View

Figure 7.6: Different Views of Experimental Setup

7.3.1 Experiment Setup

Hardware and Software

To simulate real-world scenarios, we conducted a series of experiments in a multi-human context environment over the course of three days. Our experimental setup involved the utilization of three ESP32 microcontrollers, chosen for their compact size and low energy consumption. One microcontroller served as the transmitter (Tx), while the other two functioned as receivers (Rx), all operating on the same channel. Fig. 7.6(a) and Fig. 7.6(b) depict the side and top views of our experimental setup, strategically utilizing underutilized areas as considered by users. The transmitter was positioned 81 cm away from the two receivers, with a distance of 127 cm between the receivers themselves. In this experiment, all transceivers shared the third channel, and the configuration ensured that the receiver extracted approximately 100 CSI packets per second for 52 subcarriers. These CSI frames provided information on multipath effects caused by phenomena such as reflection, scattering, power, and distance fading within the sensing area. To process the obtained CSI data, perform phase extraction, remove noise, and implement the learning models, we utilized Python 3.9.

Data Description

To assess the robustness of our system across different sessions and locations, we conducted a series of experiments that involved collecting data from two

distinct locations over a period of three days. The volunteer, while simulating desk-related activities, performed seven specific gestures on the desk surface, as illustrated in Fig. 7.2. Throughout the experiments, each gesture was captured within a one-second timeframe. The ping transmission rate was set at 100 Hz, and the dimensionality of each gesture sample was represented as $\mathbb{R}^{100\times52\times2}$, where 100 signifies the length of the signal, 52 denotes the number of subcarriers, and 2 represents the number of links. To ensure diversity in both session and location, we collected CSI data on three separate days. On the first day, the participant was placed in a real-world setting surrounded by a significant number of individuals. The second day involved a single-user environment with only the participant present, while the third day had a small number of individuals within the vicinity.

To assess the system's performance, we gathered datasets of varying durations for each gesture across the three days, specifically 20 minutes, 10 minutes, and 15 minutes. These datasets were employed to validate the system's functionality and evaluate its performance in different scenarios and over extended periods of time. The details of the three sessions are summarized in Table 7.1.

Table 7.1: Properties of Body Gesture Dataset

Day	Time(minutes)	Body Gesture							
		G1	G2	G3	G4	G5	G6	G7	Total
Day1	20	601	616	632	667	627	628	641	4412
Day2	10	309	320	311	325	311	307	316	1888
Day3	15	449	388	464	482	451	445	441	3120

7.3.2 Results

In this section, to comprehensively assess the robustness of our proposed system, we adapt the accuracy metric and utilize two validation methods as outlined below:

1. In-session cross validation: With this approach, we divide the dataset from each individual day into 70% for training and 30% for testing. This method allows us to evaluate the system's performance within the context of a single session, providing insights into its effectiveness in handling variations within a given day's dataset.

2. Leave-one-session-out cross validation: This validation method enables us to evaluate the generalization and robustness of the proposed system across different sessions and locations. It involves training the model using the dataset from one session while utilizing the dataset from another session for testing.

Table 7.2: Overall proposed system performance in terms of accuracy

Validation method	Dataset	Models			
		SVM	RF	NB	GBC
In-session	Day1	**98**	**98**	94	97.2
	Day2	98.4	**99**	95.4	97.6
	Day3	98	**98.7**	95	97
Leave-one-session-out	Day1 → Day1	57.4	**66.2**	48.3	65.8
	Day1 → Day2	57.3	**68.8**	43	58.8
	Day2 → Day1	59.4	65.4	48	**66.4**
	Day2 → Day3	56.1	**68.7**	46	68
	Day3 → Day1	57	**69**	46	65.4
	Day3 → Day2	57.5	**72**	53	61.3

To determine the most suitable classifier for our study, we evaluated multiple algorithms, including Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), and Gradient Boosting Classifier (GBC). Table 7.2 presents a comparison of the performance of these classifiers. Notably, the Random Forest algorithm outperformed the other three approaches in both in-session validation and session diversity robustness.

In the context of in-session validation, the accuracy of Support Vector Machine (SVM) and Random Forest (RF) classifiers were found to be relatively close to each other. This can be attributed to SVM's ability to classify features based on common characteristics within the same session. However, SVM failed to achieve satisfactory accuracy in session diversity robustness due to overfitting, as its training dataset's distribution differed from that of the testing dataset. On the other hand, Random Forest demonstrated the capability to capture relevant features and achieve reasonable accuracy even when faced with diverse sessions.

Therefore, despite the modest overall accuracy of 72%, Random Forest's performance surpasses that of the other classifiers, particularly in terms of session diversity robustness. Its ability to effectively capture relevant features and adapt to varying session conditions justifies its acceptability for our study, despite the relatively lower accuracy.

Table 7.3, 7.4, 7.5, and 7.6 represent the confusion matrix for in-session and out-session cross validation.

Table 7.3: Day1 confusion matrix

	Predicted							Per-class metrics		
	G1	G2	G3	G4	G5	G6	G7	PR	RE	F1
G1	1	0	0	0	0	0	0	1	1	1
G2	0	0.94	0.06	0	0	0	0	0.82	.94	0.88
G3	0	0.01	0.99	0.01	0	0	0	0.99	0.99	0.99
G4	0	0.01	0	0.99	0	0	0	0.99	0.99	0.99
G5	0	0	0	0	0.99	0.01	0	0.99	0.99	0.99
G6	0	0	0	0	0.01	0.99	0	0.99	0.99	0.99
G7	0	0	0	0	0	0	1	1	1	1

Table 7.4: Day2 confusion matrix

	Predicted							Per-class metrics		
	G1	G2	G3	G4	G5	G6	G7	PR	RE	F1
G1	1	0	0	0	0	0	0	1	1	1
G2	0	0.94	0.06	0	0	0	0	0.82	.94	0.88
G3	0	0	1	1	0	0	0	1	1	1
G4	0	0.01	0	0.99	0	0	0	0.99	0.99	0.99
G5	0	0	0	0	1	0	0	1	1	1
G6	0	0	0	0	0.01	0.99	0	0.99	0.99	0.99
G7	0	0	0	0	0	0	1	1	1	1

Table 7.5: Day3 confusion matrix

	Predicted							Per-class metrics		
	G1	G2	G3	G4	G5	G6	G7	PR	RE	F1
G1	1	0	0	0	0	0	0	1	1	1
G2	0	0.95	0.05	0	0	0	0	0.83	.94	0.88
G3	0	0	1	0	0	0	0	1	1	1
G4	0	0.01	0	0.99	0	0	0	0.99	0.99	0.99
G5	0	0	0	0	1	0	0	1	1	1
G6	0	0	0	0	0.01	0.99	0	0.99	0.99	0.99
G7	0	0	0	0	0	0	1	1	1	1

Moreover, we conducted additional analysis to explore the influence of various base signals and dimensionality reduction techniques on the robustness of session diversity.

7.3.3 Effect of Different Base Signals

In this study, we aim to investigate the influence of different base signals in Wi-Fi CSI-based systems. Traditionally, these systems have focused on

Table 7.6: Day3 → Day2 confusion matrix

	Predicted							Per-class metrics		
	G1	G2	G3	G4	G5	G6	G7	PR	RE	F1
G1	0.76	0.17	0	0	0.05	0.02	0	0.77	0.78	0.78
G2	0.02	0.78	0.01	0.01	0.16	0.02	0	0.51	0.76	0.61
G3	0.01	0.3	0.66	0.03	0.01	0	0	0.78	0.66	0.71
G4	0	0.04	0.12	0.75	0.09	0	0	0.92	0.75	0.83
G5	0.04	0.25	0.07	0.01	0.58	0.05	0	0.6	0.59	0.59
G6	0.15	0	0	0	0.05	0.79	0	0.69	0.78	0.73
G7	0	0.02	0	0.03	0	0.24	0.71	0.99	0.72	0.83

extracting the amplitude and using its variations as the base signal for feature extraction and learning algorithms. Additionally, many of these systems employ a Hampel filter, mentioned in Chapter 3.4.1 to eliminate outliers in the amplitude. Some previous research has combined both amplitude and phase as input for the system. However, to the best of our knowledge, our study is the first to extract the variations of the phase values in the wavelet domain as signatures of different body gestures in Wi-Fi CSI-based systems.

We have found that the phase values convey the most significant features of different gestures in various sessions, as observed in Figure 8. By applying DWT to the calibrated phase, we were able to improve the recognition accuracy by 10% compared to using only the calibrated phase. DWT effectively mitigates the impact of environmental noise and wireless signal interference on the same channel.

On the other hand, when utilizing the amplitude, the performance of the system degrades, achieving only 46% accuracy. This is because body gestures involve micro-scale motion, which is challenging to extract from the CSI magnitude, especially in the presence of signal interference when two receivers share the same channel. Based on our evaluation using various machine learning techniques, we observed that the performance of the calibrated phase alone slightly outperforms the calibrated phase with DWT when employing SVM and GBC algorithms. However, despite this slight difference in performance, utilizing DWT for the calibrated phase offers a significant advantage in terms of data compression without sacrificing information. This advantage allows the system to be lightweight and easily deployed while still achieving the highest accuracy among the tested algorithms, specifically 72% accuracy based on RF. We summarize the recognition accuracy of different base signals into Table 7.7.

Table 7.7: The recognition accuracy of different base signals

Base Signal	Models			
	SVM	RF	NB	GBC
Hampel Amplitude	**46.3**	46	33	43
DWT Hampel Amplitude	44.3	**47**	37	35
Calibrated Phase	59	**69**	52	68
Hampel Amplitude + Calibrated Phase	45	**67**	38	63
DWT Calibrated Phase	57.5	**72**	52.8	61.3

7.3.4 Effect of Different Dimensional Reduction Methods

We investigate the impact of various dimensional reduction (DR) techniques, including Principal Component Analysis (PCA), Dense Autoencoder (AE), and Convolutional Autoencoder (CAE), for extracting representative features from each gesture, as summarized in Table 7.8. However, the classifier's performance deteriorates when utilizing PCA since it is primarily a linear projection-based dimensionality reduction technique. In other words, PCA transforms high-dimensional data into a lower-dimensional space, potentially leading to the loss of relevant information and subsequent misclassification.

AE and CAE, on the other hand, serve as black-box feature extraction methods based on neural network architectures. Nonetheless, these techniques lack support for explainable AI and may not be well-suited for analyzing time series data due to its temporal nature. In the case of AE, the network design aims to create a compressed representation of the input data, but it fails to effectively capture the temporal dependencies that are often crucial for precise classification. Similarly, CAE focuses on spatial feature extraction and struggles to extract sequential information effectively from CSI time series data.

In contrast, the extraction of statistical features ensures an energy-efficient and lightweight classification process, promoting efficiency in resource consumption. Additionally, this approach facilitates explainable AI, enabling a clear understanding of the decision-making process within the system.

Table 7.8: The recognition accuracy of different dimensional reduction methods

DR Models	Models			
	SVM	RF	NB	GBC
Statisitcal Features	57.5	**72**	52.8	61.3
PCA	38	**56**	29.5	40.4
AE	14	**30.1**	23	28
CAE	15.7	14.3	14.8	14.2

7.4 Summary

This study introduced a promising passive body gesture recognition system that utilizes the ESP32 node as a Wi-Fi CSI tool. In contrast to conventional methods, our approach involves transforming the CSI phase through a linear transformation applied in the wavelet domain. Furthermore, we extract hand-crafted features to ensure the explainability, robustness, and practical applicability of our system in real-world scenarios.

To evaluate the system's performance, we conducted three experiments across different days and environments, aiming to assess its robustness in various sessions and in a multi-human context environment. Additionally, we investigated the influence of different machine learning classifiers, base signals, and dimensional reduction methods. Our results demonstrate the superiority of our proposed methods, surpassing other approaches and achieving an impressive recognition accuracy of 98% for in-session evaluation and 72% for session diversity robustness.

8 Conclusions

8.1 Achieved Aims and Objectives

The main objective of this thesis is to build a robust Wi-Fi CSI gesture recognition system in a multi-human context environment.

In this research, we have presented pioneering work in the field of Wi-Fi CSI, specifically focusing on enabling its use in real-world scenarios. Our study aimed to overcome the limitations of existing CSI tools and harness the potential of Wi-Fi CSI for various applications.

To achieve this, we collected data using ESP32, a powerful microcontroller, to address the shortcomings of existing CSI tools. By utilizing ESP32, we were able to capture and analyze CSI measurements with higher accuracy and precision, enabling more robust and reliable results.

Through our research, we explored the applications of Wi-Fi CSI in head gestures and body gestures, highlighting their potential for communication and interaction purposes. We introduced techniques to enhance the robustness of Wi-Fi CSI in real-world settings.

One of the key challenges we tackled the issue of location diversity, which is common in real-world scenarios. By developing adaptive methodologies, we improved the generalization capabilities of Wi-Fi CSI models across different environments, enabling consistent performance and applicability.

Furthermore, our research considered the complexities of multi-human context environments. We proposed novel algorithms and techniques to handle interference and variability introduced by multiple individuals, making Wi-Fi CSI more effective in these scenarios.

We introduced two applications with different frameworks. The first one is a communication method for quadriplegia patients based on Morse code. The

other is to passively detect employees' emotions within the workplace setting. To record CSI waveforms, we utilize the ESP32 microcontroller due to its compact size, low power consumption, and cost-effectiveness.

We begin with with the Wi-Nod system, which serves as the foundation for a communication system designed for quadriplegia patients utilizing head nodding gestures. To implement this system, we developed a specialized frame to look like a real wheelchair. Data collection was conducted in a laboratory environment involving two users with a caregiver who holding the frame behind the target. We gathered the data on two different time, morning and evening, to investigate the session diversity robustness. To process the collected data, we calculated the spectrogram using the amplitude of the CSI.

Subsequently, the spectrogram was utilized as input for an inception model, which served as the classifier which achieved an accuracy rate exceeding 95% for three distinct symbols. Furthermore, we also investigated the system's robustness concerning user diversity and time diversity. Considering that Wi-Fi signals are subject to variations due to environmental changes, it was crucial to assess the system's ability to maintain accurate performance in diverse scenarios.

After that, we introduced the HeMoFi4Q system as an extension of the Wi-Nod system, addressing its main limitation related to data collection in a fixed location that does not reflect real-world scenarios. In the HeMoFi4Q system, we mounted six ESP32 microcontrollers on a real wheelchair and collected data sets in two different environments with distinct locations. This allowed us to investigate the system's robustness in both single-use and multi-human context environments. We proposed a novel sign language approach based on head motion and Morse code.

To address the domain shift problem and enable the system to capture head motion-related features, we presented a new technique inspired by few-shot learning models. Additionally, we employed an ECA model, a powerful computer vision classifier known for its low parameter count and high performance.

We further examined the impact of different base signals and link configurations on the system's performance. Experimental results revealed that

variations in amplitude in the time domain achieved the highest accuracy, and the ECA model outperformed other state-of-the-art algorithms in terms of both accuracy and computational time.

Lastly, we presented an on-desk passive emotion recognition system based on upper body gestures. To develop this system, we collected a dataset comprising seven different body gestures in two distinct locations over three days. We utilized three ESP32 microcontrollers as a MIMO configuration, with one serving as the transmitter and the others as receivers operating on the same channel.

CSI waveforms were parsed and segmented based on their corresponding time stamps and MAC addresses of each receiver. To mitigate packet loss, we employed a linear interpolation technique. We observed that the shared channel introduced collisions and interference, impacting the CSI amplitudes. Consequently, we focused on extracting the CSI phase, as it was found to be less influenced by signal interference.

To enhance the accuracy of the phase data, we applied phase calibration techniques to mitigate CFO and SFO. Next, we utilized DWT on the filtered phase to generate phase signatures corresponding to each body gesture in the stable wavelet domain. For dimensionality reduction and feature selection, several statistical measures such as mean, standard deviation, median, lower quartile, upper quartile, minimum, maximum, skewness, and kurtosis were calculated for the wavelet phase values.

Finally, we compared the performance of various common machine learning models and found that the random forest model outperformed the others, achieving an accuracy rate exceeding 72% in terms of location diversity robustness.

Overall, our study contributes to the growing body of knowledge in Wi-Fi CSI research and establishes its potential for real-world applications. By collecting data using ESP32, we overcame the limitations of existing CSI tools, enabling more accurate and precise measurements. The applications and techniques we introduced enhance the robustness of Wi-Fi CSI, making it a valuable tool in the healthcare domain.

8.2 Future Work

Considering different HeMoFi4Q Morse code characters to form complete words, while also incorporating spelling correction capabilities in real-world scenarios. This additional challenge requires us to consider the integration of NLPalgorithms. By incorporating NLP techniques, we aim to develop a system that can effectively generate words by combining Morse code characters and ensure accurate spelling correction in practical applications.

For future on-desk emotional recognition research, recognizing a worker's body gestures during their work activities could be proposed as a future task. This will involve self-segmentation for the CSI signals and extracting relevant features associated with the body gestures to facilitate the classification task. By employing these techniques, we aim to develop a comprehensive system capable of accurately recognizing and classifying various body gestures exhibited by workers in real-time work scenarios.

Bibliography

[1] Sojeong Yun and Youn-Kyung Lim. "Potential and Challenges of DIY Smart Homes with an ML-intensive Camera Sensor". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–19 (cit. on p. 1).

[2] Akashdeep Bhardwaj, Keshav Kaushik, Salil Bharany, and SeongKi Kim. "Forensic analysis and security assessment of IoT camera firmware for smart homes". In: *Egyptian Informatics Journal* 24.4 (2023), p. 100409 (cit. on p. 1).

[3] Youjin Jang, Inbae Jeong, Moein Younesi Heravi, et al. "Multi-Camera-Based Human Activity Recognition for Human–Robot Collaboration in Construction". In: *Sensors* 23.15 (2023), p. 6997 (cit. on p. 1).

[4] Md Moniruzzaman, Zhaozheng Yin, Md Sanzid Bin Hossain, Hwan Choi, and Zhishan Guo. "Wearable motion capture: Reconstructing and predicting 3D human poses from wearable sensors". In: *IEEE Journal of Biomedical and Health Informatics* (2023) (cit. on p. 1).

[5] Adeola Bannis, Shijia Pan, Carlos Ruiz, et al. "IDIoT: Multimodal Framework for Ubiquitous Identification and Assignment of Human-carried Wearable Devices". In: *ACM Transactions on Internet of Things* 4.2 (2023), pp. 1–25 (cit. on p. 1).

[6] Yu Shi, Lan Du, Xiaoyang Chen, et al. "Robust Gait Recognition based on Deep CNNs with Camera and Radar Sensor Fusion". In: *IEEE Internet of Things Journal* (2023) (cit. on p. 1).

[7] Aaron Asael Smith, Rui Li, and Zion Tsz Ho Tse. "Reshaping healthcare with wearable biosensors". In: *Scientific Reports* 13.1 (2023), p. 4998 (cit. on p. 1).

[8] Yahia Baashar, Gamal Alkawsi, Wan Nooraishya Wan Ahmad, et al. "Towards wearable augmented reality in healthcare: a comparative survey and analysis of Head-Mounted displays". In: *International journal of environmental research and public health* 20.5 (2023), p. 3940 (cit. on p. 1).

[9] Antonio Guerrero-Ibañez, Ismael Amezcua-Valdovinos, and Juan Contreras-Castillo. “Integration of wearables and wireless technologies to improve the interaction between disabled vulnerable road users and self-driving cars”. In: *Electronics* 12.17 (2023), p. 3587 (cit. on p. 1).

[10] Alfredo J Perez, Farhan Siddiqui, Sherali Zeadally, and Derek Lane. “A review of IoT systems to enable independence for the elderly and disabled individuals”. In: *internet of Things* 21 (2023), p. 100653 (cit. on p. 1).

[11] Md Golam Morshed, Tangina Sultana, Aftab Alam, and Young-Koo Lee. “Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities”. In: *Sensors* 23.4 (2023), p. 2182 (cit. on p. 1).

[12] Xuena Chen, Li Su, Jinxiu Zhao, et al. “Sign language gesture recognition and classification based on event camera with spiking neural networks”. In: *Electronics* 12.4 (2023), p. 786 (cit. on p. 1).

[13] Wei Lv. “Gesture recognition in somatosensory game via Kinect sensor”. In: *Internet Technology Letters* 6.5 (2023), e311 (cit. on p. 1).

[14] Sungho Suh, Vitor Fortes Rey, and Paul Lukowicz. “TASKED: Transformer-based Adversarial learning for human activity recognition using wearable sensors via Self-KnowledgE Distillation”. In: *Knowledge-Based Systems* 260 (2023), p. 110143 (cit. on p. 1).

[15] Xuping Wu, Xuemei Luo, Zhuman Song, et al. “Ultra-Robust and Sensitive Flexible Strain Sensor for Real-Time and Wearable Sign Language Translation”. In: *Advanced Functional Materials* (2023), p. 2303504 (cit. on p. 1).

[16] Rayane Tchantchane, Hao Zhou, Shen Zhang, and Gursel Alici. “A review of hand gesture recognition systems based on noninvasive wearable sensors”. In: *Advanced Intelligent Systems* 5.10 (2023), p. 2300207 (cit. on p. 1).

[17] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. “Predictable 802.11 packet delivery from wireless channel measurements”. In: *ACM SIGCOMM computer communication review* 40.4 (2010), pp. 159–170 (cit. on pp. 2, 22).

[18] Yaxiong Xie, Zhenjiang Li, and Mo Li. “Precise power delay profiling with commodity WiFi”. In: *Proceedings of the 21st Annual international conference on Mobile Computing and Networking*. 2015, pp. 53–64 (cit. on pp. 2, 21–23).

[19] Jacopo Tosi, Fabrizio Taffoni, Marco Santacatterina, Roberto Sannino, and Domenico Formica. “Performance evaluation of bluetooth low energy: A systematic review”. In: *Sensors* 17.12 (2017), p. 2898 (cit. on p. 9).

[20] Gernot Bahle, Vitor Fortes Rey, Sizhen Bian, Hymalai Bello, and Paul Lukowicz. "Using privacy respecting sound analysis to improve bluetooth based proximity detection for COVID-19 exposure tracing and social distancing". In: *Sensors* 21.16 (2021), p. 5604 (cit. on p. 9).

[21] Shuangquan Wang and Gang Zhou. "A review on radio based activity recognition". In: *Digital Communications and Networks* 1.1 (2015), pp. 20–29 (cit. on p. 10).

[22] Jue Wang, Deepak Vasisht, and Dina Katabi. "RF-IDraw: Virtual touch screen in the air using RF signals". In: *ACM SIGCOMM Computer Communication Review* 44.4 (2014), pp. 235–246 (cit. on p. 10).

[23] Xinyu Li, Yuan He, and Xiaojun Jing. "A survey of deep learning-based human activity recognition in radar". In: *Remote Sensing* 11.9 (2019), p. 1068 (cit. on p. 10).

[24] Jia Zhang, Rui Xi, Yuan He, et al. "A survey of mmWave-based human sensing: Technology, platforms and applications". In: *IEEE Communications Surveys & Tutorials* (2023) (cit. on p. 11).

[25] Yangfan Sun, Renlong Hang, Zhu Li, Mouqing Jin, and Kelvin Xu. "Privacy-preserving fall detection with deep learning on mmWave radar signal". In: *2019 IEEE Visual Communications and Image Processing (VCIP)*. IEEE. 2019, pp. 1–4 (cit. on p. 11).

[26] Peijun Zhao, Chris Xiaoxuan Lu, Bing Wang, et al. "Heart rate sensing with a robot mounted mmwave radar". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 2812–2818 (cit. on p. 11).

[27] Hira Hameed, Muhammad Usman, Ahsen Tahir, et al. "Pushing the limits of remote RF sensing by reading lips under the face mask". In: *Nature Communications* 13.1 (2022), p. 5168 (cit. on p. 12).

[28] Scott Y Seidel and Theodore S Rappaport. "914 MHz path loss prediction models for indoor wireless communications in multifloored buildings". In: *IEEE transactions on Antennas and Propagation* 40.2 (1992), pp. 207–217 (cit. on p. 13).

[29] Maurizio Bocca, Ossi Kaltiokallio, Neal Patwari, and Suresh Venkatasubramanian. "Multiple target tracking with RF sensor networks". In: *IEEE Transactions on Mobile Computing* 13.8 (2013), pp. 1787–1800 (cit. on p. 14).

[30] Chenren Xu, Bernhard Firner, Robert S Moore, et al. "SCPL: Indoor device-free multi-subject counting and localization using radio signal strength". In: *Proceedings of the 12th international conference on Information Processing in Sensor Networks*. 2013, pp. 79–90 (cit. on p. 14).

[31] Heba Abdelnasser, Moustafa Youssef, and Khaled A Harras. "Wigest: A ubiquitous wifi-based gesture recognition system". In: *2015 IEEE conference on computer communications (INFOCOM)*. IEEE. 2015, pp. 1472–1480 (cit. on p. 14).

[32] Xuyu Wang, Xiangyu Wang, and Shiwen Mao. "Deep convolutional neural networks for indoor localization with CSI images". In: *IEEE Transactions on Network Science and Engineering* 7.1 (2018), pp. 316–327 (cit. on p. 15).

[33] Xiaolong Zheng, Jiliang Wang, Longfei Shangguan, Zimu Zhou, and Yunhao Liu. "Smokey: Ubiquitous smoking detection with commercial WiFi infrastructures". In: *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE. 2016, pp. 1–9 (cit. on p. 15).

[34] Jianfei Yang, Han Zou, Hao Jiang, and Lihua Xie. "CareFi: Sedentary behavior monitoring system via commodity WiFi infrastructures". In: *IEEE Transactions on Vehicular Technology* 67.8 (2018), pp. 7620–7629 (cit. on p. 15).

[35] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. "SignFi: Sign language recognition using WiFi". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.1 (2018), pp. 1–21 (cit. on pp. 15, 20, 24, 42–44, 46, 48, 49).

[36] Marwa RM Bastwesy, Nada M ElShennawy, and Mohamed T Faheem Saidahmed. "Deep learning sign language recognition system based on wi-fi csi". In: *Int. J. Intell. Syst. Appl* 12.6 (2020), pp. 33–45 (cit. on pp. 15, 24, 42–49).

[37] Hyuckjin Choi, Manato Fujimoto, Tomokazu Matsui, Shinya Misaki, and Keiichi Yasumoto. "Wi-cal: Wifi sensing and machine learning based device-free crowd counting and localization". In: *IEEE Access* 10 (2022), pp. 24395–24410 (cit. on pp. 15, 24).

[38] Guillermo Diaz, Iker Sobron, Iñaki Eizmendi, et al. "Channel phase processing in wireless networks for human activity recognition". In: *Internet of Things* 24 (2023), p. 100960 (cit. on p. 18).

[39] Yongsen Ma, Gang Zhou, and Shuangquan Wang. "WiFi sensing with channel state information: A survey". In: *ACM Computing Surveys (CSUR)* 52.3 (2019), pp. 1–36 (cit. on pp. 19, 31, 32).

[40] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. "Tool release: Gathering 802.11 n traces with channel state information". In: *ACM SIGCOMM computer communication review* 41.1 (2011), pp. 53–53 (cit. on pp. 20, 23, 43).

[41] Andrii ZHURAVCHAK. "Human Activity Recognition based on WiFi CSI data". PhD thesis. Ukrainian Catholic University, 2020 (cit. on p. 21).

[42] Matthias Schulz, Daniel Wegemer, and Matthias Hollick. “The Nexmon firmware analysis and modification framework: Empowering researchers to enhance Wi-Fi devices”. In: *Computer Communications* 129 (2018), pp. 269–285 (cit. on pp. 21–23).

[43] Fahd Abuhoureyah, Wong Yan Chiew, Ahmad Sadhiqin Bin Mohd Isira, and Mohammed Al-Andoli. “Free device location independent WiFi-based localisation using received signal strength indicator and channel state information”. In: *IET Wireless Sensor Systems* 13.5 (2023), pp. 163–177 (cit. on p. 22).

[44] Steven M Hernandez and Eyuphan Bulut. “Lightweight and standalone IoT based WiFi sensing for active repositioning and mobility”. In: *2020 IEEE 21st International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*. IEEE. 2020, pp. 277–286 (cit. on pp. 22, 23).

[45] Xuyu Wang, Lingjun Gao, and Shiwen Mao. “PhaseFi: Phase fingerprinting for indoor localization with a deep learning approach”. In: *2015 IEEE Global Communications Conference (GLOBECOM)*. IEEE. 2015, pp. 1–6 (cit. on p. 24).

[46] Yuxi Wang, Kaishun Wu, and Lionel M Ni. “Wifall: Device-free fall detection by wireless networks”. In: *IEEE Transactions on Mobile Computing* 16.2 (2016), pp. 581–594 (cit. on p. 26).

[47] Linsong Cheng and Jiliang Wang. “How can I guard my AP? Non-intrusive user identification for mobile devices using WiFi signals”. In: *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 2016, pp. 91–100 (cit. on p. 26).

[48] Tahmid Z Chowdhury, Cyril Leung, and Chun Yan Miao. “WiHACS: Leveraging WiFi for human activity classification using OFDM subcarriers’ correlation”. In: *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. 2017, pp. 338–342 (cit. on p. 26).

[49] Wenjun Jiang, Chenglin Miao, Fenglong Ma, et al. “Towards environment independent device free human activity recognition”. In: *Proceedings of the 24th annual international conference on mobile computing and networking*. 2018, pp. 289–304 (cit. on p. 26).

[50] Ishtiaque Ahmed Showmik, Tahsina Farah Sanam, and Hafiz Imtiaz. “Human Activity Recognition from Wi-Fi CSI Data Using Principal Component-Based Wavelet CNN”. In: *Digital Signal Processing* 138 (2023), p. 104056 (cit. on p. 29).

[51] Wenda Li, Mohammud Junaid Bocus, Chong Tang, et al. “On CSI and passive Wi-Fi radar for opportunistic physical activity recognition”. In: *IEEE Transactions on Wireless Communications* 21.1 (2021), pp. 607–620 (cit. on p. 29).

[52] Shuai Yang, Dongheng Zhang, Ruiyuan Song, Pengfei Yin, and Yan Chen. "Multiple WiFi Access Points Co-Localization Through Joint AoA Estimation". In: *IEEE Transactions on Mobile Computing* (2023) (cit. on p. 32).

[53] Yao Ge, Jingyan Wang, Shibo Li, et al. "WiFi sensing of Human Activity Recognition using Continuous AoA-ToF Maps". In: *2023 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE. 2023, pp. 1–6 (cit. on p. 32).

[54] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. "Spotfi: Decimeter level localization using wifi". In: *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. 2015, pp. 269–282 (cit. on p. 32).

[55] Deepa Jeevaraj et al. "Feature Selection Model using Naive Bayes ML Algorithm for WSN Intrusion Detection System". In: *International journal of electrical and computer engineering systems* 14.2 (2023), pp. 179–185 (cit. on p. 32).

[56] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000 (cit. on p. 32).

[57] Sakorn Mekruksavanich, Wikanda Phaphan, Narit Hnoohom, and Anuchit Jitpattanakul. "Attention-based hybrid deep learning network for human activity recognition using WiFi channel state information". In: *Applied Sciences* 13.15 (2023), p. 8884 (cit. on p. 34).

[58] Ankan Dash, Junyi Ye, and Guiling Wang. "A review of Generative Adversarial Networks (GANs) and its applications in a wide variety of disciplines: From Medical to Remote Sensing". In: *IEEE Access* (2023) (cit. on p. 34).

[59] Fangxin Wang, Wei Gong, and Jiangchuan Liu. "On spatial diversity in WiFi-based human activity recognition: A deep learning-based approach". In: *IEEE Internet of Things Journal* 6.2 (2018), pp. 2035–2047 (cit. on p. 34).

[60] Dazhuo Wang, Jianfei Yang, Wei Cui, Lihua Xie, and Sumei Sun. "Multimodal CSI-based human activity recognition using GANs". In: *IEEE Internet of Things Journal* 8.24 (2021), pp. 17345–17355 (cit. on pp. 34, 39).

[61] Rui Gao, Xingsong Hou, Jie Qin, et al. "Zero-VAE-GAN: Generating unseen features for generalized and transductive zero-shot learning". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 3665–3680 (cit. on p. 35).

[62] Jianfei Yang, Xinyan Chen, Han Zou, et al. "EfficientFi: Toward large-scale lightweight WiFi sensing via CSI compression". In: *IEEE Internet of Things Journal* 9.15 (2022), pp. 13086–13095 (cit. on p. 35).

[63] Minseuk Kim, Dongsoo Han, and June-Koo Kevin Rhee. “Multiview variational deep learning with application to practical indoor localization”. In: *IEEE Internet of Things Journal* 8.15 (2021), pp. 12375–12383 (cit. on p. 35).

[64] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. “Generalizing from a few examples: A survey on few-shot learning”. In: *ACM computing surveys (csur)* 53.3 (2020), pp. 1–34 (cit. on p. 35).

[65] Jianfei Yang, Han Zou, Yuxun Zhou, and Lihua Xie. “Learning gestures from WiFi: A Siamese recurrent convolutional architecture”. In: *IEEE Internet of Things Journal* 6.6 (2019), pp. 10763–10772 (cit. on p. 35).

[66] Zain Ul Abiden Akhtar, Hafiz Faiz Rasool, Muhammad Asif, Wali Ullah Khan, Md Ali, et al. “Driver’s face pose estimation using fine-grained Wi-Fi signals for next-generation Internet of Vehicles”. In: *Wireless Communications and Mobile Computing* 2022 (2022) (cit. on pp. 36, 39).

[67] Yiming Liu and Shin’ichi Konomi. “WiHead: WiFi-Based Head-Pose Estimation”. In: *International Conference on Human-Computer Interaction*. Springer. 2022, pp. 69–86 (cit. on pp. 36, 39).

[68] Vijay Kumar Singh, Pragma Kar, Ayush Madhan Sohini, et al. “Monitoring Engagement in Online Classes Through WiFi CSI”. In: *2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS)*. IEEE. 2023, pp. 462–465 (cit. on pp. 36, 39).

[69] Sameera Palipana, David Rojas, Piyush Agrawal, and Dirk Pesch. “FallDeFi: Ubiquitous fall detection using commodity Wi-Fi devices”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.4 (2018), pp. 1–25 (cit. on p. 36).

[70] Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon, and Shahrokh Valaee. “A survey on behavior recognition using WiFi channel state information”. In: *IEEE Communications Magazine* 55.10 (2017), pp. 98–104 (cit. on pp. 37, 38, 46, 48, 49).

[71] Yi Zhang, Yue Zheng, Kun Qian, et al. “Widar3. 0: Zero-effort cross-domain gesture recognition with wi-fi”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.11 (2021), pp. 8671–8688 (cit. on pp. 37, 38).

[72] Jianfei Yang, Xinyan Chen, Han Zou, Dazhuo Wang, and Lihua Xie. “AutoFi: Toward Automatic Wi-Fi Human Sensing via Geometric Self-Supervised Learning”. In: *IEEE Internet of Things Journal* 10.8 (2022), pp. 7416–7425 (cit. on p. 37).

[73] Yu Gu, Xiang Zhang, Yantong Wang, et al. “WiGRUNT: WiFi-enabled gesture recognition using dual-attention network”. In: *IEEE Transactions on Human-Machine Systems* 52.4 (2022), pp. 736–746 (cit. on p. 38).

[74] Dazhuo Wang, Jianfei Yang, Wei Cui, Lihua Xie, and Sumei Sun. “CAUTION: A Robust WiFi-based human authentication system via few-shot open-set recognition”. In: *IEEE Internet of Things Journal* 9.18 (2022), pp. 17323–17333 (cit. on pp. 38, 39).

[75] Ruiyang Gao, Mi Zhang, Jie Zhang, et al. “Towards position-independent sensing for gesture recognition with Wi-Fi”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.2 (2021), pp. 1–28 (cit. on p. 38).

[76] Niloofar Bahadori, Jonathan Ashdown, and Francesco Restuccia. “ReWiS: Reliable Wi-Fi sensing through few-shot multi-antenna multi-receiver CSI learning”. In: *2022 IEEE 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. IEEE. 2022, pp. 50–59 (cit. on pp. 39, 68).

[77] Zhenguo Shi, Qingqing Cheng, J Andrew Zhang, and Richard Yi Da Xu. “Environment-robust WiFi-based human activity recognition using enhanced CSI and deep learning”. In: *IEEE Internet of Things Journal* 9.24 (2022), pp. 24643–24654 (cit. on pp. 39, 68).

[78] Francesca Meneghello, Domenico Garlisi, Nicolò Dal Fabbro, Ilenia Tinnirello, and Michele Rossi. “Sharp: Environment and person independent activity recognition with commodity ieee 802.11 access points”. In: *IEEE Transactions on Mobile Computing* (2022) (cit. on pp. 39, 69).

[79] Marwa RM Bastwesy, Kiichiro Kai, Hyuckjin Choi, Shigemi Ishida, and Yutaka Arakawa. “Wi-Nod: Head Nodding Recognition by Wi-Fi CSI Toward Communicative Support for Quadriplegics”. In: *2023 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE. 2023, pp. 1–6 (cit. on p. 39).

[80] Marwa RM Bastwesy, Hyuckjin Choi, and Yutaka Arakawa. “HeMoFi4Q: Morse Communication Based on Wi-Fi and Head Motion for Quadriplegia with Environmental Robustness”. In: *IEEE Access* (2023) (cit. on p. 39).

[81] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. “Detecting and recognizing human-object interactions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8359–8367 (cit. on p. 41).

[82] Yu Guan and Thomas Plötz. "Ensembles of deep lstm learners for activity recognition using wearables". In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1.2 (2017), pp. 1–28 (cit. on p. 42).

[83] Bo Wei, Kai Li, Chengwen Luo, et al. "No need of data pre-processing: A general framework for radio-based device-free context awareness". In: *ACM Transactions on Internet of Things* 2.4 (2021), pp. 1–26 (cit. on pp. 44, 46, 48, 49).

[84] Zhenghua Chen, Le Zhang, Chaoyang Jiang, Zhiguang Cao, and Wei Cui. "WiFi CSI based passive human activity recognition using attention based BLSTM". In: *IEEE Transactions on Mobile Computing* 18.11 (2018), pp. 2714–2724 (cit. on pp. 46, 48, 49).

[85] Antonio Gulli and Sujit Pal. *Deep learning with Keras*. Packt Publishing Ltd, 2017 (cit. on p. 46).

[86] Yu Gu, Jinhai Zhan, Yusheng Ji, et al. "MoSense: An RF-based motion detection system via off-the-shelf WiFi devices". In: *IEEE Internet of Things Journal* 4.6 (2017), pp. 2326–2341 (cit. on p. 52).

[87] Md Shaad Mahmud, Honggang Wang, AM Esfar-E-Alam, and Hua Fang. "A wireless health monitoring system using mobile phone accessories". In: *IEEE Internet of Things Journal* 4.6 (2017), pp. 2009–2018 (cit. on p. 52).

[88] Tianben Wang, Daqing Zhang, Leye Wang, et al. "Contactless respiration monitoring using ultrasound signal with off-the-shelf audio devices". In: *IEEE Internet of Things Journal* 6.2 (2018), pp. 2959–2973 (cit. on p. 52).

[89] Timothy T Roberts, Garrett R Leonard, and Daniel J Cepela. *Classifications in brief: American spinal injury association (ASIA) impairment scale*. 2017 (cit. on p. 52).

[90] Kai Niu, Fusang Zhang, Yuhang Jiang, et al. "WiMorse: A contactless morse code text input system using ambient WiFi signals". In: *IEEE Internet of Things Journal* 6.6 (2019), pp. 9993–10008 (cit. on p. 52).

[91] Shangqing Liu, Yanchao Zhao, Fanggang Xue, Bing Chen, and Xiang Chen. "DeepCount: Crowd counting with WiFi via deep learning". In: *arXiv preprint arXiv:1903.05316* (2019) (cit. on p. 75).

[92] Qilong Wang, Banggu Wu, Pengfei Zhu, et al. "ECA-Net: Efficient channel attention for deep convolutional neural networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 11534–11542 (cit. on p. 75).

[93] Min Lin, Qiang Chen, and Shuicheng Yan. “Network in network”. In: *arXiv preprint arXiv:1312.4400* (2013) (cit. on p. 75).

[94] Akihito Shimazu, Takeo Fujiwara, Noboru Iwata, et al. “Effects of work-family life support program on the work-family interface and mental health among Japanese dual-earner couples with a preschool child: A randomized controlled trial”. In: *Journal of occupational health* 65.1 (2023), e12397 (cit. on p. 93).

[95] Da-Yee Jeung, Changsoo Kim, and Sei-Jin Chang. “Emotional labor and burnout: A review of the literature”. In: *Yonsei medical journal* 59.2 (2018), pp. 187–193 (cit. on p. 93).

[96] Sriparna Saha, Shreyasi Datta, Amit Konar, and Ramadoss Janarthanan. “A study on emotion recognition from body gestures using Kinect sensor”. In: *2014 international conference on communication and signal processing*. IEEE. 2014, pp. 056–060 (cit. on p. 94).

[97] Muhammad Asif Razzaq, Jaehun Bang, Sunmoo Svenna Kang, and Sungyoung Lee. “UnSkEm: unobtrusive skeletal-based emotion recognition for user experience”. In: *2020 International Conference on Information Networking (ICOIN)*. IEEE. 2020, pp. 92–96 (cit. on p. 94).

[98] Jie Wei, Guanyu Hu, Xinyu Yang, Anh Tuan Luu, and Yizhuo Dong. “Learning facial expression and body gesture visual information for video emotion recognition”. In: *Expert Systems with Applications* 237 (2024), p. 121419 (cit. on p. 94).

[99] Son Thai Ly, Guee-Sang Lee, Soo-Hyung Kim, and Hyung-Jeong Yang. “Emotion recognition via body gesture: Deep learning model coupled with keyframe selection”. In: *Proceedings of the 2018 international conference on machine learning and machine intelligence*. 2018, pp. 27–31 (cit. on p. 94).

[100] N Beckmann, R Viga, A Dogangün, and A Grabmaier. “Measurement and analysis of local pulse transit time for emotion recognition”. In: *IEEE Sensors Journal* 19.17 (2019), pp. 7683–7692 (cit. on p. 94).

[101] Shantanu Pal, Subhas Mukhopadhyay, and Nagender Suryadevara. “Development and progress in sensors and technologies for human emotion recognition”. In: *Sensors* 21.16 (2021), p. 5554 (cit. on p. 94).

[102] Xuyu Wang, Lingjun Gao, and Shiwen Mao. “CSI phase fingerprinting for indoor localization with a deep learning approach”. In: *IEEE Internet of Things Journal* 3.6 (2016), pp. 1113–1123 (cit. on p. 99).

Appendix

Publish Works

Journal

1. M. R. M. Bastwesy, H. Choi and Y. Arakawa, "HeMoFi4Q: Morse Communication Based on Wi-Fi and Head Motion for Quadriplegia With Environmental Robustness," in IEEE Access, vol. 11, pp. 116384-116397, 2023, doi: 10.1109/ACCESS.2023.3326259. Networking (ICMU2023), 2023.

Conference

1. M. R. M. Bastwesy, K. Kai, H. Choi, S. Ishida and Y. Arakawa, "Wi-Nod: Head Nodding Recognition by Wi-Fi CSI Toward Communicative Support for Quadriplegics," 2023 IEEE Wireless Communications and Networking Conference (WCNC), Glasgow, United Kingdom, 2023, pp. 1-6, doi: 10.1109/WCNC55385.2023.10118666.

2. Marwa R. M. Bastwesy, Hyuckjin Choi, Yutaka Arakawa, "Tracking On-Desk Gestures Based on WiFi CSI on Low-Cost Microcontroller" Proceedings Article In: The 14th International Conference on Mobile Computing and Ubiquitous Networking (ICMU2023), 2023.