Leveraging AI and Radar for Non-Invasive Fine-Grained Human Activity Monitoring

Basim Alhumaily

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

Principal Supervisor
Dr Ahmed Zoha

Advisory Team
Prof Muhammad Ali Imran
Prof Sajjad Hussain
Dr Lina Mohjazi

James Watt School of Engineering
College of Science and Engineering
University of Glasgow

July 2026

Abstract

Human-centric monitoring demands sensing technologies that are non-invasive, privacy-preserving, and capable of operating reliably across diverse environments and motion scales. Conventional approaches based on wearable sensors suffer from compliance and long-term usability constraints, while vision-based systems raise privacy concerns and degrade under poor lighting or occlusion. Alternative RF-based modalities, such as Wi-Fi channel state information (CSI), offer privacy-preserving device-free sensing but remain sensitive to environmental dynamics and offer limited spatial resolution. In contrast, frequency-modulated continuous-wave (FMCW) radar provides a favourable balance of robustness, privacy preservation, and fine-motion sensitivity, making it well-suited to the human-monitoring use cases targeted in this thesis, namely, ambient assisted living and in-vehicle driver monitoring, across behavioural and physiological motion scales. This thesis positions radar-based sensing as a unified framework for large-scale and fine-grained human activity monitoring and vital signs estimation, and further examines how behavioural and physiological cues can be complemented through multimodal fusion in realistic settings. First, large-scale human activity monitoring is addressed using FMCW radar representation maps, where multiple radar signal-processing domains and deep learning architectures are systematically evaluated under subject-wise testing. Among the evaluated configurations, MobileNetV2 combined with short-time Fourier transform (STFT) representations achieves a recognition accuracy of 96.30% with the shortest inference time of 2.57 ms per sample, while requiring only 0.22 s to preprocess and generate each STFT-based time–frequency radar map, supporting real-time, low-power deployment. Second, the thesis advances fine-grained monitoring by analysing seated upper-body and head movements using mmWave FMCW radar. Generalisation to unseen subjects is initially limited to 60% accuracy; by applying the proposed data augmentation strategy, performance improves substantially to 91.32%, demonstrating robust cross-subject recognition under data-scarce conditions. In parallel, a motion-robust radar-based vital signs estimation framework is developed, which uses variational mode decomposition to suppress motion artefacts. Under non-stationary conditions, vital signs estimation performance improves by 40.59%, highlighting radar's ability to recover physiological dynamics despite concurrent movement. Finally, the thesis extends toward multimodal human state monitoring by integrating physiological heart rate with behavioural gaze metrics in real-world driving. Analysis of approximately 480 minutes of naturalistic driving data reveals clear scenario-dependent and subject-specific patterns, and a lightweight decision-level fusion algorithm identifies abnormal driving segments with improved robustness and interpretability compared with unimodal cues, demonstrating effective operation under real-world, unconstrained conditions.

Keywords

Human Activity Monitoring, Large-Scale, Fine-Grained Activity Monitoring, Physiological-Scale Monitoring, Radar Sensing, Data Scarcity, Data Augmentation, Transfer Learning, Multimodal Human Monitoring.

Acknowledgements

First and foremost, I express my deepest gratitude to Almighty Allah for granting me the strength, perseverance, and guidance necessary to complete this thesis. Without His blessings and mercy, this journey would not have been possible.

I would like to extend my sincere and profound appreciation to my supervisor, **Dr. Ahmed Zoha**, for his tremendous and continuous support throughout my doctoral research. His exceptional guidance, unwavering encouragement, and intellectual generosity have been instrumental in shaping my research direction. I am particularly grateful for his ability to identify insightful perspectives, propose innovative solutions, and motivate me to pursue excellence with confidence and rigour.

I am deeply thankful to my co-supervisors, **Prof. Muhammad Ali Imran**, **Prof. Sajjad Hussain**, and **Dr. Lina Mohjazi**, for their valuable guidance, constructive feedback, and continued support throughout my doctoral journey. Their expertise, strategic research perspectives, and thoughtful advice significantly shaped the direction, quality, and impact of this work. I am particularly grateful for their support in strengthening my research methodology, refining my academic writing, and improving the clarity of my publications and thesis.

I would like to acknowledge and thank my colleagues **Dr. Fahad Ayaz**, **Dr. Habib Ullah Manzoor**, and **Dr. Ahsan Raza Khan** for their collaboration, support, and teamwork in data collection, experimental design, and joint publications. Working with them has been both professionally rewarding and personally enriching.

Finally, my heartfelt thanks go to my family. I am deeply grateful to my parents for their unconditional support, sacrifices, and prayers, and for giving me the opportunity and freedom to pursue my studies abroad. Their belief in me has been a constant source of strength. A very special thanks goes to my beloved wife, **Maram**, for her endless patience, understanding, and selfless support throughout this journey. Her encouragement during long working hours and periods of absence has been invaluable. I am equally grateful to my daughters, **Norah**, **Ghadah**, and **Shaden**, for their love and understanding during the many times I was working late or unexpectedly away. This achievement is as much theirs as it is mine.

Declaration

University of Glasgow

College of Science & Engineering

Statement of Originality

Name: Basim Alhumaily
Registration Number: xxxxxxx

I certify that the thesis presented here for examination for a PhD degree in the University of Glasgow is solely my own work other than where I have indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it) and that the thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice.

I declare that the thesis does not include work forming part of a thesis presented successfully for another degree.

I declare that this thesis has been produced in accordance with the University of Glasgow's Code of Good Practice in Research.

I acknowledge that if any issues are raised regarding good research practice-based on review of the thesis, the examination may be postponed pending the outcome of any investigation of the issues.

Signature:
Date: 20/07/2026

Dedication

To my beloved parents,

my wife,

and my children,

whose love and support made this journey possible.

Contents

List of Tables

List of Figures

List of Publications

Journal

Published Papers

1. F. Ayaz, **B. Alhumaily**, S. Hussain, M. A. Imran, K. Arshad and K. Assaleh, and A. Zoha, “Radar Signal Processing and Its Impact on Deep Learning-Driven Human Activity Recognition,” in *Sensors*, vol. 25, no. 3, p. 724, January 2025.
   doi: https://doi.org/10.3390/s25030724

2. F. Ayaz, **B. Alhumaily**, Ahsan Raza Khan, S. Hussain, M. A. Imran, K. Arshad and K. Assaleh, and A. Zoha, “Benchmarking radar preprocessing techniques and transfer learning models for FMCW-based human activity recognition,” in *Elsevier Pervasive and Mobile Computing*, p. 102198, February 2026.
   doi: https://doi.org/10.1016/j.pmcj.2026.102198

To Be Submitted

1. F. Ayaz, **B. Alhumaily**, S. Hussain, M. A. Imran, and A. Zoha, “Edge-optimised, Privacy-Preserving Human Activity Recognition Using FMCW Radar,” *Elsevier Internet of Things*, 2026.

2. **B. Alhumaily**, F. Ayaz, L. Mohjazi, S. Hussain, and A. Zoha, “Multimodal Fusion of Heart Rate and Gaze Data for Real-Time Driver Monitoring in Naturalistic Driving,” *MDPI Vehicles*, 2026.

3. **B. Alhumaily**, F. Ayaz, L. Mohjazi, S. Hussain, and A. Zoha, “Upper Body Movements Recognition Using mmWave Radar,” *MDPI Sensors*, 2026.

Conference Proceedings

1. **B. Alhumaily**, F. Ayaz, L. Mohjazi, S. Hussain, and A. Zoha, “Non-Invasive Driver Activity Recognition using mmWave Radar,” in Proc. IEEE 5th International Conference

on Computing and Machine Intelligence (ICMI), King Faisal University, Al-Ahsa, Saudi Arabia, Apr. 2026, pp. 1–6, doi: https://doi.org/10.1109/ICMI68585.2026.11539895.

2. F. Ayaz, **B. Alhumaily**, S. Hussain, M. A. Imran, and A. Zoha, "VMD integrated FMCW radar system for non-invasive pediatric vital sign estimation," in Proc. 28th IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Edinburgh, United Kingdom, Oct. 2023, pp. 264–269, doi: https://doi.org/10.1109/CAMAD59638.2023.10478425.

3. F. Ayaz, **B. Alhumaily**, S. Hussain, L. Mohjazi, M. A. Imran, and A. Zoha, "Integrating millimeter-wave FMCW radar for investigating multi-height vital sign monitoring," in Proc. IEEE Wireless Communications and Networking Conference (WCNC), Dubai, United Arab Emirates, Mar. 2024, pp. 1–6,
doi: https://doi.org/10.1109/WCNC57260.2024.10570899.

For multi-author publications in which I am not the first author, my contribution was substantial and included thesis-related experimental analysis, interpretation of radar representations, preparation of results, manuscript drafting, and integration of the findings into the relevant thesis chapters.

List of Abbreviations

AAL	Ambient Assisted Living
ADC	Analogue to Digital Converter
ADAM	Adaptive Moment Estimation
AI	Artificial Intelligence
BPM	Beat Per Minute
BrPM	Breath Per Minute
CAE	Convolutional Auto Encoder
CNN	Convolutional Neural Network
CSI	Channel State Information
CW	Continous Wave
DFT	Discrete Fourier Transform
ECG	Electrocardiogram
FFT	Fast Fourier Transform
FMCW	Frequency Modulated Continuous Wave
FN	False Negative
FP	False Positive
FT	Fine Tuning
GHz	Giga Hertz
GPU	Graphics Processing Unit
HAM	Human Activity Monitoring
HPF	High Pass Filter
HR	Heart Rate
I	In-phase (component)
IF	Intermediate Frequency
IMF	Intrinsic Mode Function
KNN	K-Nearest Neighbor
LPF	Low Pass Filter
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MCU	Microcontroller Unit

MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
mmWave	Millimetre Wave
MRP	Model Representation Pair
MTI	Moving Target Indicator
PIR	Passive Infrared
Q	Quadrature (component)
RD	Range Doppler
RF	Radio Frequency
RMSE	Root Mean Squared Error
RR	Respiration Rate
RSSI	Received Signal Strength Indicator
SPWVD	Smoothed Pseudo Wigner–Ville Distribution
STD	Standard Deviation
SVM	Support Vector Machine
TF	Time Frequency
TN	True Negatives
TP	True Positives
TR	Time Range
UWB	Ultra-wideband
USRP	Universal Software Radio Peripheral
WHO	World Health Organisation

List of Symbols

B	Bandwidth
T_c	Chirp duration
f_c	Carrier (starting/center) frequency
$f(t)$	Instantaneous frequency of the chirp
S	Chirp slope (Hz/s)
$s_T(t)$	Complex baseband transmitted chirp signal
A_T	Amplitude of the transmitted signal
M	Number of chirps in a frame
T_r	Chirp repetition interval
T_{idle}	Inter-chirp idle time
t	Absolute time
m	Chirp index
τ	Fast-time variable within a chirp
$u(\cdot)$	Unit step function
$R(t)$	Time-varying target range
R_0	Initial range at the beginning of the frame
v	Radial velocity
c	Speed of light
$\tau(t)$	Two-way propagation delay
τ_0	Approx. constant delay within one chirp
α	Complex reflection/propagation coefficient
$s_R(t)$	Received signal at the antenna
P	Number of point targets/scatterers
$R_p(t)$	Range of target/scatterer p
v_p	Radial velocity of target/scatterer p
α_p	Complex amplitude of target/scatterer p
$s_{mix}(t)$	Mixer output signal
$(\cdot)^*$	Complex conjugation
f_b	Beat frequency
f_D	Doppler frequency shift

λ	Wavelength
F_{s}	ADC sampling frequency
N	Number of fast-time ADC samples per chirp
T_{s}	Sampling interval
n	Fast-time sample index
$x[n,m]$	Complex baseband ADC signal
A	Complex amplitude for a single target in discrete model
A_p	Complex amplitude for target p in discrete model
k	Range-bin index
Δf	Fast-time FFT bin spacing
R	Target range estimate
ΔR_{th}	Theoretical range resolution
R_{max}	Maximum unambiguous range
k_{R}	Fixed range-bin index for Doppler processing
ℓ	Doppler-bin index
Δf_{D}	Doppler frequency bin spacing
Δv	Velocity resolution
v_{max}	Maximum unambiguous radial velocity
$u_k(t)$	k-th IMF
K	Number of IMFs
$\mathscr{I}_{\mathrm{RR}}$	Index set of IMFs retained for RR
$\mathscr{I}_{\mathrm{HR}}$	Index set of IMFs retained for HR
$\bar{X}_i$	Sample mean for scenario i
p	p-value
HR_{mean}	Mean heart rate
BD_{mean}	Mean blink duration
FD_{mean}	Mean fixation duration
SD_{mean}	Mean saccade duration
SA_{mean}	Mean saccade amplitude
SV_{mean}	Mean saccade velocity
w_{HR}	Weight for heart-rate metric
w_{Blink}	Weight for blink metric
w_{Fixation}	Weight for fixation metric
w_{Saccade}	Weight for saccade metric
S_{total}	Total weighted segment score
$I\{\cdot\}$	Indicator function
θ	Decision threshold

Chapter 1

Introduction

1.1 Background

Human Activity Monitoring (HAM) has moved from niche prototypes to a cornerstone capability for health, safety, assisted living, and human–computer interaction (HCI) [1]. Demographic and societal pressures make continuous, unobtrusive monitoring a public-health and infrastructure priority: the World Health Organisation reports that the share of people aged 60+ will rise from about one in eight in 2015 to nearly one in five by 2050, reshaping demand for long-term care and remote monitoring in homes and clinics [2]. The adage "Prevention is better than cure" underscores the importance of continuous monitoring for both healthy individuals and those at risk, as early detection through continuous care can prevent serious, high-cost health complications. Within Ambient Assisted Living (AAL), artificial intelligence (AI)-enabled sensing is positioned to support independence and early intervention [3, 4]. In parallel, transport safety has become a key area of monitoring as time spent in vehicles and the number of vehicles worldwide increase [5]. Drowsiness and inattention remain material risk factors. National Highway Traffic Safety Administration (NHTSA) estimates that in 2017, drowsy driving contributed to about 91,000 police-reported crashes and nearly 800 deaths in the United States [6, 7], underscoring the need for reliable, real-time in-cabin assessment.

To address these critical needs, a wide range of monitoring technologies has been developed, traditionally falling into three main categories: vision-based systems, contact-based wearables, and ambient sensors. In the commercial sector, these approaches are widespread. For example, in healthcare and AAL, solutions often rely on installed ambient sensors, such as the Passive Infrared (PIR)-based systems from providers like Tunstall [8], or contact-based wearables like the Apple Watch, which includes fall detection and ECG monitoring capabilities [9–11]. Vehicle-based performance metrics track indicators such as lane position and steering wheel movements [12]. Behavioural indicators include facial expressions and eye movements [13, 14]. Physiological data involves monitoring Heart Rate (HR) and other vital signs [15]. Moreover, camera-based driver monitoring systems from companies such as Seeing Machines and Mobileye are standard in modern vehicles [13, 14].

However, these traditional monitoring approaches suffer from persistent limitations that hinder their suitability for long-term and fine-grained sensing in real-world environments:

- **Camera-based:** Vision-based systems, while information-rich, are fundamentally unreliable in dynamic, real-world settings. Their performance degrades significantly in low-light conditions or when the subject is occluded by furniture or blankets [3, 16]. Furthermore, persistent social and ethical privacy barriers make them unsuitable for continuous monitoring in private spaces, such as bedrooms or care homes, thereby limiting user acceptance [17].

- **Contact-based wearable:** Body-worn sensors, such as smartwatches and chest straps, provide high-quality physiological data; however, their accuracy depends on user compliance. This "user burden" includes the need for daily charging, the potential for skin irritation, and the discomfort of continuous wear, particularly for older adults [17]. This often results in incomplete data, thereby defeating the purpose of continuous monitoring.

- **Ambient and infrastructure:** This category includes sensors such as PIR motion detectors, door contacts, and smart meters. While non-invasive, they require substantial infrastructure, with sensors often requiring hard-wiring or calibration for each room [3]. More importantly, they offer only coarse, indirect measurements, such as "a person is in the room," which lack the fine-grained detail needed to distinguish between specific activities or detect urgent events, such as falls [17].

With advancements in wireless technology, we are witnessing a shift from intrusive, contact-based methods to wireless, non-contact systems [18]. Among these, radio-frequency (RF)-based monitoring technologies have emerged as promising alternatives. Within the spectrum of RF-based solutions, radar technology stands out for its privacy-preserving, light-independent approach that eliminates the need for direct skin contact [19].

1.2 Motivation

Radar, and in particular, Frequency-Modulated Continuous Wave (FMCW) radar, offers a lightweight, environmentally independent, non-invasive, and contactless sensor. It can detect objects behind obstacles and measure multiple metrics [20], making it a compact, possibly multimodal system. Unlike other RF technologies, such as Wi-Fi, radar systems can directly provide spectrograms of motion and physiological signals using a single device [21]. Millimetre-wave (mmWave) radar, in particular, offers significant advantages, including a compact size [22], higher resolution, and lower interference than radar operating at other frequencies. These characteristics make radar-based monitoring especially suitable for various applications, such as in-vehicle monitoring [23], human motion [24], and fall detection [25].

For even more advanced applications, radar has demonstrated the ability to monitor a range of physiological metrics, including HR [26], Respiratory Rate (RR) [27], blood glucose levels [28], and blood pressure [29]. Moreover, the evolution of AI and compact sensors has enabled the development of intelligent, multimodal monitoring frameworks that integrate these physiological signals into broader health assessment systems, aligning with the trend toward personalised and decentralised healthcare [30], human activity recognition [31], and multimodal integration [32] for improved decision-making.

Moreover, road safety has been a research focus since the advent of vehicles. As mentioned earlier, several approaches have been explored. However, radar-based sensing technologies offer a compelling alternative for in-vehicle health monitoring. These systems can passively estimate vital signs without requiring physical contact or active user participation [33]. Compared with wearables, which may be uncomfortable or impractical for continuous use, radar offers a more seamless and privacy-preserving solution. When integrated with behavioural data inferred from upper-body and head movements, this approach enables a comprehensive understanding of driver state, supporting early detection of fatigue, distraction, or stress [34–36].

Assessing human behaviour, such as cognitive engagement and attention, is a critical area of research. Gaze tracking is a primary tool for this, but its reliance on contact-based equipment limits its broader application. In this regard, radar-based systems offer a non-intrusive alternative for monitoring head and fine-grained movements. However, this application presents new challenges compared to traditional HAM, which typically involve large, distinct motions. Monitoring a seated person requires detecting far subtler movements, whose low-amplitude Doppler signatures are easily obscured by noise. This requires careful signal processing to detect low-amplitude motion in constrained environments.

To systematically investigate radar as a comprehensive human monitoring platform, this thesis adopts a progressive approach across motion scales. It begins with large-scale human activities to establish robust signal representations and data-efficient AI methods for radar-based classification. Building on this foundation, the research then addresses the more challenging problem of fine-grained seated movements and motion-robust physiological sensing, culminating in a driver monitoring case study that integrates behavioural and physiological metrics. By grounding the investigation in real-world driving scenarios, this work evaluates radar not only as a general activity monitoring sensor but also as a practical tool for continuous, non-contact assessment of human state in safety-critical environments.

1.3 Problem Statement

Despite ongoing research into radar-based human monitoring systems, there are still three main challenges:

- **C1. Sensing and Interpreting Three Scales of Human Motion:** Radar's ability to sense motion across different scales is both an advantage and a significant challenge [37]. The data stream comprises a complex mix of signals that are difficult to separate and interpret using AI models. This problem exists across all scales:

1. At the **Large-Scale** (e.g., walking, falling): The high-dimensional radar signal can be represented in numerous ways, but it remains a challenge to find a representation that is both computationally efficient and highly discriminative for AI models.
2. At the **Fine-Grained Scale** (e.g., head turns, body leans): These subtle movements are critical for understanding human state, such as distraction and other behaviour, but they are often kinematically similar, resulting in highly ambiguous radar signatures that are difficult for AI models to interpret and classify.
3. At the **Physiological-Scale:** The extremely subtle radar reflections from HR and RR are easily overwhelmed and corrupted by signals from any other body movement (from large-scale walking to fine-grained head turns). These Motion Artefacts are a primary obstacle to reliable, non-contact vital sign monitoring in real-world, non-static conditions.

- **C2. Limited datasets in AI-driven Radar Sensing:** Modern Deep Learning models are data-hungry, requiring vast amounts of labelled examples to learn effectively. However, in radar-based human monitoring systems, collecting large amounts of high-quality, accurately labelled datasets is exceptionally difficult, expensive, and time-consuming. This is especially true for fine-grained movements (**C1.2**) or specific events, such as falls (**C1.1**). This fundamental problem of data scarcity is a primary bottleneck that hinders the training, performance, and generalisation of AI models for radar-based human monitoring systems.
- **C3. Comprehensive Human-State Assessment:** Sensing physical motion (**C1.1** & **C1.2**) or physiological signals (**C1.3**) in isolation provides an incomplete picture of a person's true cognitive or physical state [38]. For example, a high HR could indicate physical exertion, such as walking, or cognitive stress. A holistic understanding of an internal state, such as fatigue or cognitive load, requires integrating multiple, diverse data streams (i.e., behavioural and physiological).

1.4 Aims and Objectives

In response to these challenges, this thesis aims to:

- **Investigate whether a data-efficient AI-based radar sensing framework can interpret the three scales of human motion**, from large-scale body activities to subtle fine-grained movements, and examine how reliable physiological monitoring can be achieved under motion-induced signal corruption.

- **Evaluate data- and computationally-efficient AI methodologies for radar-based human monitoring**, jointly considering recognition accuracy, model latency, and preprocessing overhead, and understand how transfer learning and data augmentation mitigate limited radar datasets and enable robust training of deep learning models on specialised datasets.

- **Investigate multimodal approaches for comprehensive human-state assessment** by analysing how behavioural and physiological signals can be fused into a unified and interpretable decision framework capable of operating in real-time monitoring scenarios.

1.5 Thesis Contributions

The main contributions of this thesis are:

- The establishment of a data- and computationally-efficient AI pipeline for radar-based large-scale human activity monitoring [1], evaluated jointly on recognition accuracy, inference latency, and preprocessing cost. This contribution addresses the challenges of large-scale motion interpretation (**C1.1**) and limited dataset availability (**C2**). Radar echoes are transformed into three distinct two-dimensional signal representations: range–Fast Fourier Transform (FFT)-based Time–Range (TR) maps, Short-Time Fourier Transform (STFT)-based time–frequency (TF) maps, and high-resolution Smoothed Pseudo Wigner–Ville Distribution (SPWVD)-based TF maps. These radar representations vary significantly in precision and computational cost. An evaluation is conducted to determine their discriminative capacity and computational trade-offs. Moreover, to overcome data scarcity (**C2**) and prevent overfitting, this contribution applies transfer learning by evaluating four state-of-the-art, pre-trained convolutional neural network (CNN) architectures (VGG-16, VGG-19, ResNet-50, and MobileNetV2) on ImageNet [39]. A systematic comparison of the resulting 12 unique model-domain pairings is then performed, rigorously evaluating trade-offs among preprocessing complexity, feature richness, recognition accuracy, and computational efficiency, with a specific focus on preprocessing and inference times for real-time applications. This comprehensive analysis of the processing pipeline, from signal preprocessing to model inference, provides a straightforward methodology for developing a data-efficient pipeline that effectively mitigates the challenges posed by the complexity of radar data and data scarcity.

[1] In this thesis, *human activity monitoring* is used as the umbrella term for recognising observable human movement and behaviour. *Large-scale activity monitoring* refers to whole-body activities such as walking, sitting, standing, bending, drinking, and falling. *Fine-grained monitoring* refers to subtle seated head and upper-body movements. *Physiological-scale monitoring* refers to radar-based estimation of HR and RR. *Human-state monitoring* refers to the higher-level interpretation of behavioural and physiological cues, as examined in the multimodal driving study in Chapter 5.

- A framework for radar-based fine-grained human activity monitoring and motion-robust vital sign sensing of HR and RR. This contribution provides a two-part solution to the most challenging aspects of fine-grained motion interpretation (**C1.2**) and physiological sensing (**C1.3**). First, it addresses the classification of subtle, kinematically similar seated activities, including driver-related movements such as head nods, turns, and body leans. A single mmWave FMCW radar is employed to collect a dedicated fine-grained activity dataset, and a specialised signal processing pipeline based on the STFT is developed to extract discriminative TF representations from raw radar signals. To mitigate data scarcity (**C2**), a tailored data augmentation strategy is introduced. This method adapts *SpecAugment* [40], originally developed for audio processing, to STFT-derived radar TF maps. It applies frequency and time masking to generate a more robust and diverse training set. Second, to overcome motion artefact corruption (**C1.3**) during physiological sensing, this contribution integrates Variational Mode Decomposition (VMD) into the radar processing pipeline for vital sign estimation. This adaptive signal decomposition technique separates the radar signal into intrinsic modes, enabling effective suppression of motion-induced interference and isolation of the underlying physiological components. The proposed approach significantly improves radar-based vital sign estimation for both static and moving subjects.

- The development and validation of a real-time multimodal fusion framework for comprehensive human-state assessment. This contribution addresses the challenge of holistic state assessment (**C3**) by developing a multimodal monitoring framework that integrates behavioural and physiological indicators within a single, interpretable decision model. The framework combines HR, measured using a Polar H10 [41] reference sensor, with gaze-based behavioural metrics, such as fixation duration, saccade velocity, and blink frequency, captured through Pupil Labs Invisible [42] eye-tracking glasses. These signals are synchronised and processed within fixed 10-second analysis windows to ensure temporal consistency across modalities. A weighted decision-level fusion algorithm is then designed to convert heterogeneous input streams into a unified "alertness score" that provides insight into the subject's cognitive and physical state. The weighting coefficients are empirically tuned through scenario-specific and subject-level analyses, with sensitivity testing used to quantify the trade-off between missed detections and false alarms across different fusion thresholds. The system is validated using 480 minutes of real-world multimodal data collected from five participants during driving in urban and motorway environments, enabling evaluation of behavioural and physiological responses under real-world workload conditions. This contribution advances multimodal human monitoring with a real-time, interpretable, and resource-efficient fusion framework for continuous assessment in unconstrained settings.

These three contributions map directly onto the thesis chapters: Chapter 3 addresses efficient large-scale monitoring, Chapter 4 addresses fine-grained activity and motion-robust vital signs, and Chapter 5 addresses multimodal human-state assessment for driver monitoring.

1.6 Thesis Organisation

This thesis comprises six chapters that address the challenges and present the contributions. Below is a detailed description of the organisation of the thesis chapters.

Chapter 1: establishes the background and motivation, outlines the limitations of traditional sensing, and introduces radar for human monitoring. It then defines the three primary research challenges, sets out the aims and objectives, and outlines the three core contributions.

Chapter 2: provides a comprehensive literature review that situates radar-based monitoring within the broader evolution of human sensing technologies. It begins with a historical overview of HAM, tracing the progression from ambient infrastructure sensors, such as PIR detectors and environmental sensors, to vision-based systems and contact-based wearables. This discussion highlights the limitations of traditional modalities in privacy, user compliance, environmental sensitivity, and sensing granularity, thereby motivating a transition toward RF- and radar-based approaches. The chapter then reviews recent advances in radar-based human monitoring. It surveys state-of-the-art human activity recognition methods, including radar data representation, signal preprocessing techniques, and deep learning-based classification strategies. This is followed by an examination of contactless vital sign monitoring, outlining processing pipelines from subject detection and phase extraction to motion compensation, while critically discussing current technical challenges. Given that radar-based systems are often constrained by limited datasets (**C2**), the chapter further reviews data-efficient learning strategies, with an emphasis on transfer learning and AI-driven augmentation methods, and evaluates their applicability to radar sensing. Finally, it surveys multimodal fusion frameworks and their role in comprehensive human-state monitoring. The chapter concludes by identifying open research problems and synthesising the key gaps that motivate the contributions of this thesis.

Chapter 3: presents a data- and computationally-efficient AI pipeline for classifying large-scale human activities, directly addressing the challenge of effective radar-based motion representation at the large-scale level (**C1.1**) under limited dataset conditions (**C2**). Radar echoes are transformed into multiple two-dimensional signal representations and paired with both compact and deeper CNN architectures using TL. A systematic comparison of model–domain pairings is conducted under subject-wise evaluation to assess generalisation performance, while also reporting preprocessing complexity and inference latency. By jointly analysing recognition accuracy and computational cost, this chapter establishes practical guidelines for selecting radar representations and model architectures that satisfy real-time constraints while remaining robust and data-efficient.

Chapter 4: develops a framework for behavioural and physiological monitoring using mmWave

FMCW radar, addressing the challenges of fine-grained motion interpretation (**C1.2**) under limited dataset conditions (**C2**) and motion-corrupted physiological-scale sensing (**C1.3**). It first presents a signal processing and classification pipeline for seated upper-body and head movements, enabling discrimination among kinematically similar activities. To enhance generalisation with limited training data, a targeted augmentation strategy adapted from *SpecAugment* is introduced. The chapter then proposes a physiological sensing pipeline that integrates VMD into the radar processing chain to mitigate motion artefacts and stabilise vital-sign estimation during subject movement.

Chapter 5: explores a real-time multimodal framework that addresses the challenge of comprehensive human-state assessment (**C3**) by combining physiological and behavioural indicators into an interpretable state measure. HR is aligned with gaze-based behavioural metrics, including fixation duration, saccade dynamics, and blink features, using tunable analysis windows to ensure temporal consistency. A lightweight decision-level fusion algorithm converts the synchronised streams into a single alertness score with adjustable operating points. The framework is validated on real-world driving data through both scenario-wise and subject-wise analyses, demonstrating its applicability to dynamic, safety-critical environments.

Chapter 6: summarises the key findings of the thesis, revisits how the proposed contributions address the initial challenges, and reflects on the limitations of the current work. It concludes by outlining promising directions for future research and development in radar-based human monitoring.

Chapter 2

Literature Review

Human Activity Monitoring (HAM) has progressed from binary presence detection to systems that infer behavioural and physiological patterns from continuous sensor streams. Achieving reliable human-centric intelligence increasingly requires sensing modalities and analytical methods that remain robust across environments, subjects, and motion scales, while meeting usability and privacy constraints that govern real-world deployment.

Chapter 1 identified three core challenges that motivate a multi-scale investigation. **C1** spans three interconnected scales: large-scale activity recognition (**C1.1**), fine-grained motion monitoring (**C1.2**), and physiological sensing (**C1.3**). Beyond representational challenges, **C2** concerns the scarcity and heterogeneity of radar datasets, which constrain the generalisation of learning-based methods. In contrast, **C3** concerns the multimodal fusion strategies required to achieve reliable and interpretable human-state estimation.

This chapter addresses these challenges by reviewing the evolution of HAM research (section 2.1) and establishing the physical and signal-processing foundations that enable radar sensing at both large and fine-grained scales (section 2.2). It next surveys radar-based estimation of physiological dynamics and the dominant sources of motion artefacts (section 2.3). The chapter then examines AI and machine learning frameworks for radar, with emphasis on data scarcity, transfer learning, and augmentation strategies (section 2.4), before expanding to multimodal monitoring approaches for holistic human-state inference (section 2.5). It concludes by synthesising the remaining research gaps and the thesis positioning (section 2.6), explicitly linking them to **C1**, **C2**, and **C3**.

2.1 Evolution of Human Activity Monitoring

The quest for automated HAM did not originate with sophisticated deep learning models or high-resolution sensors, but rather with simple, utilitarian technologies designed for basic presence detection. During the late 1970s and throughout the 1980s, non-contact motion detection emerged as a foundational concept in security and building automation [43]. Early systems employed ultrasonic and microwave Doppler sensors that operated by illuminating the environment with acoustic or electromagnetic waves and detecting motion through phase or frequency shifts in the reflected signals [44].

Almost concurrently, PIR sensors entered the commercial landscape. By detecting changes in the infrared radiation emitted by human bodies, PIR sensors provided a cost-effective, energy-efficient solution for binary motion detection [45]. Their widespread adoption in intrusion alarms [46] and lighting control systems [46] demonstrated that reliable, non-contact human sensing was both technically and economically viable. Despite their limited functionality, constrained to distinguishing between motion and no motion, these early systems laid the conceptual groundwork for a non-intrusive form of human monitoring [45]. Even decades later, PIR sensors, along with simple ambient devices such as door contact switches and pressure mats, continued to be used in research and early smart home applications for recognising coarse activity patterns and occupancy transitions [47].

Over the past decade, however, the landscape of HAM has undergone a profound transformation [48]. This shift has been driven by the convergence of pervasive computing, ubiquitous sensing, and breakthroughs in machine learning, especially the advent of deep neural networks [49]. HAM systems are no longer limited to binary presence detection. Modern systems aim to infer not only what people are doing, but also how, when, and potentially why. To achieve these goals, researchers have explored a wide array of sensing modalities [50], each contributing unique capabilities. Contact-based sensors, such as accelerometers and gyroscopes embedded in smartphones or wearables [51], can capture fine-grained movement trajectories. Vision-based systems, leveraging cameras and computer vision techniques [52], provide rich semantic context regarding posture, gesture, and environmental interaction. Meanwhile, radio frequency (RF)-based approaches, including passive sensing through Wi-Fi or dedicated radar systems, offer the promise of privacy-preserving, through-wall, and lighting-invariant monitoring [47]. Figure 2.1 provides an illustration of modern intelligent human monitoring systems.

HAM has long been a central focus of ubiquitous computing [53]. The ability to accurately and opportunistically infer human activities, ranging from simple limb movements to complex behavioural sequences, is foundational to numerous applications. These include intelligent surveillance, wellness monitoring, personalised healthcare [54], and AAL [48]. Particularly for vulnerable populations such as older adults, robust HAM systems hold the promise of extending independent living, reducing healthcare costs through preventive care, and ensuring rapid response to adverse events such as falls [50].

Across application domains, an "ideal" HAM system is typically evaluated against four competing requirements:

- **High accuracy across motion scales:** The system should reliably recognise large-scale activities (e.g., walking, sitting, falling), fine-grained movements, and physiological signals, with performance that remains stable across subjects and environments.

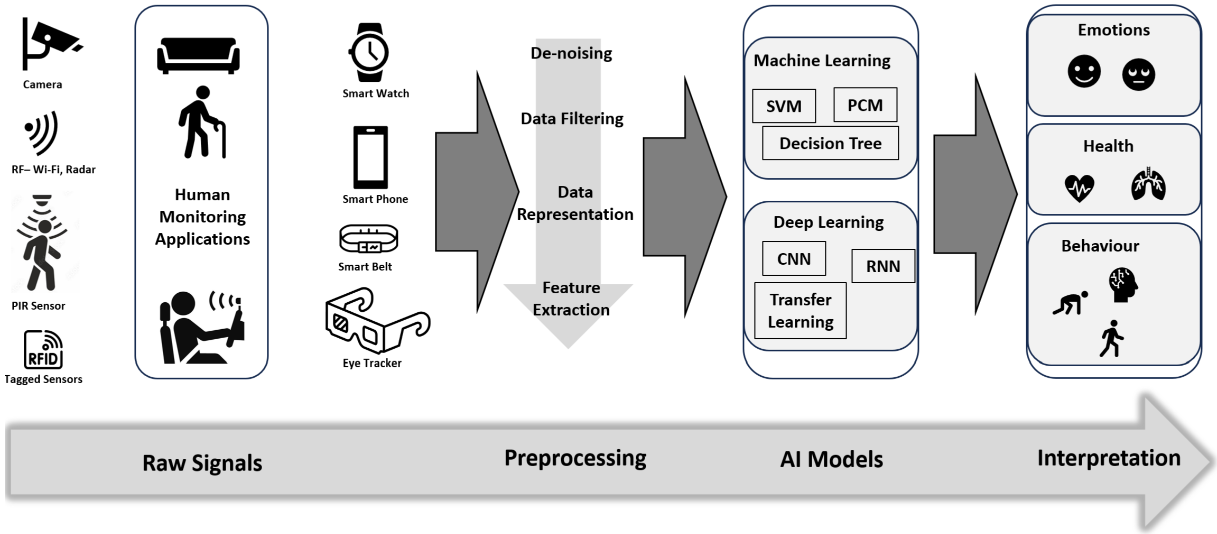


Figure 2.1: Overview of intelligent human monitoring systems, showing typical sensing modalities and inference pipelines for activity and physiological state estimation.

- **High user acceptance:** The solution should minimise behavioural burden, avoiding reliance on consistent wear, frequent charging, or user interaction so that it can support long-term monitoring with minimal compliance demands.
- **Strong privacy protection:** The sensing modality and processing pipeline should minimise capture of personally identifiable information, particularly in sensitive spaces such as bedrooms, bathrooms, and clinical settings.
- **Environmental robustness:** The system should operate reliably under uncontrolled conditions, including occlusion, lighting changes, clutter, and architectural variability.

In the subsequent sections, we analyse the evolution of the field across various technological paradigms. Starting with contact-based wearables, we move through vision-based systems, ultimately reaching RF-based sensing. Each modality is scrutinised for its effectiveness in addressing the outlined constraints. This examination goes beyond a historical overview by explicitly linking each sensing paradigm to the constraints above and showing how subsequent approaches emerged in response to specific limitations. Through iterative refinement, the field has consistently sought to identify a modality that balances accuracy, user acceptance, privacy, and robustness.

2.1.1 Contact-Based Sensing

The paradigm of human monitoring begins with a fundamental bifurcation: placing the sensor on the body or situating it within the surrounding environment. This initial distinction defines two dominant paradigms of sensing, namely contact-based and non-contact, and frames the series of trade-offs that the field has continually attempted to resolve [48, 52]. Contact-based sensing played a central role in early HAM systems because it directly captured human motion and physiology at the source [52].

Contact-based HAM relies on sensors affixed to or carried by the human body. These devices typically incorporate inertial measurement units (IMUs), which combine accelerometers and gyroscopes to measure movement and orientation [47, 55]. Smartphones, smartwatches, fitness bands, and specialised consumer wearables all include such sensors, making them common platforms for widespread activity recognition [51]. In addition to kinematic measurements, physiological signals can be obtained using devices that record electrocardiography (ECG), photoplethysmography (PPG), or skin-contact electrodes [52]. These signals enable the inference of activities linked to HR, RR, or stress levels.

During the past decade, the decreasing cost and increasing miniaturisation of these sensors have encouraged their adoption in fitness tracking, rehabilitation, workplace safety, and elderly care [10]. Their advantages are significant. They are self-contained, preserve privacy by avoiding image or audio capture, and are largely insensitive to lighting or visual obstructions [56].

Algorithmic Evolution and the Shift toward Deep Learning

The development of wearable HAM has progressed alongside advances in machine learning [19]. Early approaches used traditional classifiers such as support vector machines (SVMs) and random forests, trained on manually designed features extracted from the time or frequency domains [56]. With the rise of deep learning, the field shifted toward end-to-end models that learn features directly from raw sensor data [57]. Convolutional neural networks (CNNs) and recurrent architectures such as long short-term memory (LSTM) networks improved recognition accuracy by capturing both spatial and temporal structure within the data [51, 57].

The Burden of Wearability and Challenges of User Compliance

Despite robust technical performance, wearable-centric HAM systems encounter notable human-centred constraints. Primarily, user compliance remains a key challenge [58]. The utility of any wearable device depends on the user consistently and correctly wearing it. Missing data are common due to forgetfulness, discomfort, sensor detachment, or battery depletion [59]. These issues are especially problematic in long-term health monitoring, where consistent data collection is essential.

Wearable devices can also be uncomfortable or intrusive, particularly when multiple sensors are required. Even when worn properly, they provide information only from the specific body location where they are placed [53]. A wrist-mounted sensor may capture arm movements effectively but provide incomplete information about leg movements or interactions involving the trunk or head [1, 10]. This limitation in spatial coverage reduces the system's ability to infer complex activities.

There is also a well-documented tension between system performance and usability. Improving the fidelity of wearable HAM often requires higher sampling rates, additional sensors, or more complex processing [53, 60]. These improvements increase power consumption and reduce battery life [11]. For example, adding a gyroscope improves motion classification but consumes significantly more energy than accelerometer-only systems [61]. Short battery life has been identified as the most frequently cited reason for device abandonment and dissatisfaction [10, 62]. Thus, engineering enhancements that improve recognition accuracy often degrade the long-term user experience.

Evidence of Abandonment and the Limits of the Wearable Paradigm

User studies consistently demonstrate that long-term wearable adherence is fragile, with discontinuation driven by charging routines, device discomfort, technical reliability, and diminishing perceived value over time [59, 62–64]. These barriers are amplified for older adults in AAL contexts, where charging discipline, comfort, and stigma can directly reduce sustained usage [65]. As a result, even when wearable sensing is technically accurate, long-term monitoring performance is often constrained by human factors rather than signal quality.

In summary, contact-based sensing established the first major paradigm in human monitoring. It demonstrated that human motion and physiology can be measured directly and with high precision. However, significant challenges remain regarding user burden, comfort, and long-term compliance. These issues motivate the exploration of sensing approaches that eliminate the need for physical contact. The next step in the evolution of HAM places the sensing infrastructure in the environment, paving the way for non-contact monitoring.

2.1.2 Vision-Based Sensing

The second dominant paradigm in human monitoring emerged as a robust response to the inherent physical limitations of wearable sensors. By shifting the sensing responsibility from the user's body to the environment, the vision-based monitoring system effectively addresses issues of user compliance and device wearability [66]. Utilising RGB, depth-based, or thermal imaging, these systems passively observe the user to extract activity-related features from visual data [67]. Consequently, this approach is entirely device-free, requiring no active interaction or physical adherence from the user [68].

Technology and Algorithmic Evolution

Vision-based modalities have historically played a central role in computer vision research due to their ability to capture rich visual information about human posture, gestures, and environmental context [69]. Unlike wearable sensors that primarily provide inertial measurements, video sequences encode fine-grained spatiotemporal motion patterns across multiple body segments and capture complex human–object interactions [67]. This semantic richness makes vision-based sensing particularly effective for recognising intricate Activities of Daily Living (ADL), such as cooking, grooming, or assembly tasks, which are often difficult to infer from wearable sensors alone [70].

Early vision-based human activity recognition systems, particularly in the early 2000s, relied heavily on hand-crafted feature extraction techniques, including background subtraction, silhouette tracking, optical flow analysis, and spatiotemporal interest points [71]. While these approaches provided important foundations for activity modelling, they typically assumed static backgrounds and controlled camera viewpoints, limiting their robustness in real-world environments with dynamic scenes or camera motion [69]. Their performance was also sensitive to occlusion, illumination changes, and viewpoint variations, all of which frequently occur in practical deployments [67]. Moreover, the computational cost of processing raw video data often limits its applicability in real-time monitoring systems [72].

A major paradigm shift occurred in the mid-2010s with the emergence of deep learning. CNNs, originally designed for static image classification, were extended to video analysis through both two-dimensional and three-dimensional convolutional architectures capable of capturing spatial and temporal patterns [73]. At the same time, Recurrent Neural Networks (RNNs), particularly LSTM units, were introduced to model temporal dependencies inherent in sequential human actions [74]. These architectures enabled models to learn hierarchical representations directly from raw visual data, significantly reducing reliance on manually engineered features [49].

Consequently, modern vision-based HAM systems leverage deep neural networks and large datasets to achieve high recognition accuracy, particularly in controlled environments, and have demonstrated strong capability in detecting complex behaviours and detailed motion patterns.

Practical and Deployment Challenges of Vision-Based HAM

Despite substantial technical progress, vision-based HAM faces significant barriers to reliable real-world deployment. The most critical limitation is privacy. Video-based monitoring captures identifiable visual information and can record individuals in private or sensitive environments [75]. This concern is widely recognised as unacceptable in contexts such as in-home assisted living, healthcare facilities, and long-term behavioural monitoring [21, 76, 77]. Even privacy-oriented solutions highlight the severity of this issue. For instance, the PrivHAR framework [78] introduces hardware-level optical degradation to prevent cameras from capturing clear imagery, motivated by the risk that unprocessed visual data could be intercepted before anonymisation. This approach illustrates a fundamental tension: accurate activity recognition typically requires clear visual information, whereas privacy preservation requires obscuring it. Consequently, privacy-preserving modifications reduce identifiability but often also degrade the visual fidelity on which recognition pipelines depend, meaning the trade-off can only be mitigated rather than fully resolved.

Environmental robustness represents another major challenge. Vision-based systems are inherently sensitive to occlusion caused by furniture, other occupants, pets, or architectural barriers [79]. Variations in illumination, including shadows and low-light conditions, further degrade recognition performance [80]. In addition, background clutter and dynamic scenes can confuse models trained on simplified datasets [69]. As a result, performance achieved under controlled laboratory conditions often fails to generalise reliably to complex real-world environments [81, 82].

Beyond sensing limitations, vision-based systems impose substantial infrastructural and computational requirements that constrain scalability. Deploying such systems typically requires installing camera networks across the monitored environment, which increases both capital cost and logistical complexity related to installation, maintenance, and network configuration [66]. Continuous video acquisition also generates large volumes of data. Streaming high-resolution video requires considerable network bandwidth and significant storage, which can make cloud-centric processing impractical for long-term monitoring scenarios [83]. In addition, modern deep learning models for vision-based monitoring systems are computationally intensive, often requiring GPUs or specialised edge accelerators to perform real-time inference [84]. This reliance on high-performance hardware increases both operational cost and energy consumption, particularly in large-scale deployments [85].

In summary, vision-based HAM offers high recognition accuracy and enables rich interpretation of human activities without relying on wearable or contact sensors. However, its reliance on visual sensing introduces persistent challenges related to privacy, environmental robustness, and computational demands. These limitations are largely intrinsic to image-based sensing and therefore difficult to address solely through incremental improvements.

These constraints have motivated the exploration of alternative sensing paradigms that remain contactless while preserving privacy. RF-based sensing has emerged as a promising direction. Instead of capturing images, RF-based systems infer human activity by analysing variations in electromagnetic signals reflected from the body [86]. Such approaches do not record visual information, do not require wearable devices, and can operate under challenging lighting or partial occlusion conditions. The following section provides a more detailed review of RF-based human monitoring.

2.1.3 RF-Based Sensing

RF-based sensing has emerged as a compelling alternative to wearables and vision, particularly for privacy-sensitive, device-free monitoring. Compared with wearables, RF sensing reduces dependence on user compliance and can operate continuously without body-worn devices. Compared with cameras, RF sensing typically reduces privacy risk because it does not capture visually identifiable imagery, and it remains effective under poor illumination and partial visual occlusions [21, 80]. This paradigm includes approaches based on Wi-Fi channel state infor-

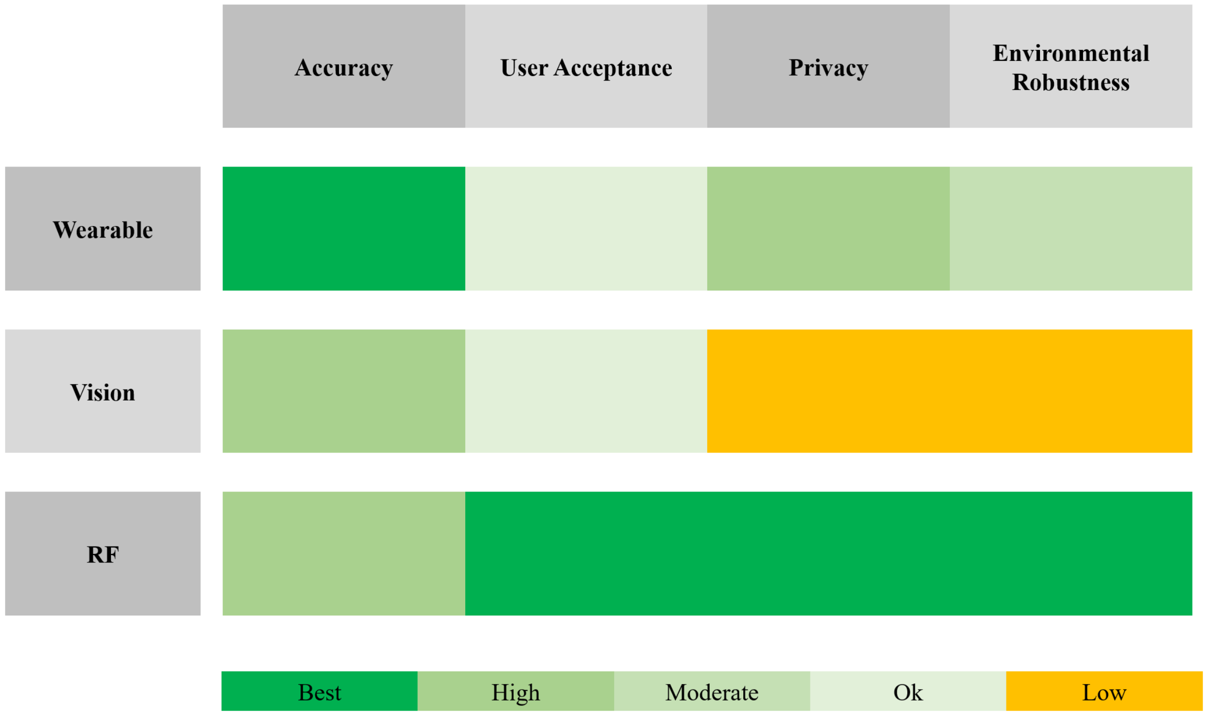


Figure 2.2: Qualitative comparison of contact-based, vision-based, and RF-based monitoring across four requirements: accuracy, user acceptance, privacy, and environmental robustness.

mation (CSI), research-grade platforms such as Universal Software Radio Peripheral (USRP), ultra-wideband (UWB) radar, and FMCW radars, each offering distinct trade-offs in spatial resolution, robustness, and implementation complexity.

Wi-Fi-Based Sensing Using CSI

Among RF approaches, Wi-Fi-based sensing has gained significant attention due to its cost-effectiveness and ubiquity. The CSI, obtained from the physical layer of Wi-Fi devices, captures the amplitude and phase of multiple subcarriers as a Wi-Fi signal traverses an indoor environment [87]. Human motion perturbs the multipath propagation of these signals, producing micro-dynamic variations in CSI measurements that can be mapped to activity labels using machine learning models [88].

Pioneering studies showed that off-the-shelf Wi-Fi routers could be repurposed as motion sensors, enabling recognition of actions such as walking, sitting, falling, and even specific gait patterns [89–91]. These systems often use signal processing techniques, such as the STFT, to construct time–frequency representations, which are then used to extract features for classifiers such as SVMs or deep learning models [89]. The appeal of this approach lies in its passive, device-free operation and its ability to operate without visual data, making it more acceptable for privacy-sensitive environments [92].

Despite these strengths, Wi-Fi sensing exhibits fundamental limitations that compromise its scalability and reliability. The CSI signature is highly sensitive to environmental changes, a phenomenon referred to as the multipath curse [55]. Even small alterations in the room layout, such as moving a chair or opening a door, can significantly change the CSI pattern for the same activity, leading to degraded recognition performance [93–95]. This environmental brittleness makes cross-location generalisation difficult and often requires site-specific training or recalibration. Moreover, commercial Wi-Fi hardware provides limited resolution due to the number of available subcarriers and the relatively long wavelength at 2.4 or 5 GHz, which restricts the granularity of motion detection [96, 97].

Research has also shown that Wi-Fi is inadequate for capturing fine-grained physiological motions. Its wavelength and signal structure limit sensitivity to sub-millimetre displacements, rendering it insufficient for medical-grade monitoring [93, 98]. Furthermore, the proliferation of Wi-Fi routers in real-world environments introduces interference and complexity, and the bulkiness of typical Wi-Fi access points may not suit embedded applications [98].

USRP and UWB Radar Systems

Beyond Wi-Fi, software-defined radios such as the USRP have enabled researchers to access full CSI data from more subcarriers and experiment with controlled RF signal designs [96, 99]. These platforms offer greater flexibility and resolution than commodity Wi-Fi, but at a higher cost and greater complexity. UWB radars, in particular, have shown promise due to their high time resolution and ability to detect fine movements such as limb gestures or falls [100]. However, their short range and susceptibility to hardware noise have posed challenges in practical deployments [86].

FMCW radar

Within RF-based modalities for device-free monitoring, radar is widely adopted when privacy, illumination invariance, and reduced user burden are primary requirements, and it has been applied extensively to ambient HAM, including fall detection and assisted-living scenarios [101–105]. Compared with other RF-based approaches such as Wi-Fi CSI, radar benefits from tighter control over the transmitted waveform, which can yield more stable motion signatures under matched conditions [21]. Among radars, FMCW has become a leading choice because it transmits a controlled waveform and provides range-resolved measurements that support clutter suppression and target isolation in indoor and in-cabin environments [106, 107]. These properties also enable operation under limited non-line-of-sight geometries, although performance degrades under severe attenuation and multipath [108]. In privacy-sensitive deployments, FMCW radar remains compatible with responsible sensing because it does not capture visually identifiable imagery, although data governance and threat modelling remain necessary [21, 76].

FMCW baseband processing yields structured representations that map naturally to both activity and physiological inference tasks. In this thesis, these are described using the terminology of Chapter 1, namely *time–range (TR)* maps and *time–frequency (TF)* maps computed using the

STFT and SPWVD. In the broader literature, the term *micro-Doppler* is often used to describe the underlying micro-motion Doppler phenomenon and its visualisation, which in our notation corresponds to TF maps derived from the slow-time signal, as will be detailed and explained later in this chapter.

Table 2.1: Comparison of RF-based sensing modalities for non-contact human monitoring, highlighting typical hardware, motion granularity, and key trade-offs.

RF Modality	Typical Hardware / Band	Motion or State Granularity	Key Advantages	Key Limitations and Typical Use
Wi-Fi	Commodity Wi-Fi access points and network cards, 2.4 and 5 GHz bands, extraction of CSI time series [109]	Medium granularity, whole body activities and some coarse gestures, limited vital sign capability in controlled conditions [110]	Reuses existing infrastructure, works through walls in many layouts, moderate deployment cost, good coverage in residential and office environments [111]	Sensitive to multipath and network traffic patterns, often environment-specific models and calibration are required, latency and reliability depend on communication load [112]
USRP	Software-defined radio front ends such as USRP devices with flexible carrier frequency and waveform [99]	From coarse to fine, depending on waveform and processing, can capture micro-Doppler of walking, arm swing and in some systems, breathing and subtle motion [96]	Very high flexibility in waveform design, carrier band and processing chain, ideal for rapid prototyping of new RF sensing concepts and custom passive radar architectures [96]	Higher hardware cost and power consumption, larger form factor, more complex synchronisation and calibration, mainly used in research testbeds rather than embedded products [99]
UWB radar	Short pulse ultra wide-band transceivers, typically 3–10 GHz with sub-nanosecond pulses and wide instantaneous bandwidth [113]	Fine range resolution, sensitive to small chest wall displacement, supports RR and HR monitoring in addition to gross movement and posture [114]	Good penetration through clothing and some obstacles, precise ranging, robust short-range vital sign sensing, suitable for indoor patient and occupancy monitoring [115]	Subject to regulatory limits on bandwidth and power, specialised hardware, usually short-range line of sight [113]
FMCW radar (non-mmWave)	C-band FMCW radars around 5.8 GHz with hundreds of MHz bandwidth, often single-input single-output [22]	Good micro-Doppler resolution of walking, sitting, hand gestures and continuous activities, basic vital sign extraction possible at short range [84]	Larger unambiguous range than mmWave for the same sweep parameters, less attenuation by clothing and some building materials, well-suited for large-scale HAM [22]	Coarser range and angular resolution than mmWave for a given antenna aperture, antennas physically larger, many implementations remain research prototypes rather than commercial modules [22]
mmWave FMCW radar	60–81 GHz FMCW radar with GHz-level bandwidth and compact MIMO arrays [116]	Very fine range and Doppler resolution, supports detailed micro-Doppler of limbs and head, as well as RR and HR, enabling fine-grained HAM and vital sign monitoring [117]	High spatial resolution with compact antennas compared with non-mmWave radars, already available in automotive and consumer radar chipsets [117]	Higher propagation loss and sensitivity to blockage than lower GHz bands, more demanding RF design and calibration, performance can degrade with severe self-occlusion or non-line-of-sight conditions [22]

Advances in radar-based algorithms and embedded implementations

Early radar-based HAM relied on classical signal processing and handcrafted feature extraction from spectrogram-like TF representations [24, 25]. Since 2016, there has been a surge in deep learning models applied to radar signals, paralleling the evolution seen in vision and wearable domains [19]. In many studies, CNNs are applied to TF maps as image-like inputs, while recurrent models such as LSTMs capture temporal dependencies in activity sequences [118].

Hybrid CNN–LSTM architectures have reported activity-classification accuracy beyond 90% in multiple scenarios [101, 119].

Modern radar HAM systems also increasingly target real-time and embedded deployment [120, 121]. As radar sensors, mmWave in particular, become smaller and more integrated into IoT platforms, there is growing interest in lightweight neural networks and efficient attention mechanisms that reduce computational load without sacrificing accuracy [19, 122]. While deep models can achieve strong recognition performance [58], their complexity often prevents real-time inference on embedded hardware [10, 123], motivating architecture optimisation and careful representation design.

Comparative performance

Comparative evaluations often report that FMCW radar achieves higher activity-recognition accuracy than Wi-Fi CSI under matched experimental conditions, largely due to stronger control over the transmitted waveform and improved motion and range resolution. For example, a study comparing a 77 GHz mmWave FMCW radar with a Wi-Fi CSI-based system for a seven-class activity task reported accuracies of 97.78% and 65.09%, respectively [21]. While such results depend on deployment geometry, interference, and dataset design, they reflect a recurring trend: FMCW radar sensing can yield more stable motion representations than CSI measurements in the same environment.

Physiological-scale monitoring with radar

Radar sensing extends beyond activity recognition to contactless estimation of physiological metrics. At short range, small periodic thoracic displacements caused by respiration and cardiac activity modulate the phase of the reflected signal, enabling estimation of RR and HR under appropriate conditions [124, 125]. This capability complements activity recognition in scenarios where both behavioural and physiological context are relevant. For example, in assisted-living settings, a radar system that detects a fall event may also support observation of respiratory and cardiac rhythms, provided that the subject remains within a stable range and motion artefacts are sufficiently controlled. Physiological monitoring is more sensitive to motion than activity recognition, and HR estimation typically degrades under posture changes, limb motion, or multi-person interference, motivating the dedicated treatment of vital sign processing in Section 2.3.

Synthesis and positioning within RF sensing

The progression from wearables to vision and then to RF sensing can be interpreted as an iterative attempt to balance four competing requirements: accuracy across motion scales, user acceptance, privacy, and robustness under real-world conditions. Wearables provide high-fidelity kinematic and physiological measurements but often face limitations in long-term adherence and body-location coverage [10,64]. Vision-based methods provide rich semantic context but remain constrained by privacy concerns, occlusions, lighting variation, and infrastructural and computational burden [69, 82]. RF-based methods such as Wi-Fi CSI reduce privacy risk and avoid illumination dependence, but their signatures are often sensitive to environmental change and interference, and typically offer limited motion granularity compared with radar systems [55,93,95].

Within RF-based modalities, FMCW radar offers a favourable balance for device-free monitoring because it actively controls the transmitted waveform and provides range-resolved measurements that support clutter suppression and target isolation [106, 107]. The motion induced by different body segments introduces time-varying Doppler modulations, commonly described as micro-Doppler in the literature, and captured in practice through TF maps [24, 64, 126]. At the same time, performance depends on placement, aspect angle, range, multipath, and interfering reflectors, and these effects can be amplified at mmWave frequencies where propagation loss and blockage may reduce signal quality under severe self-occlusion or non-line-of-sight conditions [22], which require a careful handling of its raw data and signal processing as will be detailed later in Section 2.2.

FMCW radar should not be interpreted as a universal replacement for wearable, vision-based, or Wi-Fi CSI sensing; rather, it is selected in this thesis because the targeted use cases require non-contact operation, reduced visual privacy exposure, robustness to illumination variation, and sensitivity to motion at both behavioural and physiological scales, whereas Wi-Fi CSI may be preferable for low-cost coarse device-free occupancy or activity sensing using existing communication infrastructure, vision-based sensing for detailed semantic interpretation where privacy constraints are manageable, and wearables for direct contact-based physiological measurement when user compliance is acceptable. By enabling range-resolved measurements, FMCW radar supports structured signal representations, including range profiles, TR maps, and TF maps derived using STFT and SPWVD. These representations support monitoring across large-scale and fine-grained motion, while phase-sensitive processing supports physiological estimation when motion artefacts are sufficiently mitigated. Figure 2.2 summarises qualitative trade-offs among contact-based, vision-based, and RF-based paradigms. The remainder of this chapter uses FMCW radar as a sensing foundation to address **C1.1** and **C1.2** through TR and TF representations, and **C1.3** through phase-sensitive vital sign processing under motion artefacts.

2.2 FMCW Radar Principles and Signal Model for Human Monitoring

Building on the modality-level comparison in the previous section, this section formalises the FMCW radar signal model and the baseband processing chain that produce the TR and TF (STFT/SPWVD) representations used throughout the thesis. Figure 2.3 illustrates the FMCW operating principle and the main processing blocks. FMCW radars achieve this by transmitting

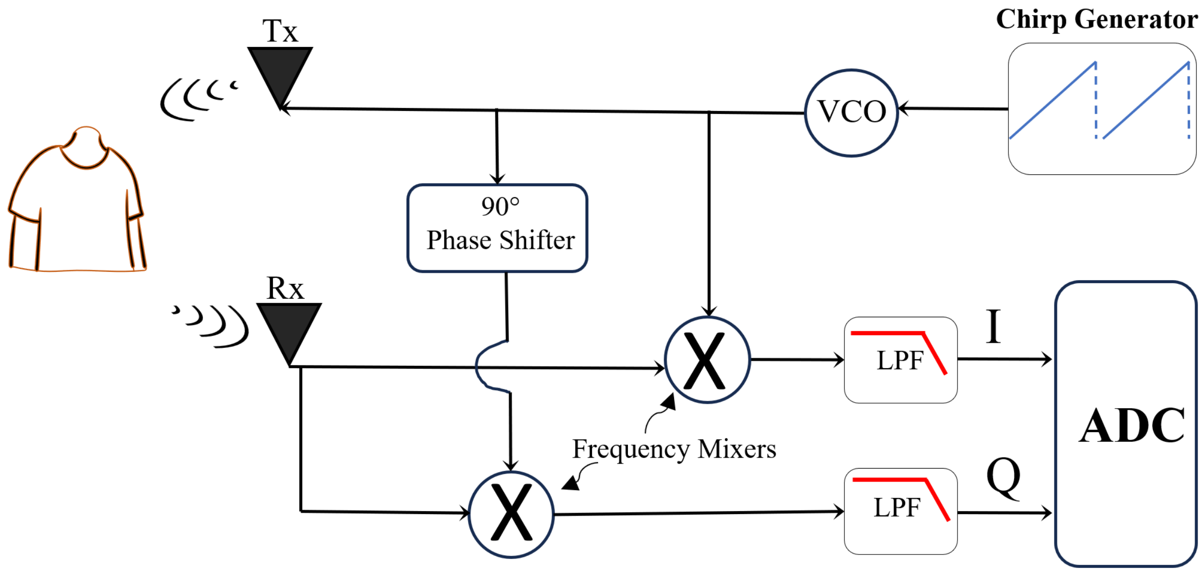


Figure 2.3: Working principle of an FMCW radar system, showing chirp generation, transmission, echo reception, and front-end processing to produce complex ADC samples.

linear chirps whose beat frequency encodes target range, enabling range-resolved motion characterisation that is not available in conventional continuous-wave Doppler radars [127]. This range–Doppler capability, together with compact solid-state integration and relatively low power consumption, makes FMCW suitable for real-time deployment in homes, clinics, vehicles, and mobile platforms [101].

Many FMCW devices operate at mmWave frequencies such as 24 GHz [128], 61 GHz [129], and 77 GHz [130]. These carrier frequencies support bandwidths from several hundred megahertz to multiple gigahertz, yielding fine range resolution that helps separate closely spaced body parts and capture subtle variations in motion [95, 120]. This capability depends directly on the underlying signal model and the processing steps used to extract range and TF information from the received echoes. The following subsection introduces the FMCW waveform structure and the baseband processing chain that form the basis of all subsequent radar representations used in this thesis.

2.2.1 FMCW Radar Signal Model and Baseband Processing

This subsection introduces the FMCW waveform and provides a mathematical description of the transmit and receive signals, followed by the essential baseband processing steps. The derivation culminates in the complex analogue-to-digital converted (ADC) signal, which represents the raw radar data used throughout this thesis for human monitoring across multiple motion scales. In particular, this signal serves as the starting point for large-scale activity recognition, fine-grained upper-body and head movement analysis, and radar-based physiological estimation.

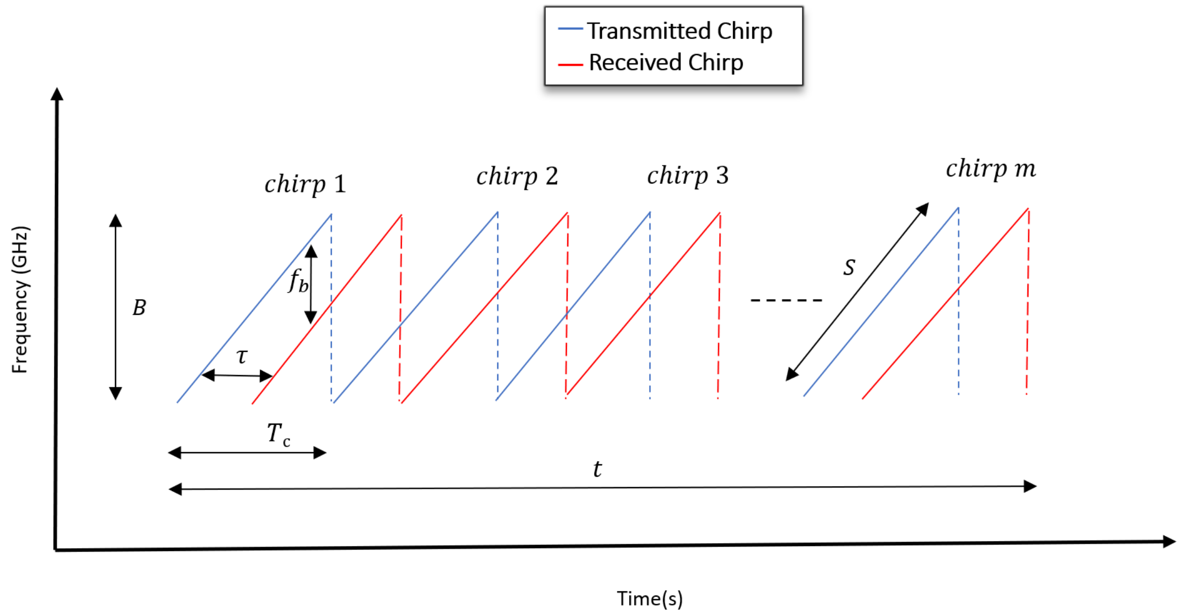


Figure 2.4: Illustration of an FMCW chirp sequence, where linear frequency modulation occurs over fast time, and chirps repeat over slow time.

FMCW Chirp Signal

FMCW radar estimates the range and radial velocity of targets by transmitting a sequence of linearly frequency-modulated chirps and processing the frequency difference between the transmitted and received signals [131–133]. Consider a single up-chirp of duration T_c and bandwidth B, as illustrated in Figure 2.4. The instantaneous frequency increases linearly with time from a starting (carrier) frequency f_c according to

$$f(t) = f_\mathrm{c} + St, \qquad 0 \le t < T_\mathrm{c}, \tag{2.1}$$

where

$$S = \frac{B}{T_\mathrm{c}} \tag{2.2}$$

is the chirp slope in Hz/s. The complex baseband representation of the transmitted chirp can be written as

$$s_\mathrm{T}(t) = A_\mathrm{T} \exp\left\{ j2\pi \left(f_\mathrm{c}t + \frac{S}{2}t^2 \right) \right\}, \qquad 0 \le t < T_\mathrm{c}, \tag{2.3}$$

where A_T is the amplitude of the transmitted signal. In practical radar operation, a frame (or coherent processing interval) consists of a sequence of M chirps repeated with a chirp repetition interval T_r (often $T_\mathrm{r} = T_\mathrm{c} + T_\mathrm{idle}$ where T_idle is an inter-chirp idle time). The absolute continuous time can be expressed as

$$t = mT_\mathrm{r} + \tau, \qquad m = 0, 1, \ldots, M-1,\ 0 \le \tau < T_\mathrm{c}, \tag{2.4}$$

where m is the chirp index, and τ is the fast-time variable within each chirp. The transmitted signal over one frame is then

$$s_{\mathrm{T}}(t) = \sum_{m=0}^{M-1} A_{\mathrm{T}} \exp\left\{ j2\pi \left(f_{\mathrm{c}}\tau + \frac{S}{2}\tau^2 \right) \right\} u(\tau)\, u(T_{\mathrm{c}} - \tau), \tag{2.5}$$

where $u(\cdot)$ is the unit step function enforcing $0 \leq \tau < T_{\mathrm{c}}$.

Propagation and Received Signal

For simplicity, consider a point target at a time-varying range $R(t)$ with radial velocity v relative to the radar. The two-way propagation delay is

$$\tau(t) = \frac{2R(t)}{c}, \tag{2.6}$$

where c is the speed of light. For slow motion within one chirp, the range is approximated as

$$R(t) \approx R_0 + vt, \tag{2.7}$$

where R_0 is the initial range at the beginning of the frame. The received signal at the antenna, after propagation and reflection with complex amplitude α, is a delayed and attenuated version of the transmit signal,

$$s_{\mathrm{R}}(t) = \alpha\, s_{\mathrm{T}}(t - \tau(t)). \tag{2.8}$$

Substituting (2.3) into (2.8) for a single chirp and writing t as the local fast-time τ for notational simplicity gives

$$s_{\mathrm{R}}(\tau) = \alpha A_{\mathrm{T}} \exp\left\{ j2\pi \left[f_{\mathrm{c}}(\tau - \tau_0) + \frac{S}{2}(\tau - \tau_0)^2 \right] \right\}, \tag{2.9}$$

where, within one chirp, we approximate the delay as constant $\tau(t) \approx \tau_0 = \frac{2R_0}{c}$ for short-range operation. This approximation applies within a single chirp; motion is then captured predominantly through the inter-chirp phase progression that produces the Doppler term in slow time. This approximation is commonly adopted in human monitoring because the target range variation within a single chirp is negligible for typical human motion and short chirp durations; motion is then captured primarily through the inter-chirp phase progression in slow time.

In realistic scenarios, there are multiple scatterers (e.g., different body parts and a static background). For P point targets with ranges $\{R_p\}$, velocities $\{v_p\}$ and complex amplitudes $\{\alpha_p\}$, the received signal becomes a sum of such components

$$s_{\mathrm{R}}(t) = \sum_{p=1}^{P} \alpha_p\, s_{\mathrm{T}}(t - \tau_p(t)), \qquad \tau_p(t) = \frac{2R_p(t)}{c}. \tag{2.10}$$

For notational clarity, the following derivation is first given for a single dominant target.

Homodyne Mixing and Low-Pass Filtering

FMCW radar front-ends typically use homodyne mixing, in which the received signal is mixed with a copy of the transmitted signal, as illustrated in Figure 2.3. The mixer output is

$$s_{\text{mix}}(t) = s_{\text{R}}(t)\, s_{\text{T}}^{*}(t), \tag{2.11}$$

where $(\cdot)^*$ denotes complex conjugation.

For a single target and within one chirp,

$$\begin{aligned} s_{\text{mix}}(\tau) &= \alpha A_{\text{T}}^2 \exp\left\{ j2\pi \left[f_{\text{c}}(\tau-\tau_0) + \frac{S}{2}(\tau-\tau_0)^2 \right] \right\} \exp\left\{ -j2\pi \left(f_{\text{c}}\tau + \frac{S}{2}\tau^2 \right) \right\} && (2.12)\\ &= \alpha A_{\text{T}}^2 \exp\left\{ j2\pi \left(-f_{\text{c}}\tau_0 - S\tau\tau_0 + \frac{S}{2}\tau_0^2 \right) \right\}. && (2.13) \end{aligned}$$

Neglecting constant phase terms that do not affect frequency content, the dominant term is

$$s_{\text{mix}}(\tau) \approx A_{\text{IF}} \exp\{ j2\pi f_{\text{b}} \tau \}, \qquad f_{\text{b}} = S\tau_0 = S\frac{2R_0}{c}, \tag{2.14}$$

where A_{IF} is a complex amplitude that absorbs constant phase terms. The signal $s_{\text{mix}}(\tau)$ contains a high-frequency component at approximately $2f_{\text{c}}$, which is removed by a low-pass filter (LPF). The remaining component is the intermediate frequency (IF) or beat signal, which is a complex sinusoid at beat frequency f_{b} proportional to the target range.

When the target has a non-zero radial velocity v, the range changes slightly from chirp to chirp. Writing the time as $t = mT_{\text{r}} + \tau$, and considering the slow-time (across chirps time as illustrated in Figure 2.5) index m explicitly, the received signal from a moving target introduces an additional phase rotation across chirps. After mixing and low-pass filtering, the IF signal becomes

$$s_{\text{IF}}(\tau, m) \approx A_{\text{IF}} \exp\{ j2\pi (f_{\text{b}}\tau + f_{\text{D}} m T_{\text{r}}) \}, \tag{2.15}$$

where

$$f_{\text{D}} = \frac{2v}{\lambda} \tag{2.16}$$

is the Doppler frequency shift, and $\lambda = c/f_{\text{c}}$ is the radar wavelength. The term f_{b} encodes the target range, while f_{D} introduces a linear phase progression along the chirp index m that encodes the radial velocity.

I/Q Demodulation and Complex ADC Signal

In many FMCW radar front-ends, the IF signal is split into in-phase (I) and quadrature (Q) components using two mixers driven by cosine and sine local oscillators that are phase shifted by 90° as shown in Figure 2.3. After mixing and low-pass filtering, the analogue I and Q components are sampled by ADCs at a sampling frequency F_{s}.

Let N denote the number of fast-time ADC samples per chirp. The sampling interval is

$$T_{\mathrm{s}} = \frac{1}{F_{\mathrm{s}}}, \qquad n = 0, 1, \ldots, N-1. \tag{2.17}$$

The sampled I and Q sequences for chirp m can be written as

$$I[n,m] = \Re\{s_{\mathrm{IF}}(nT_{\mathrm{s}}, m)\}, \tag{2.18}$$
$$Q[n,m] = \Im\{s_{\mathrm{IF}}(nT_{\mathrm{s}}, m)\}, \tag{2.19}$$

and the complex baseband ADC signal is

$$x[n,m] = I[n,m] + jQ[n,m]. \tag{2.20}$$

Using the model in (2.15), the discrete-time complex ADC samples for a single target are approximated as

$$x[n,m] \approx A \exp\{j2\pi\,(f_{\mathrm{b}} nT_{\mathrm{s}} + f_{\mathrm{D}} mT_{\mathrm{r}})\}, \tag{2.21}$$

where A is a complex amplitude capturing reflection strength and phase.

For P targets, the ADC signal becomes a superposition

$$x[n,m] = \sum_{p=1}^{P} A_p \exp\left\{j2\pi\left(f_{\mathrm{b},p} nT_{\mathrm{s}} + f_{\mathrm{D},p} mT_{\mathrm{r}}\right)\right\} + w[n,m], \tag{2.22}$$

where $f_{\mathrm{b},p}$ and $f_{\mathrm{D},p}$ denote the beat and Doppler frequencies of target p, and $w[n,m]$ captures noise and interference. The matrix $x[n,m] \in \mathbb{C}^{N\times M}$, illustrated in Figure 2.5, is the fundamental radar data used for further processing of radar-based human monitoring.

Range Estimation and Range Resolution

From (2.14), the beat frequency for a static target is

$$f_{\mathrm{b}} = S\frac{2R}{c}. \tag{2.23}$$

Solving for the range yields

$$R = \frac{c}{2S} f_{\mathrm{b}}. \tag{2.24}$$

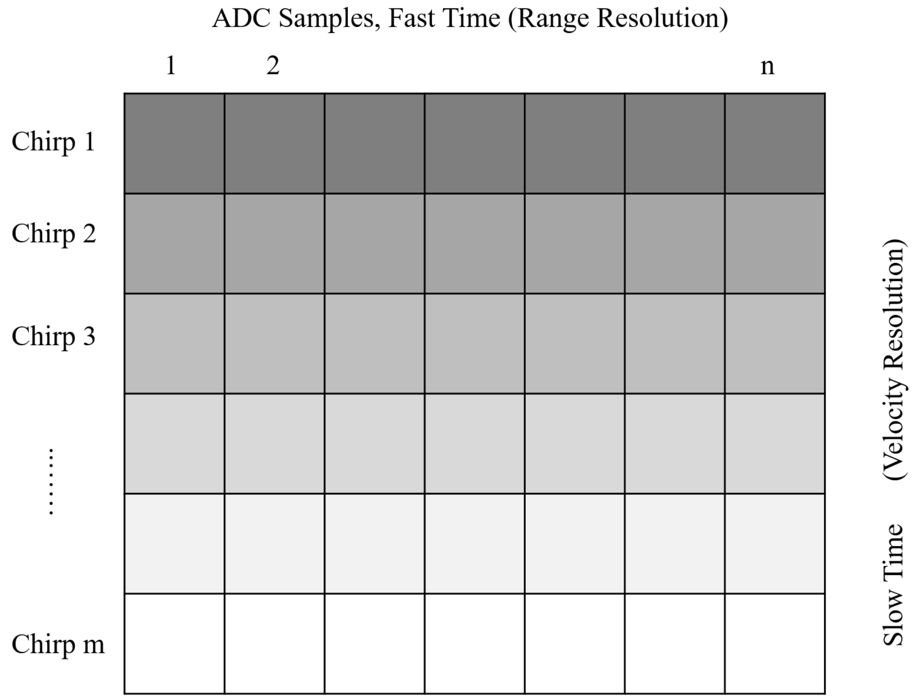


Figure 2.5: FMCW radar data cube represented as complex ADC samples, ready for subsequent preprocessing.

In practice, f_{b} is estimated by applying a discrete Fourier transform (DFT) or fast Fourier transform (FFT) over the fast-time index n for each chirp m:

$$X[k,m] = \sum_{n=0}^{N-1} x[n,m] \exp\left(-j\frac{2\pi kn}{N}\right), \qquad k = 0,1,\ldots,N-1. \tag{2.25}$$

The frequency spacing between adjacent FFT bins is

$$\Delta f = \frac{F_{\mathrm{s}}}{N}. \tag{2.26}$$

Combining this with (2.24), the range spacing between adjacent bins is

$$\Delta R = \frac{c}{2S}\Delta f = \frac{c}{2S}\frac{F_{\mathrm{s}}}{N}. \tag{2.27}$$

On the other hand, classical radar theory shows that the fundamental range resolution is determined by the transmitted bandwidth [19, 132]:

$$\Delta R_{\mathrm{th}} \approx \frac{c}{2B}. \tag{2.28}$$

For appropriately chosen F_{s} and N, the discrete bin spacing in (2.27) is consistent with the theoretical resolution (2.28).

The maximum unambiguous beat frequency is limited by the Nyquist frequency of the ADC:

$$|f_\mathrm{b}| \leq \frac{F_\mathrm{s}}{2}. \tag{2.29}$$

Using (2.24), this imposes a maximum unambiguous range

$$R_\mathrm{max} \approx \frac{c}{2S}\frac{F_\mathrm{s}}{2} = \frac{cF_\mathrm{s}}{4S}. \tag{2.30}$$

In radar-based human monitoring applications, R_max is often set to a few meters.

Velocity Estimation and Velocity Resolution

Radial velocity information is obtained from the slow-time dimension m of the ADC data. For a target at a fixed range bin (i.e., fixed f_b), the Doppler-induced phase factor in (2.21) is

$$\exp\{j2\pi f_\mathrm{D} m T_\mathrm{r}\}. \tag{2.31}$$

An FFT across chirps for a fixed range bin yields the Doppler spectrum:

$$X[k_\mathrm{R},\ell] = \sum_{m=0}^{M-1} X[k_\mathrm{R},m]\exp\left(-j\frac{2\pi\ell m}{M}\right), \qquad \ell = 0,1,\ldots,M-1, \tag{2.32}$$

where k_R is a fixed range-bin index. The Doppler frequency spacing between bins is

$$\Delta f_\mathrm{D} = \frac{1}{MT_\mathrm{r}}. \tag{2.33}$$

Using the relationship in (2.16), the corresponding velocity spacing (velocity resolution) is

$$\Delta v = \frac{\lambda}{2}\Delta f_\mathrm{D} = \frac{\lambda}{2MT_\mathrm{r}}. \tag{2.34}$$

The maximum unambiguous Doppler frequency is constrained by the Nyquist limit in slow time:

$$|f_\mathrm{D}| \leq \frac{1}{2T_\mathrm{r}}, \tag{2.35}$$

which implies a maximum unambiguous radial velocity

$$v_\mathrm{max} \approx \frac{\lambda}{4T_\mathrm{r}}. \tag{2.36}$$

In large-scale human monitoring (**C1.1**), the velocity resolution Δv and the maximum unambiguous velocity v_max determine how effectively the system can resolve walking speeds and limb-induced motions. For fine-grained monitoring (**C1.2**) and physiological sensing (**C1.3**), the target velocity is significantly smaller, and the corresponding Doppler signature manifests as slow phase variations rather than distinct Doppler peaks. In this regime, analysis focuses on phase evolution within one or a small number of range bins over time, rather than on classical Doppler FFT processing. These distinctions are examined in detail in Chapters 3 and 4.

Relevance to Radar-Based Human Monitoring

The complex ADC signal $x[n,m]$ defined in (2.22) forms the starting point for subsequent radar-based human monitoring tasks:

- **Large-scale and fine-grained monitoring:** The radar data cube $x[n,m]$ is commonly transformed into two-dimensional or three-dimensional representations, such as Range–Doppler (RD) maps, TF maps, and TR images. These representations exploit the range resolution in (2.28) and the velocity resolution in (2.34) to characterise large-scale movements, such as walking and limb motion, as well as fine-grained movements, including seated upper-body and head motions.

  In this work, large-scale activities are investigated using a 5.8 GHz FMCW radar dataset, while fine-grained upper-body and head movements are analysed using a custom-collected 77 GHz mmWave radar dataset. The benefit of higher carrier frequency for fine-grained motion sensing can be illustrated through a simple micro-motion example. Consider a radial velocity of $v = 1$ cm/s (0.01 m/s). At 5.8 GHz, corresponding to a wavelength of $\lambda \approx 5.17$ cm, the Doppler frequency, from (2.16), is

  $$f_{\mathrm{D}} = \frac{2v}{\lambda} \approx \frac{2 \cdot 0.01}{0.0517} \approx 0.39 \text{ Hz}.$$

  Such a small frequency shift lies close to the DC clutter region and is easily obscured by system noise and phase instability; however, it remains sufficient for large-scale motion monitoring where higher velocities and stronger Doppler components dominate the radar return. In contrast, at 77 GHz with $\lambda \approx 0.39$ cm, the same motion produces

  $$f_{\mathrm{D}} \approx \frac{2 \cdot 0.01}{0.0039} \approx 5.1 \text{ Hz},$$

  which is substantially more separable from static clutter and therefore more suitable for detecting subtle fine-grained movements.

- **Physiological-Scale Monitoring:** For HR and RR estimation, the subject is positioned at a short range and occupies a small number of dominant range bins. The phase and very low-frequency micro-Doppler components of $x[n,m]$ within these bins are analysed over long slow-time windows. Periodic chest wall displacements introduce phase modulation in the complex samples, from which respiratory and cardiac rhythms can be extracted using appropriate signal processing techniques, such as filtering, phase unwrapping, and decomposition-based methods. Accurate range bin selection, sampling rates, and a precise understanding of the relationship between motion and phase, all derived from the FMCW signal model, are essential for robust vital signs estimation.

In summary, FMCW radar encodes target range in the beat frequency f_b and radial velocity in the Doppler frequency f_D, both of which are inherently present in the complex ADC samples $x[n,m]$. This unified signal model forms the basis for designing signal processing pipelines that address large-scale activity recognition, fine-grained motion analysis, and physiological parameter estimation. Building upon this foundation, the following subsection details the radar signal processing techniques used to extract informative representations for HAM.

2.2.2 Radar Representations for Activity Monitoring

The FMCW signal model established in the previous section provides the theoretical foundation for radar-based HAM. After homodyne mixing and low-pass filtering, the received signal from a moving subject takes the form shown in (2.15), where the beat frequency f_b encodes target range through (2.14), and the Doppler frequency f_D captures radial velocity as defined in (2.16). Sampling this IF signal along the fast-time index n and the slow-time (chirp) index m yields the complex baseband measurements.

$$x[n,m] = \sum_{p=1}^{P} A_p \exp\left\{ j2\pi \left(f_{\mathrm{b},p} n T_\mathrm{s} + f_{\mathrm{D},p} m T_\mathrm{r} \right) \right\} + w[n,m], \tag{2.37}$$

which matches the multi-target formulation in (2.22). Here, $f_{\mathrm{b},p}$ and $f_{\mathrm{D},p}$ denote the beat and Doppler frequencies associated with reflector p, corresponding to a specific body segment. Using the relations

$$f_{\mathrm{b},p} = S\frac{2R_p}{c}, \qquad f_{\mathrm{D},p} = \frac{2v_p}{\lambda}, \tag{2.38}$$

each term in (2.37) directly reflects the instantaneous range R_p and radial velocity v_p of that segment.

Equation (2.37) shows that human motion manifests as structured variations in f_b across fast time and f_D across slow time. These variations form the basis of the canonical radar signal representations used in HAM. Applying a fast-time FFT along n converts $x[n,m]$ into range profiles according to (2.24). A subsequent Doppler FFT across chirps exploits the phase progression $\exp(j2\pi f_\mathrm{D} m T_\mathrm{r})$ to reveal velocity content. Finally, TF analysis along the slow-time dimension extracts micro-Doppler signatures, or TF representations as we call them in this thesis, that capture limb motion and other dynamic components of human behaviour.

Together, these transformations convert the FMCW measurement model into interpretable signal representations that underpin radar-based human monitoring across multiple motion scales, as summarised in Table 2.2 and illustrated conceptually in Figure 2.6.

Range processing. An FFT applied along the fast-time dimension converts each chirp into a range profile [125, 134]. Peaks in the resulting spectrum correspond to reflectors at distances given by

$$R = \frac{c}{2S} f_{\mathrm{b}}. \tag{2.39}$$

Stacking these profiles over consecutive chirps produces a TR map that visualises translational motion along the radial dimension.

Doppler processing. Velocity information is extracted by applying a second FFT across chirps for each range bin [64,122]. This Doppler FFT leverages the slow-time phase evolution in (2.37), producing an RD map in which one axis corresponds to range and the other to Doppler frequency, or equivalently radial velocity using $v = \frac{\lambda}{2} f_{\mathrm{D}}$ [123,127]. Moving body segments appear as localised energy concentrations whose Doppler shifts track their instantaneous motion [135].

Micro-Doppler signatures. Human motion typically involves multiple body parts moving with different velocities, which introduces fine-scale frequency modulations superimposed on the bulk Doppler shift [84, 126, 136]. Applying a TF analysis, such as STFT or SPWVD, along the slow-time dimension yields a micro-Doppler spectrogram that captures the temporal evolution of these motions [24, 137]. Periodic activities such as walking generate oscillatory sidebands, while static postures or slow gestures produce narrow-band or smoothly varying patterns [37, 138]. In this thesis, we refer to these techniques as TF representations, emphasising that two distinct time–frequency approaches are examined.

Common radar representations for HAM. A range of two-dimensional and three-dimensional radar representations can be derived from $x[n,m]$ and serve as inputs to HAM algorithms, including:

- **TR maps**, which capture translational motion over time.
- **RD maps**, which jointly encode location and radial velocity.
- **Angle-of-arrival maps**, obtained using multi-antenna arrays to resolve azimuth or elevation.
- **TF maps**, which reveal periodic or subtle motion components.

Each representation emphasises different aspects of human motion, and prior studies have shown that combining complementary representations can improve monitoring performance [104, 118, 122, 139]. Most representations rely on single-input single-output (SISO) configurations, whereas angle-related representations require multi-antenna or multiple-input multiple-output

(MIMO) systems. While MIMO enables angular discrimination, it substantially increases signal dimensionality and computational cost, which is a critical consideration for edge-deployed systems. Table 2.2 summarises these radar representations with representative references from the literature.

Table 2.2: Examples of radar signal representations for activity monitoring.

Radar Representation	**Learning Model**	**Reference**
TR Map	DCNN	[140, 141]
RD Map	3D-CNN	[106]
RA Map	MobileNetV2	[142]
TF Map (STFT)	1D-CNN + LSTM	[139]
TF Map (CWT)	Range-Distributed-Convolutional Neural Network (RD-CNN)	[126]
TF Map (SPWVD)	Mixed CNN	[143]

General HAM pipeline. Independent of the downstream learning model, radar-based HAM typically follows these processing stages:

1. Preprocessing of raw IF signal (ADC samples represented in (2.21)) to suppress noise and static clutter.
2. Fast-time FFT to estimate f_{b} and derive range profiles.
3. Slow-time Doppler FFT to estimate f_{D} at each range bin.
4. TF analysis, such as STFT, to extract micro-Doppler signatures.

The resulting multidimensional radar tensor, commonly organised as time $\times$ range $\times$ velocity and optionally angle, encodes characteristic structures associated with different actions. Alternating Doppler patterns from walking limbs, broadband bursts from falls, and localised signatures from subtle gestures emerge naturally from this representation [25, 144]. In this way, the FMCW signal model establishes a direct link between human motion and measurable modulations in f_{b} and f_{D}, enabling unified monitoring of large-scale and fine-grained activities.

Large-Scale Monitoring

FMCW radar is particularly well-suited for recognising activities that involve large-scale body motion. Gait analysis provides a clear example of this capability. During walking, the human torso produces a dominant radar reflection with relatively small Doppler variation, while the limbs generate periodic Doppler sidebands due to their cyclic motion. These oscillatory components arise from the inherent periodicity of human gait and appear clearly in TF and RD maps [53, 126, 145].

This micro-Doppler behaviour, seen in the TF representation maps, enables radar systems to distinguish between different motion patterns. Prior studies have demonstrated that radar signatures can differentiate between normal walking [138], limping gait [24], and walking while carrying objects [146]. In fall detection scenarios, a rapid decrease in range combined with a broadband Doppler burst produces a distinctive signature that allows reliable identification of fall events [25, 100, 116]. In smart-home monitoring environments, radar systems can also continuously identify coarse postural states, such as walking, sitting, or lying down, by analysing the stability and evolution of torso reflections in the range and Doppler domains [147].

Experimental studies consistently report strong classification performance for such large-scale activities. For example, a 25 GHz FMCW radar with a 2 GHz bandwidth achieved over 95% accuracy in distinguishing activities such as walking, sitting, and standing using spectrogram-based features [147]. Similarly, a 60 GHz FMCW radar system reported over 97% accuracy for identifying walking versus non-walking periods and more than 95% accuracy for multi-class activity recognition [148].

These results highlight how the underlying physics of FMCW radar, together with appropriate signal processing techniques, enable different body segments to generate distinct beat-frequency (f_{b}) and Doppler (f_{D}) signatures. As a result, radar offers a powerful, privacy-preserving sensing modality for large-scale human activity monitoring.

Fine-Grained Monitoring

FMCW radar, particularly in the mmWave band, can also detect fine-grained human movements due to its high range resolution and sensitivity to micro-Doppler modulation [131, 149]. Hand gesture recognition provides a clear demonstration of this capability [120]. High-frequency radar systems, such as those operating at 60 GHz or 77 GHz, can detect small finger motions that generate low-amplitude Doppler shifts and millimetre-scale range variations [121, 131]. These subtle movements appear as characteristic streaks or localised patterns in TF maps, enabling real-time gesture recognition [79, 123].

Academic studies have extended this capability to more complex motion patterns. For example, a 77 GHz FMCW radar has been used to classify American Sign Language (ASL) gestures by analysing the RD trajectories of hand and finger motion [86]. Despite the small displacement amplitudes involved, the resulting Doppler modulations are sufficiently distinctive for classification models. Initial research has also explored whether mmWave radar can capture extremely subtle facial and speech-related micro-motions under controlled experimental conditions, typically requiring strict constraints on range, orientation, and environmental interference [150, 151]. These results demonstrate feasibility; however, the robustness of such signals in unconstrained

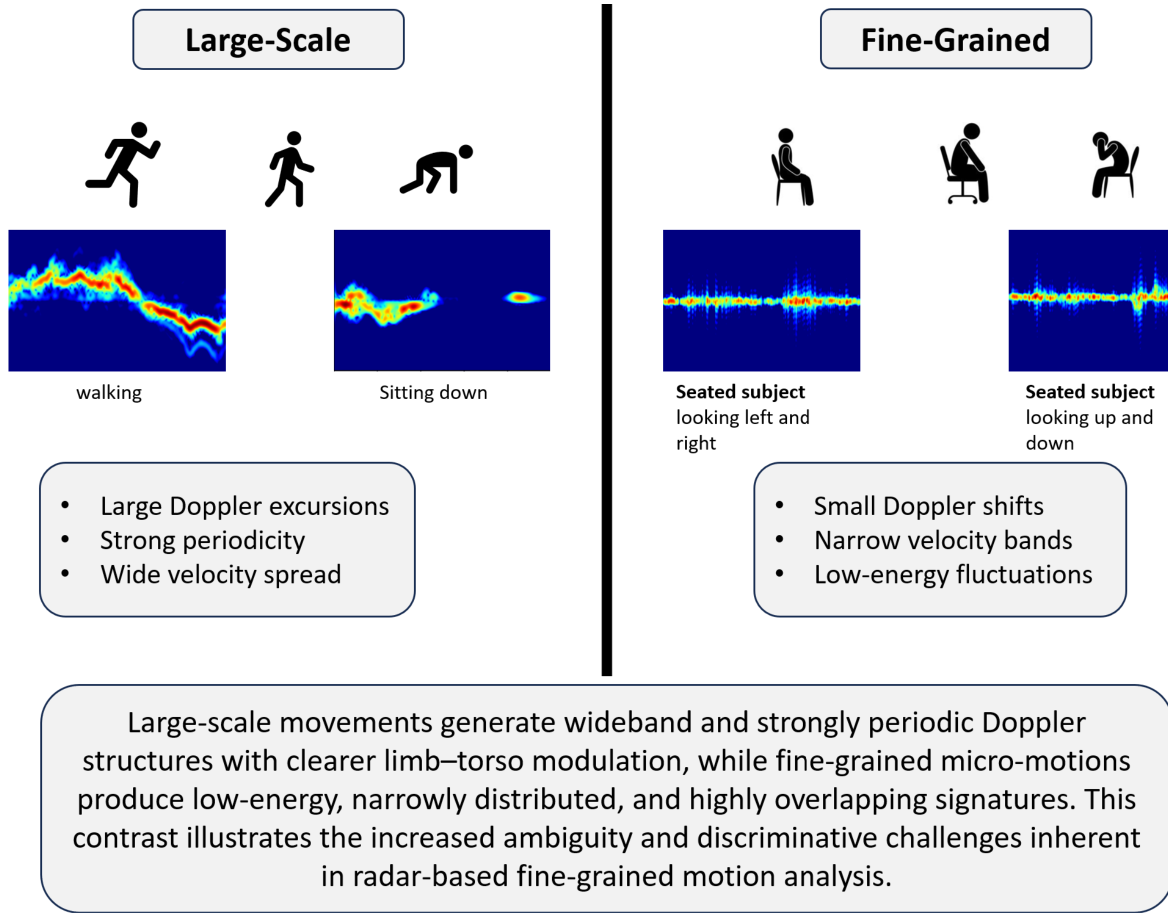


Figure 2.6: Comparison of radar-based large-scale activity monitoring and fine-grained motion monitoring, highlighting typical micro-Doppler characteristics in each regime.

environments remains less established than for larger body movements or thoracic vital-sign monitoring. Nevertheless, fine-grained radar sensing significantly expands the applicability of FMCW systems to domains such as human–computer interaction [123, 152], head movement recognition [153, 154], sign language interpretation [86], eye-gaze tracking [142], and driver behavioural monitoring [155]. These capabilities complement large-scale activity recognition and highlight the unique role of FMCW radar in multi-scale human motion monitoring.

Figure 2.6 illustrates a comparison between radar-based large-scale and fine-grained activity monitoring, including representative TF examples. The large-scale examples are obtained using a 5.8 GHz radar dataset, while the fine-grained examples are derived from data collected with a 77 GHz radar.

2.3 Physiological-Scale Monitoring Using Radar

Radar sensitivity to sub-millimetre displacements makes it a natural candidate not only for activity monitoring discussed earlier but also for contactless vital signs monitoring. Respiration and cardiac activity manifest as periodic micro-motions of the thorax: the chest wall typically moves several millimetres with each breath, while heart-induced vibrations are on the order of tens of micrometres. These displacements modulate the phase or instantaneous frequency of the reflected radar signal and can be exploited for non-contact estimation of RR and HR. This principle has been demonstrated across several radar families, including narrowband mmWave FMCW sensors at 60–77 GHz [125, 156–158], UWB radar in the few-GHz band [159–161], and more recent multi-radar fusion systems that combine FMCW and UWB for estimation in dynamic scenes [162].

Compared with contact-based ECG or chest straps, radar does not require skin contact and can operate through loose clothing, bedding, or even building materials in through-wall life-sign scenarios [159]. Early vital sign studies with CW and UWB radars focused on single, cooperative subjects in static indoor settings [160, 161]. More recent work has extended radar vital sign monitoring to longer ranges using 77 GHz MIMO arrays [157], to specialised clinical environments such as operating rooms [158], and to in-vehicle driver monitoring and cardiac-cycle reconstruction in real driving [33, 98, 114, 163]. The remainder of this section reviews the underlying principles, challenges associated with motion artefacts, and advanced processing strategies that have been proposed in the literature, before summarising representative systems in Table 2.3.

2.3.1 Principles of Radar-Based Vital Signs Estimation

The detection of vital signs using FMCW radar relies on the fact that small respiratory and cardiac motions cause measurable variations in the phase of the reflected signal. As shown in Section 2.2.1, the received FMCW signal is a delayed version of the transmitted chirp,

$$s_{\mathrm{R}}(t) = \alpha\, s_{\mathrm{T}}(t - \tau(t)), \tag{2.40}$$

with two-way propagation delay

$$\tau(t) = \frac{2R(t)}{c}, \tag{2.41}$$

where $R(t)$ is the instantaneous range between the radar and the reflecting surface, as illustrated in Figure 2.7. When the radar output is mixed and low-pass filtered, the remaining baseband signal retains a phase term proportional to this round-trip distance. Using $\lambda = c/f_{\mathrm{c}}$, the dominant phase contribution becomes

$$\phi(t) \approx \frac{4\pi}{\lambda} R(t), \tag{2.42}$$

which follows directly from substituting the delay $\tau(t)$ into the carrier term of the received signal.

Small displacements, therefore, translate linearly into phase variations,

$$\Delta\phi \approx \frac{4\pi}{\lambda} \Delta R, \tag{2.43}$$

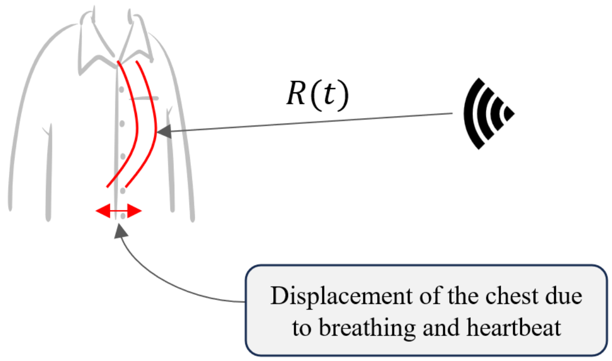


Figure 2.7: Simple concept of radar-based vital signs estimation.

which forms the basis of radar-based respiration and heartbeat extraction. At 77 GHz (with $\lambda \approx 4\,\text{mm}$), a chest displacement of 1 mm produces a full 2π phase rotation. Cardiac micro-motions on the order of 50–100 μm still yield measurable phase excursions of a few degrees [156].

Breathing typically oscillates at about 0.2–0.5 Hz, corresponding to approximately 12–30 breaths per minute (BrPM), while cardiac activity oscillates around 1–2 Hz, equivalent to 60–120 beats per minute (BPM). Because these components occupy distinct frequency bands, they can be separated from the phase time series once the chest reflection is isolated. The standard radar-based vital signs processing chain is:

1. Perform a range FFT to identify the target range.

2. Select the range bin corresponding to the chest reflection [125, 164, 165].

3. Extract the complex baseband signal for that bin and compute its phase $\phi(t)$.

4. Apply phase unwrapping to remove 2π discontinuities and obtain a continuous displacement signal [156].

The resulting unwrapped phase contains superimposed respiratory and cardiac components [156, 157].

From the unwrapped phase signal, the amplitudes and spectral peaks of the low- and high-frequency oscillations yield:

- RR,

- HR,

- Random body motion (for example, slow drift associated with posture changes).

This principle has been validated at short range with single-channel 77 GHz FMCW radars using high-resolution spectral estimators, such as differential recursive least squares multiple classification (DR-MUSIC), to separate heartbeat peaks from respiration harmonics [156], with large MIMO arrays that improve signal-to-noise ratio (SNR) at longer distances [157].

The inherent range resolution of FMCW radar significantly improves robustness compared with single-channel CW Doppler systems, which are prone to strong static clutter. Isolation of the chest range bin suppresses background reflections and improves vital sign detectability [125, 165]. Although UWB radars can also measure small motions and operate through walls, they often require more complex pulsed front-ends and specialised antennas [159]. FMCW radar, therefore, offers a practical balance between hardware simplicity, range resolution, and motion sensitivity, and is now widely used for bedside, surgical, and in-vehicle vital sign monitoring [98, 158, 163].

2.3.2 Challenges in Motion Artefact Removal

Although the vital-sign-induced phase variation in (2.42) is measurable, it is extremely small compared with the phase changes caused by ordinary body motion. Any voluntary or involuntary displacement that alters the instantaneous range $R(t)$ in the propagation delay $\tau(t) = \frac{2R(t)}{c}$ produces a phase shift through the same relation $\phi(t) = \frac{4\pi}{\lambda}R(t)$. Even millimetre-scale posture adjustments can therefore generate phase excursions several orders of magnitude larger than heartbeat signals [125, 160]. Talking, limb gestures, or shifting in a seat can dominate the micro-motion band and obscure physiological oscillations.

Indoor environments introduce additional challenges. Static clutter, multipath reflections, and nearby objects at similar range add strong components to the received signal that interfere with the small chest-induced modulation. Respiration harmonics may also overlap with the heartbeat band when breathing is irregular, which complicates simple band-pass filtering [157, 159]. These effects are moderate in controlled laboratory experiments where the subject is asked to remain still [125, 156, 159], but they become dominant in realistic environments such as operating rooms and vehicle cabins where both the subject and surrounding staff or passengers move [114, 158, 162].

Several strategies have been developed to mitigate these artefacts:

Static Clutter Removal

Static clutter arises from reflectors whose range $R(t)$ is constant, so their phase contribution remains constant in (2.42). High-pass filtering or Moving Target Indication (MTI) techniques suppress these static terms. Coherent background subtraction and MTI are utilised for life-sign detection and through-wall sensing [159]. In FMCW implementations, mean-phase cancellation or subtraction of the average complex return over time can suppress static reflections while preserving the oscillatory chest motion [125, 164]. Multi-channel fusion with adaptive combining further improves robustness to static multipath [157].

Motion-Tolerant Processing

When the subject undergoes slow or moderate motion, adaptive compensation is required to separate the large-scale change in $R(t)$ from the micro-scale variations of interest. Examples include complex demodulation and random-motion-cancellation terms to reduce the impact of body sway on HR estimation [160], and autocorrelation-based processing that is intrinsically less sensitive to rigid-body motion than pure phase demodulation [161]. For FMCW radar, the model-based schemes framework explicitly decomposes the measured displacement into a slow movement component and a high-frequency cardiac term, and uses Kalman filtering to track and suppress motion-induced variance during moderate arm and torso movements [166]. In vehicular radar-based vital sign estimation, machine learning classifiers have been trained to identify and discard, or correct, motion-corrupted segments before HR estimation, thereby significantly improving robustness in driving scenarios [114].

Harmonic Separation

Because respiration and heartbeat occupy distinct fundamental frequency bands, frequency-domain filtering is often applied as a first step. However, breathing produces harmonics that may overlap with the cardiac band, especially under irregular or heavy breathing. High-resolution spectral estimators such as DR-MUSIC can resolve closely spaced peaks and reduce interference between respiration harmonics and the heartbeat tone [156]. TF analysis via continuous wavelet transforms (CWT) or Hilbert–Huang spectra provides further discrimination between components that overlap in frequency but differ in temporal evolution, and has been used in FMCW radars [157].

Multi-Sensor and Multi-Channel Approaches

Multi-sensor configurations provide additional degrees of freedom for mitigating motion artefacts. Studies have shown that combining multiple spatial channels can enhance the SNR and reduce the impact of localised clutter [159]. Similarly, MIMO FMCW radar arrays exploit virtual receive channels and adaptive beamforming or channel combining to improve SNR at longer ranges and for non-frontal subject orientations [157]. In dynamic indoor and vehicular environments, multi-radar fusion and cross-modality fusion with cameras have also been explored to isolate the vital signs of a specific subject from interfering motion and multipath reflections [162, 167]. Such designs are particularly relevant for driver monitoring and multi-occupant scenarios, where requiring the subject to remain stationary is impractical. However, these approaches introduce additional system complexity, either through the use of multiple radar sensors or by incorporating privacy-invasive modalities such as cameras.

Despite these advances, large-scale body motion remains the dominant challenge for radar-based vital sign monitoring. Heartbeat detection is most reliable when the subject is located at short range and remains relatively stationary, consistent with the small-amplitude phase model described in (2.42)–(2.43). Systems designed for realistic daily-life conditions must therefore either tolerate reduced accuracy or incorporate advanced signal processing and learning-based strategies to improve robustness under motion, as discussed in the following section.

2.3.3 Advanced Signal Processing Techniques

Advanced processing methods have been introduced to improve the separation of respiration and heartbeat, especially when motion artefacts distort the phase signal or when measurement conditions deviate from the ideal small-motion model. Because the measured phase $\phi(t)$ is a nonlinear and nonstationary function of $R(t)$, classical linear filtering has limited effectiveness in challenging environments.

Mode Decomposition

Empirical Mode Decomposition (EMD) and VMD decompose the phase-derived displacement signal into intrinsic mode functions (IMFs). In the context of (2.42), low-frequency IMFs correspond to slow variations in $R(t)$ that are dominated by respiration and postural drift, while higher-frequency IMFs represent the small superimposed oscillations generated by cardiac activity. Radar-based vital sign systems have used VMD to separate RR and HR modes and to enhance the HR SNR before spectral estimation [159, 168]. In multi-radar fusion frameworks, hierarchical or weighted VMD has been combined with region-of-interest selection to recover heartbeat for moving people in cluttered rooms [162]. Similar hierarchical decomposition ideas have been applied in mmECG, where the radar phase is modelled as a mixture of motion sources and cardiac micro-motions, and a hierarchical VMD-based separation is used to reconstruct ECG-like cardiac cycles from 77 GHz FMCW radar in real driving conditions [98].

Machine Learning and Deep Learning Approaches

Some approaches use learning to evaluate the reliability of spectral estimates and identify motion-corrupted segments; for example, vehicular radar systems train classifiers on time-, frequency-, and TF-domain features to distinguish reliable from unreliable HR segments, suppressing corrupted intervals to improve robustness in realistic driving scenarios [114]. Similar strategies appear in multi-person monitoring, where auxiliary activity cues or Doppler trajectory analysis help reject segments dominated by gross body motion [162]. Other methods learn direct mappings from radar signals to physiological variables. Contactless seismocardiography (SCG) systems using dual 60 GHz mmWave radars employ CNNs to map phase-derived micro-vibration waveforms to SCG signals, and HR [169]. Multi-radar fusion frameworks such as A-FuseNet combine CNN and LSTM modules with adversarial learning to fuse FMCW and dual-UWB RoIs and separate vital signs from body motion and multipath [162], while vision–radar fusion integrates CNN-based person detection with radar beamforming and multi-channel decomposition to maintain vital sign accuracy in multi-person dynamic environments [167].

Table 2.3 summarises representative radar-based vital sign systems and highlights how radar type, carrier frequency, signal domain, geometry, subject motion, and movement compensation strategies vary across the literature.

Table 2.3: Representative radar-based systems for contactless vital signs estimation, summarising radar type, operating frequency, processing domain, motion conditions, and compensation strategies.

Radar Type	**Carrier Freq.**	**Radar Domain / Processing**	**Radar Position**	**Subject Positions**	**Moving?**	**Movement Compensation**	**Reference**
FMCW	77 GHz	Range FFT for chest range gating, phase time series, median filter, DR-MUSIC to localise RR/HR peaks	Short-range frontal placement in front of the chest (indoor lab)	Single subject facing the radar in a relaxed posture	Static	Static clutter removal, baseline and respiration-harmonic suppression, high resolution DR-MUSIC spectrum	[156]
FMCW	77 GHz	Range FFT and chest-bin selection on many virtual channels, CWT-based TF analysis, Adaptive multi-channel combining (AD-MRC) for vital signs	Side or frontal placement at larger stand-off distances	Single subject lying, sitting, or standing with different orientations	Static	AD-MRC to improve SNR and mitigate multipath, CWT to emphasise periodic components	[157]
FMCW	60 GHz	Range FFT, MIMO beamforming, phase extraction, time-domain separation and spectral estimation for RR and HR in theatre	Mounted on the backside of the surgical bed, illuminating the patient through the bed	Patient lying prone on surgical bed; surgeon may move around the bed	Patient static	Range FFT and beamforming to focus on the patient and suppress the surgeon reflections; modelling of the IF signal for better RR/HR separation	[158]
FMCW	24–77 GHz	Range FFT, chest-bin phase demodulation, band-pass filtering, refined spectral peak selection and multi-stage separation of respiration and heartbeat	Front-facing radar at 0.6–2 m in office or lab environments	Subjects sitting or standing, cooperative and instructed to minimise motion	Mostly static	Optimised separation of respiration and heartbeat, impulse-noise suppression; no explicit compensation of large random motion	[125, 170]
FMCW / UWB	77 GHz	Range FFT and phase demodulation for in-cabin FMCW; time, frequency and TF features with ML reliability classifiers for UWB	FMCW integrated in dashboard or cabin; UWB modules near seat or dashboard	Occupants seated in the driver or passenger seat during driving or driving-like tasks	Quasi-static	ML-based detection and rejection of motion-corrupted segments (UWB), range-bin isolation and filtering (FMCW) for better HR estimates	[114, 163]
FMCW	77 GHz	Range FFT and chest-bin phase extraction, movement mixture modelling, hierarchical VMD, template-based reconstruction of cardiac cycle (mmECG)	In-cabin radar near the steering wheel or dashboard, facing the driver's chest	Drivers seated in a vehicle; urban and highway driving	Yes	Movement mixture model plus hierarchical VMD to isolate heart-induced phase component; template-based optimisation for ECG-like waveform	[98]

Radar Type	Carrier Freq.	Radar Domain / Processing	Radar Position	Subject Positions	Moving?	Movement Compensation	Reference
FMCW + UWB	77 GHz + 3.1–5.3 GHz	FMCW RD maps and phase regions of interest; UWB distance–time regions; wavelet packet features; A-FuseNet (CNN–LSTM–GAN) for joint activity and vital sign estimation	Linear array of one FMCW and two UWB nodes at 1.45 m height in the tent, room, and lobby	Two people simultaneously sitting, standing, or lying with still, limb-motion, upper-body-waggling and turning modes	Yes	Region-of-interest selection and multi-radar fusion; GAN-guided CNN–LSTM network that disentangles heartbeat from complex body motion and multipath	[162]
FMCW / UWB	60 GHz / 3.1–5.3 GHz	Phase-based micro-vibration signals into CNN for contactless SCG, or two-stream (time + frequency) CNN with attention (DeepVS) that regresses RR and HR	Radars placed near chest (bedside) or on furniture facing torso at 0.5–2 m in home-like environments	Subjects lying supine or sitting/standing; incidental body movement allowed	Mostly static	No explicit analytical cancellation; deep models learn mapping from noisy radar signals to SCG or RR/HR and improve robustness to noise and moderate motion	[169, 171]

2.3.4 Limitations and Open Challenges

Despite the encouraging results reported in Table 2.3, reliable radar-based vital signs estimation remains difficult outside controlled laboratory conditions. Most reported systems assume cooperative subjects and limited body motion, and performance often degrades significantly when involuntary or random movement is present. Motion interference introduces strong phase and amplitude variations that overlap with the frequency bands of respiration and heartbeat, masking the comparatively weak physiological signatures and reducing estimation accuracy [167, 172]. This limitation has also been observed in recent mmWave vital signs datasets, where accuracy drops markedly when subjects change posture, perform small body movements, or operate in realistic environments [173]. Furthermore, there is no single compensation strategy that consistently handles all sources of motion artefacts. Classical approaches rely on filtering, range gating, or adaptive decomposition, whereas recent work increasingly employs machine learning to detect or suppress corrupted segments. However, these methods typically require additional assumptions, training data, or computational cost, which may limit real-time deployment in embedded radar platforms.

These limitations indicate that motion-robust physiological sensing remains an open research problem. Developing algorithms that can reliably separate weak vital-sign components from strong body-motion interference while remaining computationally and data-efficient is therefore a key requirement for practical radar-based human monitoring systems. This challenge (**C1.3**) directly motivates the motion-compensation and signal decomposition framework proposed in Chapter 4, which uses VMD as a post-filtering stage after conventional signal processing.

2.4 AI and Deep Learning for Radar-Based Activity Monitoring

Machine learning, and in particular deep learning, has become central to transforming FMCW radar signals into semantic information for both large-scale (**C1.1**) and fine-grained (**C1.2**) human activities monitoring. The radar products derived from the signal-processing pipeline in Section 2.2.1, such as TR maps, STFT-based TF maps, and range–angle (RA) maps, are high-dimensional and structured. These representations contain rich spatial, temporal, and kinematic patterns that are well-suited to learning-based inference, ranging from large-scale activities such as walking and falls to fine-grained behaviours such as head movements, hand gestures, and facial micro-motions.

This section first reviews deep learning architectures that operate directly on radar data, covering both large-scale and fine-grained recognition. It then discusses the fundamental challenge of dataset scarcity in radar-based sensing and surveys transfer learning, augmentation, and generative approaches proposed in the literature to improve generalisation under limited data.

2.4.1 Deep Learning Architectures for Radar Data

Early radar-based human monitoring systems relied on handcrafted features computed from low-dimensional projections such as TR, RD, or TF representation maps. Examples include TF features for fall detection [25], feature-level fusion of TR and TF signatures for six-class human activities with SVM and related classifiers [118], or handcrafted micro-Doppler descriptors combined with Random Forests for ADL recognition in cluttered living rooms [174]. These approaches demonstrated that simple classifiers can extract useful structure from radar, but they depend heavily on manual feature engineering and often generalise poorly beyond the training environment.

Modern systems instead learn features directly from 2D or 3D radar tensors. A common strategy is to treat radar representations as images. Radar representation maps are formed by applying an FFT on fast time to generate TR and a slow-time FFT to generate RD, while TF maps are obtained via STFT or related TF transforms, as discussed earlier in Subsection 2.2.2. Large-scale activities such as walking, sitting, bending, and different types of falls produce characteristic patterns in these images, for example, arc-like Doppler streaks for gait, broadband bursts for impacts, and quasi-static blobs for standing [103, 106]. Deep 2D CNNs have been successfully applied to such images for large-scale monitoring and fall detection [100,144,175]. In these works, radar frames or short frame sequences are mapped to activity labels using architectures ranging from shallow custom CNNs to advanced complex backbones.

Beyond large-scale monitoring, deep CNNs have also been deployed for fine-grained and task-specific radar sensing. TF maps from mmWave FMCW radars have been used for driver head-movement recognition and drowsiness-related motion analysis, where small changes in head pose and neck flexion can be distinguished [154, 176, 177]. Similar image-based CNN pipelines appear in gesture recognition, where TF maps images support recognition of up to 10–14 hand gestures in indoor scenes [121,123,144]. In more advanced settings, RA and RAE maps are converted into 2D projections or compact tensors and used for object classification [146], skeletal pose estimation [178], and 3D facial reconstruction from mmWave radar [150].

Fine-grained monitoring and cooperative head–hand action recognition adds another layer of complexity, because the network must differentiate between closely related motion patterns. Lightweight spatiotemporal CNNs with temporal dynamic convolution, attention, and spectral augmentation have been shown to improve robustness to intra-class variability and noise while remaining deployable on embedded platforms [179, 180]. For real-time gesture control on low-power edge devices, shallow CNNs operating on compact feature cubes that compress range, Doppler, and angle information can achieve high accuracy with very low latency [123].

Another trend is multi-view and multi-representation fusion. Some works process several radar domains in parallel, for example, TR, TF, and RD maps in a multidomain fusion network [122,139], or TR, TF, and RA maps in a multi-branch CNN–LSTM for gesture recognition in complex scenes [121]. Others exploit spatial diversity across multiple radars or antenna sectors, fusing features from separate CNN branches before temporal modelling [137, 181]. This is particularly beneficial for indoor human monitoring and fall detection, where occlusions and aspect-angle variations degrade single-view performance. At the volumetric end of the spectrum, 3D MIMO or 4-D mmWave radars provide point-cloud sequences that can be processed by hybrid autoencoder–RNN models for anomaly detection (for example, fall versus normal ADL) [116], or by dual-branch CNNs for joint 3D pose estimation from multiple radars [178]. These designs move beyond direct reuse of vision-based architectures by tailoring the network parameters, architecture, and fusion mechanisms to the radar application.

In summary, deep learning for radar sensing has evolved from simple CNN classifiers applied to single-domain spectrograms to rich spatiotemporal, multi-branch, and radar-aware architectures. These models span large-scale, healthcare monitoring, in-vehicle driver monitoring, gesture and sign-language recognition, and emerging tasks such as radar-only pose and face estimation. However, their performance is often constrained by the limited size and diversity of radar datasets, which motivates the data-efficient strategies reviewed next.

2.4.2 Data Scarcity in Radar-Based Datasets

Compared with image or speech domains, radar-based datasets are typically small, heterogeneous, and expensive to acquire. This is one of the challenges stated in Section 1.3, specifically **C2**. Collecting labelled radar data requires specialised hardware, controlled experimental protocols, and often multiple sessions across rooms, vehicle interiors, or clinical spaces. Annotation is labour-intensive, particularly for continuous activity streams or multi-occupant scenes. In addition, radar data are strongly device- and scenario-dependent: carrier frequency, bandwidth, antenna geometry, mounting height, and room layout all influence the distributions of TR and TF maps. As a result, models trained on one dataset tend to overfit to that specific environment and degrade under cross-room, cross-frequency, or cross-device transfer [103, 106, 182].

These challenges are especially acute for fine-grained tasks such as monitoring subtle head and seated subject motions. In such cases, the inter-class differences are small and often confounded by subject-specific style, clothing, and clutter. Many published datasets involve only a handful of subjects, limited activity classes, and highly controlled conditions, which is insufficient for training deep models from scratch. This subsection reviews three main strategies that have been proposed in the literature to mitigate data scarcity: generative models for synthetic data, transfer learning, and data augmentation.

Generative Models and GAN-based Synthesis

Generative models, particularly Generative Adversarial Networks (GANs) and related adversarial frameworks, have become a powerful tool for addressing data scarcity and domain shift in radar sensing [183]. The core idea is to learn a generative mapping from a latent space (or from one radar domain) to TF maps or other radar representations that are indistinguishable from real data according to a discriminator network [184]. Synthetic samples can then augment the training set to solve data scarcity or support cross-domain adaptation.

One line of work focuses on unconditional or class-conditioned micro-Doppler synthesis. Wasserstein Refined GAN with Gradient Penalty (WRGAN-GP) has been used to generate additional spectrograms for both in-house and public FMCW HAM datasets, with the synthetic data improving classification accuracy compared with classical augmentation alone [136]. Physics-aware GANs such as PhGAN and MBGAN go further by incorporating kinematic envelopes, stride patterns, and subject-level variability into the generator loss, to enforce physical plausibility of walking and gait signatures [183]. Training classifiers on a mixture of real and physics-aware synthetic data yields significant gains in 14-class ambulatory activity recognition from mmWave micro-Doppler, and better preservation of gait asymmetries than with naively generated spectrograms [183].

A second line of work uses conditional GANs for cross-frequency and cross-modality adaptation. Auxiliary Classifier GAN (ACGAN) and Cycle-Consistent GAN (CycleGAN) variants have been trained to transform micro-Doppler signatures across 77 GHz, 24 GHz, and UWB sensors, thereby reducing the performance drop when training on one frequency and testing on another [182, 184]. These models learn to match not only marginal distributions but also the class-conditional structure of spectrograms, enabling pretraining on abundant data from one radar and fine-tuning on much smaller datasets from another. Similar ideas have been applied to RF-based sign-language recognition, where multibranch GANs generate synthetic micro-Doppler for different signers and fluency levels [185].

Despite their promise, GAN-based and other generative approaches introduce additional challenges that must be carefully considered. Synthetic radar data may not fully capture complex multipath propagation, dynamic clutter, or multi-person interactions encountered in real environments, and subtle biases introduced during generation can propagate into downstream learning models. Consequently, rigorous validation on real-world data remains essential. For these reasons, although generative models constitute an important and rapidly evolving research direction for alleviating radar data scarcity, this work prioritises transfer learning and data augmentation as more controlled and practically robust solutions. These methods, which will be explored next, retain a direct connection to measured radar data while still enabling effective learning under limited data conditions.

Transfer Learning

Instead of generating synthetic data, another approach is to leverage data from other domains. The most straightforward transfer-learning strategy is to reuse CNN backbones pretrained on large image datasets such as ImageNet and fine-tune them on radar-based representations [186]. This approach directly leverages the significant advances in computer vision discussed earlier in the context of vision-based human monitoring in Section 2.1. The success of deep CNNs in visual perception tasks has resulted in robust hierarchical feature extractors that can be repurposed beyond natural imagery.

Despite the domain mismatch between natural images and radar representation maps, the lower convolutional layers of pretrained CNNs tend to learn generic features such as edges, blobs, and texture primitives that transfer effectively to radar representations. As a result, fine-tuning pretrained architectures often leads to improved classification accuracy and faster convergence compared to training models from scratch. Several studies have demonstrated these benefits by adapting architectures such as VGG, ResNet, and GoogLeNet to radar maps for activities of daily living and fall detection [182, 184, 187]. Transfer learning has also proven effective across radar operating frequencies. In dual-frequency FMCW configurations operating at 5.8 and 25 GHz, bottleneck features extracted from a CNN trained at one carrier frequency can be successfully adapted to the other, consistently outperforming models trained from scratch on the target frequency [187]. This observation highlights the frequency-agnostic nature of learned high-level representations and further supports the suitability of transfer learning for data-scarce radar-based human monitoring.

Beyond generic vision pretraining, a second line of work focuses on transferring between radar datasets or RF-based modalities. Supervised domain adaptation has been proposed to align latent feature spaces across four heterogeneous radar datasets, enabling few-shot recognition in a new domain with only a handful of labelled samples [188]. Multi-frequency fusion networks for RF-based sign-language recognition learn shared representations across 77 GHz, 24 GHz, and UWB sensors, and then fine-tune sensor-specific branches for each modality [86]. Similar multi-branch architectures have been used for cross-room and cross-layout adaptation in indoor

ADL datasets, where environment-specific layers capture local clutter statistics while shared layers encode motion-invariant micro-Doppler patterns [103].

Self-supervised and autoencoder-based pretraining provide a third strategy. Convolutional autoencoders (CAEs) and variational autoencoders (VAEs) are trained to reconstruct radar maps, after which their encoders initialise downstream classifiers [116, 182, 189]. This approach encourages the network to model generic structure in radar images without labels, which is particularly useful when only a subset of the acquired data can be annotated. Hybrid architectures such as the Hybrid Variational Recurrent Autoencoder (HVRAE) for 4-D mmWave fall detection use VAEs at the frame level and RNN autoencoders across time to capture both spatial and temporal regularities [116]. When fine-tuned for anomaly detection, these models can recognise rare events (falls) with high sensitivity despite limited labelled examples.

Finally, transfer learning can also be applied across tasks rather than only across datasets. For example, networks trained for large-scale activities or gait recognition can be adapted to more specialised tasks such as sign fluency assessment [185], or head-pose regression [190], by replacing only the final layers and fine-tuning on small target datasets. Therefore, transfer learning, leveraging models trained on larger datasets from other domains (natural images) for radar-based human monitoring across large-scale (**C1.1**) and fine-grained (**C1.2**) tasks, is a promising direction that can be adopted without the complications of GAN-based approaches.

Data Augmentation

Conventional data augmentation remains a simple and effective tool to increase the apparent size and diversity of radar datasets. For image-like TR or TF maps, standard transformations include random cropping, flipping of Doppler sign (to simulate motion in the opposite direction), additive Gaussian noise, amplitude scaling, random erasure, and small affine transformations. These operations introduce variability in target range, orientation, and SNR, and have been shown to improve accuracy and robustness for radar-based HAM and fall detection [103, 135, 175]. In multi-domain frameworks, different augmentations can be applied per domain to improve the complementarity among the TR, TF, and RD branches [122, 191].

More recent work adapts techniques from speech processing, such as SpecAugment [40], to radar spectrograms [180]. Time- and frequency-masking operations randomly zero out contiguous bands in TF map images, forcing the network to rely on global structure rather than narrow-band artefacts. Temporal dynamic convolution combined with spectral masking has improved robustness to corruption in radar-based swimming activity recognition and in head–hand action recognition for in-vehicle monitoring [179, 180]. These results suggest that spectral augmentation is particularly useful for fine-grained tasks, where small local corruptions or sensor noise can otherwise lead to misclassification. Beyond purely image-level transformations, physics-aware augmentation synthesises new spectrograms or point clouds by perturbing kinematic and geometric parameters. Examples include simulating different approach angles,

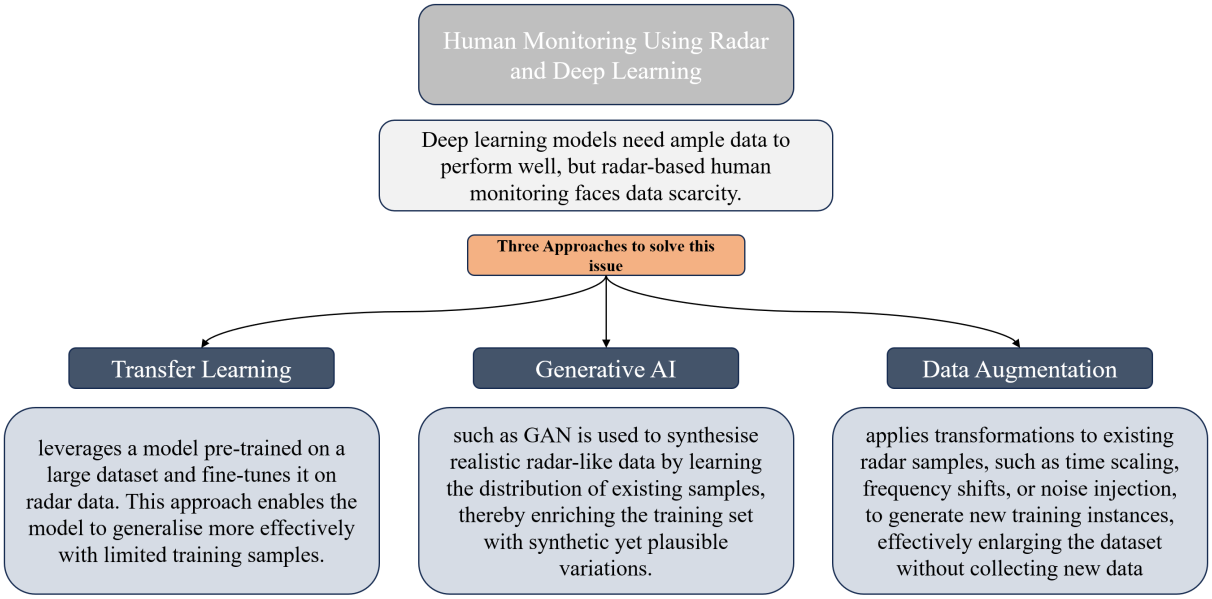


Figure 2.8: Strategies for addressing data scarcity in deep learning for radar-based human monitoring.

ranges, and walking speeds in micro-Doppler signatures based on simple motion models, or mixing trajectories from several recordings to emulate multi-person scenes [182,183]. In multi-radar setups, replaying trajectories through different virtual sensor configurations can approximate the effect of changing radar placement without re-collecting data [137, 181].

Figure 2.8 summarises the three main strategies that have been explored in the literature to address radar data scarcity. Among these, transfer learning and data augmentation are adopted and systematically investigated in this thesis, as they offer a favourable balance between effectiveness, implementation complexity, and robustness to radar maps and advanced learning architectures surveyed in Section 2.4.1. These approaches enable data-efficient learning without introducing synthetic domain bias and are particularly well-suited to scenarios where labelled radar data are limited.

Table 2.4: Representative radar-based large-scale and fine-grained studies, and their use of transfer learning, generative models, and data augmentation to address data scarcity.

Preprocessing	Learning Model	Application		Data Scarcity			Reference
		Large-Scale	Fine-Grained	TL	GAN	Aug.	
STFT	AlexNet, VGG-16, VGG-19	✓		✓			[118]
STFT	CNN		✓				[144]
STFT	AlexNet, GoogLeNet, custom CNN	✓		✓			[187]

Preprocessing	Learning Model	Application		Data Scarcity			Reference
		Large-Scale	Fine-Grained	TL	GAN	Aug.	
STFT	CNN feature extractor with supervised domain adaptation (few-shot)	✓		✓			[188]
STFT	CNN, stacked AE, classical Machine Learning	✓		✓			[103]
STFT	VGG-16, CAE, multi-modal Deep NN with ACGAN pre-training	✓		✓	✓	✓	[182]
STFT	WRGAN-GP + DCNN classifier	✓			✓	✓	[136]
STFT	Physics-aware GAN + CNN	✓			✓	✓	[183]
STFT	CNN ASL recogniser + multibranch GAN		✓	✓	✓	✓	[185]
TR, STFT	Deep CNN with Temporal Dynamic Convolution (TDC)		✓			✓	[179]

2.4.3 Limitations and Open Challenges

The preceding literature review highlights a rapid maturation of radar-based HAM across a wide spectrum of scenarios. Existing work has convincingly shown that FMCW radar can discriminate human motion patterns at different spatial scales when appropriate signal processing and learning models are used. However, when this body of work is examined through the lens of the challenges defined in Section 1.3, several systematic gaps remain in how the community addresses the three scales of motion described in **C1** and the problem of data scarcity in **C2**.

First, most large-scale radar-based HAM studies focus on a limited set of coarse activities such as walking, sitting, standing, bending, and falling, often in controlled indoor environments with a single radar and a restricted geometry. Many methods rely either on handcrafted TF features or on a single radar representation, such as an STFT TF-based map or a TR map. Only a subset of works explore multiple radar domains, and very few provide a systematic comparison of how different representations and CNN architectures trade off discriminative performance versus computational cost. As a result, the community still lacks a clear understanding of which combinations of radar representation map and model are most suitable for real-time, resource-constrained deployments, a central component of the representational challenge in **C1.1**.

Second, fine-grained human monitoring, especially at the level of head, upper-body and hand motions, is even less mature. The studies summarised in Tables 2.4 and 2.5 demonstrate that mmWave FMCW radar can recognise subtle driver head movements, hand gestures, and small postural changes. Yet these works typically involve small participant pools, a narrow set of motion classes, complexity of using multiple sensors, and highly constrained laboratory or simulator setups. Many methods do not rigorously enforce subject-independent evaluation. Confusion between kinematically similar classes is common, particularly for head and torso movements that share similar velocity envelopes. This directly reflects the difficulty, articulated in **C1.2**, of finding representations and learning strategies that can reliably separate subtle, behaviourally important motions whose radar representation maps are inherently ambiguous.

Third, across both large-scale and fine-grained monitoring, almost all studies are constrained by limited datasets. Radar data collection requires specialised hardware, careful calibration, and often safety considerations in the case of falls or clinically relevant behaviours. Consequently, many datasets comprise only a few subjects and a modest number of repetitions per class. Deep learning methods are therefore exposed to a severe imbalance between model capacity and data volume, which leads to overfitting and poor generalisation across subjects, rooms, and deployment conditions. This is precisely the bottleneck formalised in **C2**, and it is especially acute for fine-grained actions where collecting many labelled examples of rare or transient behaviours is particularly difficult.

Within this landscape, the work presented in Chapters 3 and 4 aims to address these gaps in a structured manner. Chapter 3 targets large-scale human monitoring and investigates how different radar representations and transfer learning models jointly address **C1.1** and **C2** in a public 5.8 GHz FMCW dataset. Chapter 4 focuses on fine-grained head and upper-body motions and examines how a carefully designed micro-Doppler pipeline and spectrogram augmentation can mitigate the ambiguity and data scarcity associated with **C1.2** and **C2**. Together, these chapters provide an evidence-based perspective on how to choose radar signal representations, deep learning architectures, and data augmentation strategies for the three scales of human activity monitoring.

Large-Scale Human Monitoring

For large-scale motions such as walking, sitting, standing, bending to pick up an object, and falling, the core challenge in **C1.1** is to identify radar signal representations and deep learning models that are both discriminative and computationally efficient. The literature shows that many studies use a single radar domain, typically TR maps or STFT-based TF maps, and report performance mainly in terms of classification accuracy. Only a few studies explore richer TF representations, such as SPWVD [143], and even fewer report the computational cost of preprocessing and inference. Moreover, many methods employ relatively heavy CNN architectures without quantifying the trade-off between accuracy and inference time, even though practical human monitoring systems are expected to run on embedded hardware with stringent latency and power constraints. This means that, at the level of large-scale motion (**C1.1**), it remains only partially addressed: radar can represent human motion in many ways, but has not yet converged on principled design rules for selecting representations and models suitable for real-time human monitoring.

Chapter 3 directly tackles this gap by systematically analysing a broad set of model–representation pairs on a well-established 5.8 GHz FMCW radar dataset of six large-scale activities (walking, sitting, standing, bending to pick an object, drinking, and falling). The study considers three radar representations, namely TR maps, STFT-based TF maps, and high-resolution SPWVD-based TF maps, and combines each of them with four transfer learning architectures (VGG-16, VGG-19, ResNet-50, and MobileNetV2). This yields twelve model–representation pairs whose performance is evaluated not only in terms of accuracy, precision, recall, and F1-score, but also in terms of preprocessing time, training time, and inference latency. This comprehensive analysis provides concrete evidence that directly addresses the representational and computational aspects of **C1.1**.

In addition, the use of transfer learning in Chapter 3 offers a response to the data scarcity problem in **C2**. Rather than training deep CNNs from scratch on a relatively small radar dataset, the study leverages pre-trained visual models and fine-tunes them on radar-derived images. This approach allows the models to reuse generic visual features while adapting only the higher-level layers to the radar domain, which reduces the amount of task-specific data required [186]. Although this does not eliminate the need for larger radar datasets, it demonstrates that careful model and domain selection can produce robust large-scale human monitoring systems under realistic data constraints. The resulting design guidelines from Chapter 3 serve as a conceptual foundation for the fine-grained human monitoring system in Chapter 4.

Fine-Grained Human Monitoring

Fine-grained human monitoring aims to detect subtle head and upper-body movements that are crucial for understanding human state, such as driver distraction, drowsiness, or risky behaviours. In the context of **C1.2**, the primary difficulty is that many of these movements are kinematically similar and therefore produce micro-Doppler signatures that are highly overlapping when observed from a single radar viewpoint. The literature on radar-based driver monitoring, head motion sensing, and gesture recognition confirms this challenge. Several studies report high accuracy on a narrow set of activity classes [176], limited subject diversity [154], and utilised complex multiple sensors [192]. Confusions between motions such as nodding versus lowering the head, or turning versus leaning, are frequently noted. Furthermore, almost all works are constrained by small datasets and infrequent subject-independent evaluation, which amplify the concerns raised in **C2** about model generalisation on limited datasets.

The first part of Chapter 4 addresses these issues by designing a fine-grained human monitoring system based on a single 77 GHz mmWave FMCW radar that observes a seated subject performing six kinematically similar head and upper-body activities: looking front, turning the head, nodding, leaning the body, moving the upper body forward and backwards, and looking back over the shoulder. The chapter introduces a signal processing pipeline that converts the raw IF radar data, detailed in Section 2.2, into noise-suppressed TF maps using STFT, followed by Gaussian smoothing. This representation is carefully chosen to emphasise the time-varying velocity content of small head and torso motions, while suppressing static clutter from the seat and surrounding cabin. The dataset comprises 4400 seconds of recordings from 15 participants across two indoor environments, and the evaluation is performed using a subject-independent split that directly probes the generalisation aspect of **C2**.

On the learning side, the fine-grained component of this work employs a lightweight MobileNetV2 backbone with a two-stage fine-tuning strategy. This approach follows the same transfer learning principles used for large-scale human monitoring in Chapter 3, while being adapted to the fine-grained dataset and associated motion characteristics, as detailed in Chapter 4. A key contribution of this chapter is the introduction of a tailored data augmentation pipeline inspired by SpecAugment [40]. The proposed pipeline applies realistic time- and frequency-masking, along with slight rotations, to radar representation maps to increase data diversity while preserving the underlying physical characteristics of the radar signatures.

Table 2.5 summarises the large-scale and fine-grained literature by outlining the contributions and limitations of prior works.

Table 2.5: Key outcomes and limitations of representative radar-based human activity monitoring studies.

Key Outcomes	**Limitations**	**Reference**
Feature-level fusion of TR and FT representations for FMCW radar yields higher recognition accuracy than single-domain features, with SVM on fused features achieving about 98.5% accuracy for six indoor activities.	Relies on hand-crafted features and classical machine learning rather than end-to-end deep learning, and is evaluated only in a controlled single-room setup.	[118]
A shallow CNN operating on RD maps from a single low-cost FMCW radar enables four-class gait analysis (normal and pathological gaits) with a reported accuracy of 91.75%, demonstrating the feasibility of low-cost radar-based gait assessment.	The dataset is limited to 10 subjects in a controlled environment, with only frontal walking; lateral and indirect views, and more realistic conditions, are not considered, so generalisation is uncertain.	[60]
Comprehensive analysis of continuous-wave and pulse-Doppler radar TF features for elderly fall detection shows that TF representations can separate fall from non-fall events, and that multi-sensor configurations reduce occlusion effects.	Confusion remains between kinematically similar motions such as sitting or kneeling, weak Doppler signatures for some fall types, and vulnerability to occlusion, ghost targets and pet movement; deployment in cluttered real homes is challenging. The multi-sensor setup increases the system's complexity and cost.	[25]
Patient activity recognition in mock and real hospital rooms shows that CNNs trained on RD maps can achieve very high accuracy (around 95%) within a given environment, whereas CNNs trained on TF maps generalise better across environments.	Generalisation across rooms and layouts remains difficult as range distributions and clutter change; bed activities are more confusable than in-room motions; some elderly participants could not perform floor-level activities, leading to class imbalance; and performance is sensitive to radar placement and SNR.	[106]
Multidomain fusion of TR, TF and RD maps with a hybrid deep model significantly boosts large-scale Human monitoring accuracy compared with single-domain inputs, and domain-specific feature extraction improves robustness across different indoor locations.	The method requires all three radar domains and is computationally intensive; the dataset is relatively small with limited activity diversity, and performance is assessed only in controlled indoor scenarios.	[122]

Key Outcomes	Limitations	Reference
A distributed radar network combined with an RNN for sequence classification significantly improves continuous large-scale human activity and fall-related monitoring compared with single-node and feature-level fusion; multi-static geometry helps to capture diverse motion aspects in a room.	Experiments are restricted to single-person scenarios in one indoor environment, with class imbalance and annotation challenges for continuous sequences; multi-person scenes and transfer to other RF modalities or layouts are not evaluated.	[181]
Cross-frequency training of deep CNNs on TF maps shows that pre-training with large amounts of ACGAN-synthesised 77 GHz data can reduce cross-frequency accuracy loss by about 50% and improve overall large-scale human monitoring accuracy across sensors compared with random initialisation or transfer learning.	The study uses data from only six participants in controlled line-of-sight indoor trajectories, focuses mainly on frequency mismatch, and does not fully examine other domain shifts (rooms, clutter, multi-person scenes). GAN-synthesised signatures still exhibit kinematic mismatches that degrade performance.	[182]
The mmFall system uses 4D mmWave radar point clouds and a hybrid variational recurrent autoencoder to learn the distribution of normal motion and detect falls as anomalies, combining anomaly scores with centroid-height changes to achieve around 98% detection of simulated falls with few false alarms.	Evaluation is limited to a single apartment, a small number of young subjects and a restricted set of fall types and daily activities; clutter conditions are relatively simple and computational complexity and embedded deployment aspects are not fully analysed.	[116]
This work architecture performs mid-level fusion of RD, TF and TR views with auxiliary tracking features, achieving macro-F1 scores around 0.7–0.75 for four primitive actions and up to about 90% accuracy when sit and stand transitions are merged, showing improved robustness to cluttered rooms and unseen participants.	The system still relies on only four base actions, which makes separation, such as sit versus stand, difficult; it has not been evaluated on richer activity classes or multi-person scenarios, and CAE pretraining with multi-view LSTM fusion increases training and inference complexity.	[189]
Adaptive thresholding of TF maps to extract regions of interest and extend masks into the phase and unwrapped-phase domains, followed by multi-domain handcrafted features and SVM classification, achieves about 93.1% accuracy while reducing parameters, memory, and computation by more than an order of magnitude relative to deep CNN baselines for large-scale human monitoring.	The approach is evaluated using a random dataset split, which raises concerns regarding generalisation. In addition, no explicit strategy is reported to address the imbalance in fall events. The reliance on handcrafted features further limits adaptability, as re-tuning may be required when deploying the system in different environments or with alternative sensor configurations.	[58]
In-cabin mmWave FMCW TF maps are shown to support reliable classification of multiple fine-grained driver head motions, with a deep CNN achieving high accuracy in stationary vehicles, indicating the feasibility of radar-based driver head-motion monitoring.	The dataset comprises a limited number of subjects, performance degrades under moving-vehicle conditions, and experiments rely on scripted motions with clear line-of-sight, so generalisation to naturalistic driving and different seating geometries is still open.	[154]
A 3D mmWave imaging radar mounted near the driver's visor, combined with a ResNet-BiLSTM network, enables volumetric driver fatigue detection, achieving about 93% accuracy on simulator data and up to 98% on a smaller real-driving dataset.	The approach uses a specialised 3D imaging radar that is not representative of standard commercial radars; real-driving data volume is limited, and scenarios remain relatively controlled. Generalisation to diverse drivers, vehicles and long-term naturalistic driving has not yet been demonstrated.	[155]
Through-wall head-movement sensing using UWB radar and Wi-Fi CSI demonstrates that deep learning on TF maps can recognise five fine-grained subtle head gestures in both line-of-sight and non-line-of-sight conditions, and that feature-level fusion of radar and Wi-Fi improves accuracy compared with single-modality systems.	Experiments involve a limited number of subjects and head-movement types, which may restrict generalisability to more diverse behaviours. The reliance on multiple sensing modalities further increases system complexity, potentially limiting practical deployment.	[192]
A 60 GHz FMCW radar in front of the subject's face, combined with angle spectrograms and a MobileNet-based classifier, shows that fine-grained head-orientation based eye-gaze tracking can reach over 90% accuracy for multiple gaze directions, demonstrating the feasibility of camera-free radar-based gaze tracking on commercial chips.	The study uses a single controlled indoor setup with a limited number of subjects and viewing configurations, focuses mainly on horizontal head orientation, and does not include a dynamic environment.	[142]

Key Outcomes	Limitations	Reference
A 61 GHz FMCW radar mounted on the steering wheel and analysed via TF maps with a CNN enables one-shot learning of four fine-grained driver head-movement classes, achieving high classification accuracy even with limited samples and outperforming a conventional CNN on an additional dataset.	Experiments are mainly conducted in stationary or tightly controlled conditions with a small subject pool and only four coarse movement classes.	[177]
The mmAssist system uses a single off-the-shelf mmWave radar to monitor fine-grained driver upper body and head from RD maps, training three supervised models that reach around 88% weighted F1 in lab conditions and about 79% in real driving for detecting lack of attention across several distraction scenarios.	The system is evaluated with one vehicle, and performance drops under real-road vibrations compared with the lab. The relationship between performance and computational cost is not reported.	[193]
A real-time driver state detection framework based on a single in-cabin mmWave FMCW radar fuses spatial and temporal features from TR, RD and related representations via a spatiotemporal network with cross-gated fusion, achieving about 94% average F1 for driver state classification and about 97% F1 for high-risk behaviours, while remaining deployable on Jetson-class edge devices.	The behavioural classes employed have been developed and evaluated primarily within narrowly defined, simulated use cases, and thus their capacity to generalise to other vehicle layouts, diverse driver populations, highly cluttered environments, or situations involving multiple occupants remains unproven.	[127]
A VGG16-LSTM network applied to TF maps from a single radar achieves 99.16% accuracy for seven head–hand cooperative driver actions and reduces VGG16 parameters by about 95% while slightly improving accuracy compared with several CNN and CNN-LSTM baselines, indicating the potential of lightweight attention-based networks for fine-grained human monitoring.	Experiments are conducted in a simulated driving environment with a small number of subjects and relatively simple scripted actions; more complex driver behaviours, interference from passengers and naturalistic on-road conditions are not yet studied.	[179]

Toward Multimodal Monitoring

This chapter has shown that FMCW radar can support human monitoring across multiple motion scales, from large-scale activities to fine-grained movement and physiological dynamics. However, reliable human-state inference is rarely achievable with a single modality, because behavioural and physiological cues provide complementary information. Radar can contribute to both, but this thesis focuses on a practical multimodal pairing: physiological metrics are combined with gaze metrics as a behavioural measure to enable more comprehensive and robust monitoring. The next section, therefore, reviews multimodal fusion approaches, with emphasis on integrating gaze-derived attention indicators with heart-related measures.

2.5 Multimodal Human-State Monitoring

The evolutionary trajectory of human-centric computing has shifted fundamentally from rudimentary, binary presence detection to the granular estimation of complex physiological, cognitive, and affective states [194]. In high-stakes, safety-critical domains such as automotive safety [155], the reliability bottlenecks and robustness limitations inherent to unimodal sensing architectures have become increasingly evident, necessitating the adoption of multimodal frameworks [195]. Unimodal systems are constrained by their sensing principles and are generally unable to provide a fully comprehensive assessment of subjects. As a result, multimodal monitoring has emerged not simply as an enhancement, but as a necessity for robust human-state inference. By integrating the complementary strengths of heterogeneous sensing modalities, multimodal systems construct a more holistic and resilient representation of human behaviour and health [172]. This approach typically combines several sensor modalities, including visual, physiological, environmental, and behavioural, to jointly observe and interpret human state, rather than relying on any single channel [196, 197].

Fusing these modalities through intelligent signal processing and AI enables a system to compensate for the shortcomings of each source and to amplify overall perceptual fidelity. For example, combining eye-tracking data with HR variability enables more accurate detection of cognitive load or fatigue than either modality alone [38]. The integration of these disparate modalities, facilitated by advanced signal processing and AI algorithms, enables the system to counterbalance the intrinsic limitations of individual sensors. This synergistic fusion is critical for enhancing overall perceptual fidelity by leveraging the complementary nature of heterogeneous data streams to resolve ambiguity [198]. For instance, empirical studies demonstrate that the concurrent analysis of oculometric data (eye-tracking) and physiological signals (HR) yields significantly higher accuracy in detecting cognitive load and driver fatigue compared to unimodal baselines [199].

This multimodal approach also supports redundancy and adaptability, which are essential for real-world deployments. In a smart home, cameras may be deployed in public areas for detailed behavioural analysis, while privacy-preserving radar sensors monitor movement and respiration in bedrooms [200]. In driver monitoring, visual cues (eyelid closure, head pose) can be paired with physiological indicators (HR, RR) and vehicle telemetry (steering patterns, lane deviation) to infer fatigue, distraction, or emotional stress more reliably than any single modality alone [201].

Ultimately, the convergence of multimodal sensing technologies provides a pathway toward more robust, context-aware, and human-aligned systems that directly address the challenge of holistic, reliable human-state understanding (**C3**). It supports intelligent environments that can continuously assess human state not only in terms of activity but across a broader spectrum of health and cognition.

2.5.1 Fusion of Gaze and Vital Signs for Multimodal Human-State Monitoring

Combining eye gaze information with vital signs is a powerful multimodal strategy for assessing a person's cognitive and physiological state. Gaze metrics, such as eye fixation, blink rate, fixation duration, and saccade patterns, provide direct evidence of visual attention and mental engagement. In contrast, vital signs such as HR, heart rate variability, and RR indicate underlying physiological arousal, stress, or fatigue levels [202]. Fusing these two modalities can therefore provide a more complete picture of the human state than either alone. For instance, an elevated HR or changes in heart rate variability may signal stress or high cognitive workload. In contrast, changes in gaze behaviour, such as more frequent long blinks or erratic eye movements, may indicate waning attention or emerging fatigue [203]. By analysing these signals together, it becomes possible to detect subtle psychophysiological patterns, such as the onset of drowsiness, cognitive overload, or distraction, with greater confidence and at an earlier stage [204]. Gaze and vital signs are thus complementary indicators: one monitors outward attentional focus, the other monitors internal physiological state, and their fusion yields a more robust measure of overall alertness or well-being.

Research has shown the combined effectiveness of gaze and physiological data in various human monitoring applications contexts. In cognitive workload studies, increases in task difficulty have been associated with both broadened gaze dispersion and elevated HR, suggesting that visual scanning patterns and cardiac responses together reflect rising mental effort [205]. In aviation and control-room settings, eye-tracking combined with HR monitoring has been used to assess operators' attention and stress levels, providing a multidimensional view of operator state that can trigger timely alerts or interventions [206]. Fusing gaze with vital signs is particularly useful for detecting fatigue and distraction. Gaze alone might indicate whether a person's eyes are off-target or drooping. Still, when these observations are combined with physiological cues, the assessment of fatigue or overload becomes far more reliable [202].

2.5.2 Applications in Driver Monitoring

Driver monitoring is the primary applied use case examined in Chapter 5 and is used throughout this thesis as a concrete setting in which behavioural and physiological cues must be fused in real time. Ensuring that a driver remains alert, attentive, and in good condition is crucial for road safety, and modern Driver Monitoring Systems (DMS) increasingly leverage multiple data streams to achieve this. In practice, a comprehensive driver monitoring approach can incorporate:

- **Visual cues** (cameras monitoring the driver's face and eyes) to detect gaze direction, eyelid closure (drowsiness via PERCLOS), head pose, and facial expressions of distraction or distress.
- **Physiological signals** (contact or contactless biosensors) to track HR, heart rate variability, and RR as indicators of fatigue, stress, or cognitive load [207]. For instance, wearable ECG/PPG sensors or in-seat sensors measure the driver's heart activity, while emerging in-cabin radar devices can monitor heart and breathing rate without any contact [163].
- **Vehicle dynamics** (vehicle-mounted sensors) to observe driving behaviour such as steering patterns, lane deviations, speed variance, and pedal inputs, which often change when a driver is drowsy or distracted [207].

- **Audio or vocal cues** (microphones monitoring the driver's voice or in-cabin sounds) to detect signs such as yawning, slurred speech, or changes in conversational engagement.

By fusing these modalities, driver monitoring systems create a comprehensive view of the driver's state. For example, a system may consider a driver's eye gaze to determine whether they are looking at the road, HR to assess stress or fatigue, and lane position to detect lane drift. Studies have shown that incorporating physiological data, such as heart activity measured by ECG or PPG, alongside camera-based observations provides valuable insights into the driver's cognitive and emotional state [208]. It is not sufficient to track only outward behaviour; internal signals and contextual data about the vehicle and environment are also needed to understand and predict driver safety risks [209]. For instance, an increase in HR accompanied by erratic steering may indicate rising stress or fatigue even if the driver's facial expression appears normal.

Moreover, multimodal driver monitoring has shown clear benefits in research. Simulator-based studies combining HR sensing with eye-tracking have demonstrated that drivers exhibit significantly higher HRs and more frequent gaze indicators of mental workload during demanding urban driving than on motorways [210]. Based on such observations, fused alertness scores that integrate HR with gaze features (blink duration, fixation length, saccade metrics) have been proposed to distinguish between alert and fatigued states more reliably than single-modality cues [210]. Similarly, datasets such as DROZY [211] and related work that combine facial video with EEG and ECG show that fusing visual and physiological signals can achieve very high accuracy in recognising driver fatigue by capturing both outward signs and underlying physiological changes. In one such system, integrating EEG, ECG, and eye-camera data yielded fatigue detection accuracies above 98%, clearly outperforming unimodal approaches [207].

In addition to improving accuracy, multimodal fusion enhances robustness under varying conditions. A camera-based DMS may fail in poor lighting or when the driver's face is partially obstructed. In contrast, radar or infrared physiological sensors remain unaffected and can continue to monitor HR or RR [127]. Radar, in particular, has emerged as a valuable complementary modality in vehicles. It offers a non-intrusive, privacy-preserving means of sensing occupant presence, gestures, and vital signs through clothing and under challenging lighting conditions [127, 212]. Automakers and researchers are exploring sensor fusion setups in which an in-cabin mmWave radar is integrated with a driver-facing camera [133, 213]. In these systems, the radar tracks HR, RR, and gross body motion, while the camera tracks gaze and facial cues, so that each modality covers the other's blind spots.

Multimodal monitoring is becoming a cornerstone of advanced driver assistance and safety systems. It enables early detection of dangerous states such as drowsiness, distraction, or sudden illness by cross-verifying signals. The fusion of gaze tracking and vital sign monitoring is a concrete example of how human-centric multimodal systems can enhance safety by continuously assessing a driver's state and enabling timely interventions before an incident occurs.

2.5.3 Limitations and Open Challenges

The challenge of **C3** in Section 1.3 emphasises that sensing physical motion (**C1.1** and **C1.2**) or physiological signals (**C1.3**) in isolation provides an incomplete picture of a person's actual cognitive or physical state. The literature reviewed above confirms this limitation. Most driver monitoring systems are unimodal or only weakly multimodal, for example, using either camera-based gaze features or HR alone, and in many cases, they are evaluated in controlled laboratory or simulator environments as illustrated in Table 2.6. Even when multiple modalities are used, they are often fused at a coarse level, with limited attention to how individual metrics such as blink dynamics, fixation patterns, saccade behaviour, and HR interact over time to reflect underlying cognitive load or fatigue.

Table 2.6 highlights this gap in the specific context of fusing HR and gaze metrics. Several studies incorporate HR into real driving, but use only global gaze indicators or none at all. Others use detailed gaze features such as blinks and fixations, but do not jointly model HR, or are restricted to simulated or controlled settings. Among the studies summarised in Table 2.6, we did not find any that jointly model HR with blink, fixation, and saccade features in real-world on-road driving using an interpretable fusion framework. Existing work typically either emphasises HR with limited gaze features or uses richer gaze features without joint modelling of HR under naturalistic driving. As a consequence, existing systems only partially meet the requirement in **C3** for comprehensive human-state assessment under realistic, dynamic conditions.

Moreover, current multimodal frameworks often treat individual metrics as independent predictors and rely on black-box fusion strategies that are difficult to interpret and to adapt across scenarios. There is a lack of frameworks that explicitly quantify how changes in HR and specific gaze metrics co-evolve during different driving conditions and that translate these joint patterns into interpretable scores of driver state. Addressing this gap requires multimodal algorithms that are both statistically grounded and practically deployable, capable of fusing multiple gaze features and HR in real driving environments, and directly aligned with the holistic assessment (**C3**).

2.6 Identified Research Gaps and Thesis Positioning

The literature reviewed in this chapter shows that human monitoring has progressed from simple presence detection to sophisticated activity monitoring and state estimation using wearable, vision-based, and RF-based systems. Within RF sensing, and radar in particular, multiple works have demonstrated that both large-scale and fine-grained activities can be recognised with high accuracy under controlled conditions, and that physiological-scale parameters such as RR and HR can be estimated in constrained settings. Multimodal systems that integrate visual, physiological, and contextual signals have also emerged, particularly in safety-critical domains such as driving. However, when these developments are viewed through the problem formulation in

Table 2.6: Comparison of driver monitoring studies using HR and gaze metrics.

Study	HR	Gaze Metrics			On-road driving
		Blink	Fixation	Saccade	
[214]	✓	✗	✗	✗	✓
[215]	✓	✗	✗	✗	✓
[216]	✓	✗	✗	✗	✗
[217]	✗	✗	✓	✗	✓
[218]	✓	✓	✓	✗	✗
[38]	✓	✓	✓	✗	✗
[210]	✓	✓	✓	✓	✗

Section 1.3, several persistent research gaps remain that motivate the contributions of this thesis.

First, at the level of large-scale human monitoring, the core challenge in **C1.1** is to sense and interpret gross activities such as walking, sitting, bending, and falling in a discriminative and computationally efficient manner. Existing radar-based HAM studies, as reviewed in Section 2.4.3, typically select a single radar representation and a single learning model, and optimise mainly for recognition accuracy. There is no consensus on how to match different two-dimensional representations, such as TR, TF, or higher-resolution SPWVD maps, to lightweight or deeper CNNs for real-time deployment. Very few works report preprocessing and inference budgets, and performance is often evaluated on limited datasets with restricted geometries and single-room scenarios. As a result, systematic design rules for representation and model selection can satisfy the requirements of **C1.1** under realistic resource and generalisation constraints.

Second, fine-grained human monitoring remains significantly underdeveloped compared with large-scale human monitoring. The challenge in **C1.2** is to distinguish subtle, kinematically similar movements, such as different seated upper-body and head poses, which are critical for applications like driver monitoring. The literature reviewed in Section 2.4.3 shows that mmWave radar can detect such motions. Still, most studies rely on small participant pools, narrow activity classes, and highly controlled laboratory or simulator settings. Confusion between similar motion classes is common, primarily when relying on a single TF representation from a single radar viewpoint. Furthermore, these works rarely adopt a strict subject-wise evaluation, which is necessary to assess generalisation across individuals. Consequently, the core difficulty in **C1.2**, namely learning robust patterns from an ambiguous, small dataset of TF signatures of subtle movements, remains only partially addressed.

Third, at the physiological-scale, **C1.3** highlights that radar-based vital sign estimation is heavily affected by motion artefacts arising from both large-scale and fine-grained movements. While there is a growing body of work on contactless RR and HR monitoring, much of it assumes near-static conditions or focuses on algorithmic improvements in limited-motion scenarios. Robust extraction of vital signs in the presence of realistic body movements remains an open problem. There is a clear need for radar processing chains that explicitly incorporate motion artefact suppression while remaining compatible with practical monitoring conditions.

Across all these scales, the data scarcity problem formalised in **C2** persists. Collecting large, well-annotated radar datasets is expensive and time-consuming, particularly for events such as falls or for fine-grained movements that are rare or difficult to script. Many published studies rely on in-house datasets, often with fewer than a few tens of subjects, and only a subset of these datasets are publicly available. This limits the training and benchmarking of modern deep learning models and hinders fair comparison between architectures. It also makes it challenging to study transfer across environments, sensors, and populations, which is essential for robust human monitoring systems.

Finally, with respect to multimodal monitoring, **C3** emphasises that sensing physical motion (**C1.1** and **C1.2**) or physiological signals (**C1.3**) in isolation provides only a partial view of a person's actual state. The survey in Section 2.5 and the comparison in Table 2.6 show that most driver-monitoring studies either focus on HR without detailed gaze metrics, or use blink and fixation features without jointly modelling HR, and often do so in simulated or constrained environments. Only a small number of studies use a richer set of gaze metrics, and none combine HR with blink, fixation, and saccade features in real-world driving using an interpretable fusion framework. Existing multimodal systems therefore only partially satisfy **C3**, as they do not yet provide transparent, adjustable mechanisms for combining behavioural and physiological indicators into a comprehensive, real-time measure of driver state in naturalistic conditions.

Within this landscape, the thesis is positioned to address these gaps through three contributions that are tightly aligned with **C1**, **C2**, and **C3**:

- **Contribution 1 (Chapter 3)** tackles the large-scale component of **C1.1** and the data-efficiency aspect of **C2**. It develops a data-efficient AI pipeline for radar-based human activity recognition using a public 5.8 GHz radar dataset. It systematically compares a set of radar representations with both compact and deeper CNN architectures under strict subject-wise evaluation. By jointly reporting recognition performance, preprocessing, and inference budgets, this contribution provides concrete design guidelines for selecting model–representation pairings that satisfy real-time constraints while remaining data-efficient. This directly addresses the need for principled model–representation selection in large-scale radar-based human monitoring.

- **Contribution 2 (Chapter 4)** addresses both the fine-grained component of **C1.2** and the physiological-scale metrics in **C1.3**, again under the data scarcity constraints of **C2**. The first part of the chapter develops a pipeline for classifying seated upper-body and head movements using a single mmWave FMCW radar. It introduces a targeted augmentation technique inspired by SpecAugment [40] to improve generalisation under limited data and subject-wise evaluation. This directly tackles the ambiguity of fine-grained TF signatures in **C1.2**. The second part of Chapter 4 integrates VMD into the radar processing chain to estimate physiological-scale metrics, mitigating motion artefacts and improving the stability of HR and RR estimation in the presence of motion, thereby addressing the challenge in **C1.3**. Together, these elements provide a unified radar framework for detailed behavioural and physiological monitoring.

- **Contribution 3 (Chapter 5)** responds to the multimodal challenge in **C3**. It proposes a real-time framework that fuses HR with gaze-based behavioural metrics, namely blink, fixation, and saccade features, into a single interpretable alertness score. HR, acquired from a dedicated physiological sensor, is aligned with gaze features over tunable analysis windows, and a lightweight decision-level fusion algorithm with adjustable weights and thresholds is used to generate an alertness measure that can be tuned to different operating points. The framework is validated on real-world driving data using both scenario-wise and subject-wise analyses, directly addressing the need for comprehensive, multimodal human-state assessment in dynamic, real-world environments.

In summary, Chapter 2 has identified that existing radar-based human monitoring solutions do not yet fully meet the requirements of **C1**, **C2**, and **C3**. They often operate at a single motion scale, rely on narrow datasets, and explore only limited forms of multimodal fusion. The remainder of this thesis is structured to address these gaps in a staged manner: Chapter 3 focuses on efficient and principled model–representation design for large-scale radar-based activity recognition, Chapter 4 extends this to fine-grained behaviour and motion-robust physiological-scale sensing, and Chapter 5 builds on these foundations to construct an interpretable multimodal framework for human-state monitoring in real dynamic environments.

Chapter 3

Radar-Based Large-Scale Human Activity Monitoring

This chapter presents the first contribution of this thesis and addresses challenges **C1.1** and **C2** outlined in Section 1.3 by evaluating a radar-based pipeline for large-scale Human Activity Monitoring (HAM) using FMCW radar. The chapter investigates different radar representations and learning model configurations to identify data-efficient solutions suitable for real-time activity recognition. Section 3.3 presents the system model and signal preprocessing, Section 3.4 describes the supervised learning framework, and Section 3.5 reports the experimental results and comparative evaluation.

3.1 Introduction

Large-scale HAM has become a core enabling capability across a wide range of application domains, including medical rehabilitation [219], security and surveillance systems [210], and ambient assisted living environments [3, 220]. Over the past decades, radar-based large-scale HAM has relied on a variety of sensing modalities, including video-based systems [221], infrared imaging [222], wearable devices [57], and infrastructure-based ambient sensors [223]. In smart environments, ambient sensing solutions such as motion detectors, energy consumption meters, and environmental sensors have been widely deployed to infer daily activities using indirect indicators, including room occupancy, appliance usage, and temperature variations.

As reviewed in Section 2.1, these sensing paradigms suffer from fundamental limitations that restrict their suitability for continuous, scalable, and robust monitoring. Vision-based systems are sensitive to illumination conditions, occlusions, and monitoring distance, while infrared sensors are affected by environmental temperature variations and exhibit limited sensitivity to radial motion components. Wearable devices, despite providing rich motion and physiological information, impose usability constraints related to user compliance, comfort, and battery lifetime. Ambient sensing approaches further require extensive infrastructure installation and calibration, and their reliance on indirect measurements often leads to an ambiguous interpretation of complex or overlapping activities, reducing robustness in dynamic environments.

On the other hand, radar-based sensing, detailed in Section 2.2, has emerged as an effective alternative to address these challenges, particularly for short-range indoor monitoring scenarios [84, 103, 104, 224]. Radar sensors enable contactless, privacy-preserving monitoring by capturing human motion without requiring wearable devices or line-of-sight. In addition to activity monitoring, radar systems have demonstrated the capability to capture physiological motion components such as respiration and heart rate [225, 226], making them well-suited for deployment in sensitive environments, including bedrooms, bathrooms, and healthcare facilities.

Radar echoes inherently encode information across time, range, and Doppler dimensions. As reviewed in Section 2.2.2 and evidenced by the literature surveyed in Table 2.2, radar-based large-scale HAM commonly relies on transforming raw radar signals into structured two-dimensional representations. Doppler maps obtained via time–frequency (TF) analysis [136, 224, 227] provide a compact abstraction of motion dynamics and are widely adopted for learning-based classification. Complementary spatial information can be captured using time–range (TR) representations derived from range FFT [126, 139]. Each representation offers distinct trade-offs between discriminative capability and computational cost.

In this chapter, multiple radar representations are investigated to support robust large-scale human monitoring. Specifically, three two-dimensional representations are considered: TR maps generated via range FFT, TF maps obtained using STFT [147], and high-resolution TF representations derived using SPWVD [143]. While STFT provides interpretable and computationally efficient TF representations, it is constrained by a fixed window length, leading to an inherent trade-off between time and frequency resolution. SPWVD offers improved resolution and reduced cross-term interference, but at the expense of significantly increased computational complexity, casting doubt on its suitability for real-time deployment.

The effectiveness of radar-based large-scale HAM is strongly influenced by both the selected signal representation and the adopted feature extraction strategy. Handcrafted feature extraction methods require domain expertise and are sensitive to noise and signal variability, limiting scalability and generalisation [84]. In contrast, deep learning approaches based on CNNs enable automated feature learning by jointly performing representation learning and classification, making them well-suited for processing radar-generated two-dimensional maps.

The success of CNNs in camera-based activity recognition and large-scale image classification, as overviewed in Section 2.1, has driven the development of several established architectures. Early studies demonstrated the effectiveness of deep convolutional models for visual recognition tasks [228], followed by substantial performance gains on benchmark datasets such as ImageNet [39]. Consequently, architectures such as VGGNet, ResNet, and MobileNet have become standard backbones and have been increasingly adopted for radar-based activity monitoring. Radar TF and TR representations can be processed analogously to image data, enabling the reuse of these architectures. However, training deep CNNs from scratch remains impractical for radar-based large-scale HAM due to limited labelled data, which often leads to overfitting and poor generalisation.

To address this limitation, this chapter adopts a transfer learning strategy [186, 229, 230], whereby CNNs pre-trained on large-scale image datasets are fine-tuned for radar-based activity classification. Transfer learning reduces data requirements while retaining strong representational capacity. Four CNN architectures are considered: VGG-16 [231], VGG-19 [232], ResNet-50 [233], and MobileNetV2 [234]. These models span a range of architectural complexities, enabling a systematic evaluation of accuracy–efficiency trade-offs for real-time radar-based large-scale HAM.

3.2 Chapter Contributions

This chapter investigates radar-based large-scale HAM through a systematic evaluation of radar preprocessing strategies and deep learning architectures, with particular emphasis on real-time feasibility and deployment on resource-constrained platforms. The proposed framework analyses the interaction between radar signal representation and learning model selection to identify configurations that balance recognition performance with computational efficiency. Three radar preprocessing techniques are combined with four CNN architectures, resulting in twelve distinct model–representation pairings evaluated under consistent experimental conditions.

The main contributions of this chapter are summarised as follows:

- **Comparative evaluation of radar representations:** TR maps derived from range FFT and TF representations generated using STFT and SPWVD are systematically evaluated. Their discriminative capability and computational cost are quantified to assess suitability for real-time radar-based large-scale Human Activity Monitoring.

- **Data-efficient learning via transfer learning:** Four CNN architectures, namely VGG-16, VGG-19, ResNet-50, and MobileNetV2, are fine-tuned using transfer learning to mitigate radar data scarcity. The impact of architectural complexity on recognition accuracy and computational efficiency is explicitly analysed.

- **End-to-end performance and efficiency analysis:** A comprehensive comparison of twelve model–representation pairings is conducted, considering recognition accuracy, precision, recall, and F1-score. Preprocessing time and inference latency are also analysed to identify configurations that satisfy real-time operational constraints.

- **All comparisons are performed under identical conditions**: Using a consistent subject-wise protocol and a fixed runtime environment to ensure reproducibility and fairness of comparison.

3.3 System Model for Large-Scale Human Activity Monitoring

The following discussion details the system model for radar-based large-scale HAM. It describes the end-to-end processing pipeline, beginning with radar data acquisition and progressing through signal transformation into multiple two-dimensional radar representations. Three complementary preprocessing strategies are investigated: TR maps, STFT-based TF representations, and SPWVD-based TF maps. These representations serve as structured inputs to deep learning models, where transfer learning is employed to enable accurate and computationally efficient classification of large-scale human movements under real-time constraints.

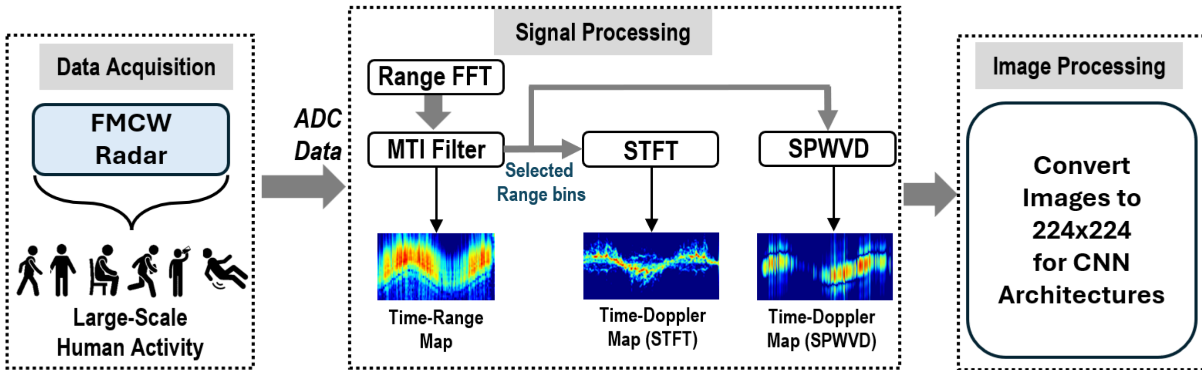


Figure 3.1: Radar data acquisition and signal processing for radar-based large-scale human activity monitoring.

3.3.1 Radar Data Acquisition

Despite substantial progress in radar-based large-scale HAM, a persistent limitation remains the scarcity of large, standardised datasets suitable for benchmarking classification algorithms. Many existing studies rely on proprietary datasets collected using different radar platforms and involving small, homogeneous participant groups, often limited to university students. This lack of demographic and environmental diversity constrains the development of robust, generalisable models and complicates fair comparisons across studies.

In this thesis, a publicly available FMCW radar dataset collected by the James Watt School of Engineering at the University of Glasgow, United Kingdom, is utilised [235]. The dataset has been widely adopted in recent literature [236], making it a suitable benchmark for evaluating radar-based large-scale HAM pipelines. Data were collected using a single-input single-output (SISO) FMCW radar operating at a carrier frequency of 5.8 GHz with a chirp bandwidth of 400 MHz.

A key strength of this dataset lies in its environmental and demographic diversity. Recordings were obtained from nursing home residents in their natural living environments, as well as from students and staff in laboratory and communal settings. Specifically, data were collected at the Glasgow Nursing Home (three rooms), the West Cumbria Elderly Centre (two rooms), and multiple locations within the University of Glasgow, including laboratories and common rooms. This diversity introduces realistic variations in background clutter, room geometry, and radar-to-subject distance, all of which are known to significantly influence radar signal characteristics. Unlike datasets collected exclusively under controlled laboratory conditions [182], this dataset captures non-ideal conditions, making it well-suited for evaluating model robustness.

In addition, the dataset includes recordings from 105 subjects, including 26 female participants, aged 21 to 98 years. This demographic spread enables the analysis of age- and sex-related variations in gait and activity patterns, which are often under-represented in datasets dominated by younger participants [84]. Although the dataset contains 1,748 labelled samples, which is modest relative to typical deep learning benchmarks, it captures substantial inter-subject variability, a critical factor for developing generalisable radar-based large-scale HAM models.

Table 3.1: Details of human activity classes.

Class	**Activity Description**	**Samples**	**Duration**
A1	Walking back and forth	312	10 s
A2	Sitting on a chair	311	5 s
A3	Standing	311	5 s
A4	Bending to pick an object	309	5 s
A5	Drinking water	311	5 s
A6	Fall	197	5 s

The dataset comprises six large-scale human activity classes, as detailed in Table 3.1, covering both routine daily movements and safety-critical events such as falls. A high-level overview of the complete radar-based processing pipeline, from data acquisition to radar representation generation, is illustrated in Figure 3.1.

Radar Configuration

Each radar chirp has a duration of 1 ms and consists of 128 ADC samples acquired at a sampling frequency of 128 kHz. The pulse repetition frequency (PRF) is 1 kHz, yielding a Doppler resolution of 1.25 Hz, which corresponds to a velocity resolution of approximately 3.2 cm/s. The achieved range resolution is 37.5 cm. Such a radar and its settings are sufficient for large-scale activities such as walking around the room. The complete set of radar configuration parameters used during data collection is summarised in Table 3.2.

Table 3.2: FMCW radar configuration used for data collection.

Radar parameter	**Value**
Chirp duration	1 ms
ADC samples per chirp	128
Pulse repetition frequency (PRF)	1 kHz
Range resolution	37.5 cm
Doppler resolution	1.25 Hz
Velocity resolution	3.2 cm/s
Maximum Doppler frequency	500 Hz
Maximum range	24 m

3.3.2 Radar Signal Model and Preprocessing Overview

The radar signal processing pipeline follows the framework introduced in Section 2.2.2. After analogue-to-digital conversion (ADC), the received intermediate-frequency (IF) signal is represented as a complex baseband signal composed of in-phase (I) and quadrature (Q) components. For a sequence of transmitted chirps, the digitised IF signal is organised into a two-dimensional structure indexed by fast time (range samples) and slow time (chirp index).

The discrete-time IF signal can be expressed as

$$x[n,m] = \sum_{p=1}^{P} A_p \exp\left\{ j2\pi \left(f_{\mathrm{b},p} n T_{\mathrm{s}} + f_{\mathrm{D},p} m T_{\mathrm{r}} \right) \right\} + w[n,m], \tag{3.1}$$

where n denotes the fast-time sample index, m is the slow-time (chirp) index, A_p represents the complex amplitude of the p-th scatterer, $f_{\mathrm{b},p}$ is the beat frequency associated with target range, and $f_{\mathrm{D},p}$ denotes the Doppler frequency associated with target motion. T_{s} and T_{r} represent the ADC sampling period and chirp repetition interval, respectively, and $w[n,m]$ denotes additive noise.

Following the formulation in Chapter 2, the complex IF samples are reshaped into a two-dimensional matrix of size $N_{\mathrm{ADC}} \times N_{\mathrm{chirp}}$, where each column corresponds to a single chirp. This representation enables systematic processing along the range and Doppler dimensions. Subsequent preprocessing steps, including range FFT, time–frequency analysis, and clutter suppression, are applied to extract discriminative motion signatures suitable for radar-based large-scale HAM. These steps form the basis for generating the TR, STFT-based TF, and SPWVD representations examined in the following subsections.

TR Map Generation

Starting from the discrete-time IF signal model in Eq. (3.1), TR map generation focuses on extracting range information as a function of time. As illustrated in Figure 3.1, this is achieved by applying a windowed FFT along the fast-time dimension (ADC samples) for each transmitted chirp. Prior to the FFT, a Hamming window is applied to reduce spectral leakage.

Let $x[n,m]$ denote the complex IF signal, where $n = 0,\ldots,N_{\text{ADC}} - 1$ indexes fast time and $m = 1,\ldots,N_{\text{tot}}$ indexes chirps. The range profile for the m-th chirp is obtained as

$$X_m[k] = \sum_{n=0}^{N_{\text{ADC}}-1} w[n]\, x[n,m]\, e^{-j2\pi \frac{kn}{N_{\text{ADC}}}}, \tag{3.2}$$

where k denotes the range bin index and $w[n]$ is the Hamming window. The resulting sequence $\{X_m[k]\}$ represents the time-varying range information between the radar and the subject.

To suppress static clutter and isolate motion-related components, a Moving Target Indicator (MTI) filter is applied along the slow-time dimension for each range bin. The MTI is implemented as a fourth-order Butterworth high-pass filter with a cut-off frequency of 0.0075 Hz. After MTI filtering, the TR maps are constructed by stacking the filtered range profiles across successive chirps. Representative TR maps for the six activity classes are shown in Figure 3.2.

By visually inspecting the TR maps, a subset of range bins k^* corresponding to the dominant target reflections is identified. These range bins are subsequently used as inputs to the TF analysis methods described in the following, as indicated in Figure 3.1.

STFT-based TF Map Generation

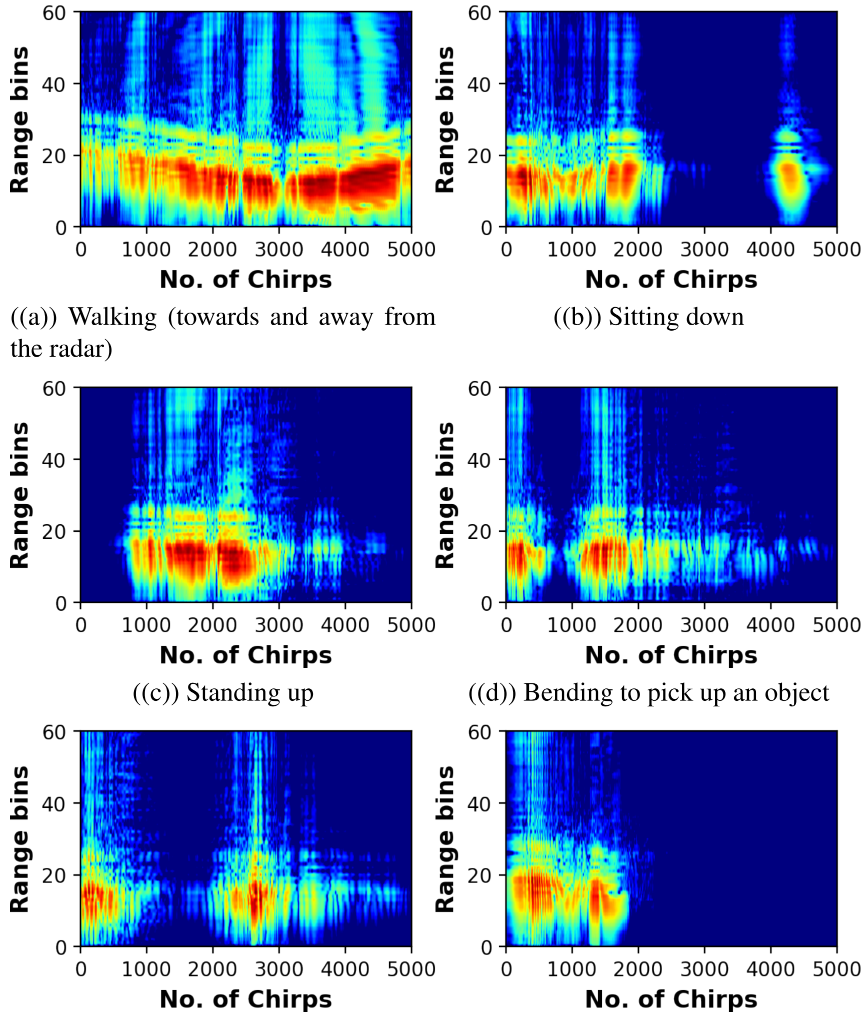
To capture the velocity dynamics of human motion, TF maps are generated by applying STFT to the MTI-filtered range bins of interest k^*. This step corresponds to the time–frequency processing block in Figure 3.1.

For each selected range bin k^*, the slow-time signal $X_m[k^*]$ is segmented using a Hann window of length $s = 200$ samples, with a hop size of $R = 10$ samples (corresponding to 95% overlap). Zero-padding is applied to achieve an FFT length of 800 samples, yielding improved frequency resolution. The STFT is computed as [143]

$$\Theta_{\text{STFT}}[m,f] = \sum_{n=-\infty}^{\infty} X_n[k^*]\, g_s[n-mR]\, e^{-j2\pi f n}, \tag{3.3}$$

where $g_s[\cdot]$ denotes the Hann window and f is the Doppler frequency index. The TF map is obtained by taking the magnitude of $\Theta_{\text{STFT}}[m,f]$.

STFT-based TF maps provide an interpretable representation of motion-induced Doppler variations over time. However, their resolution is constrained by the fixed window length, resulting in an inherent trade-off between time and frequency resolution. Representative STFT-based



((a)) Walking (towards and away from the radar)

((b)) Sitting down

((c)) Standing up

((d)) Bending to pick up an object

((e)) Drinking water

((f)) Fall

Figure 3.2: Representative TR maps for the six activity classes from a young adult participant.

TF maps for the six activity classes are shown in Figure 3.3.

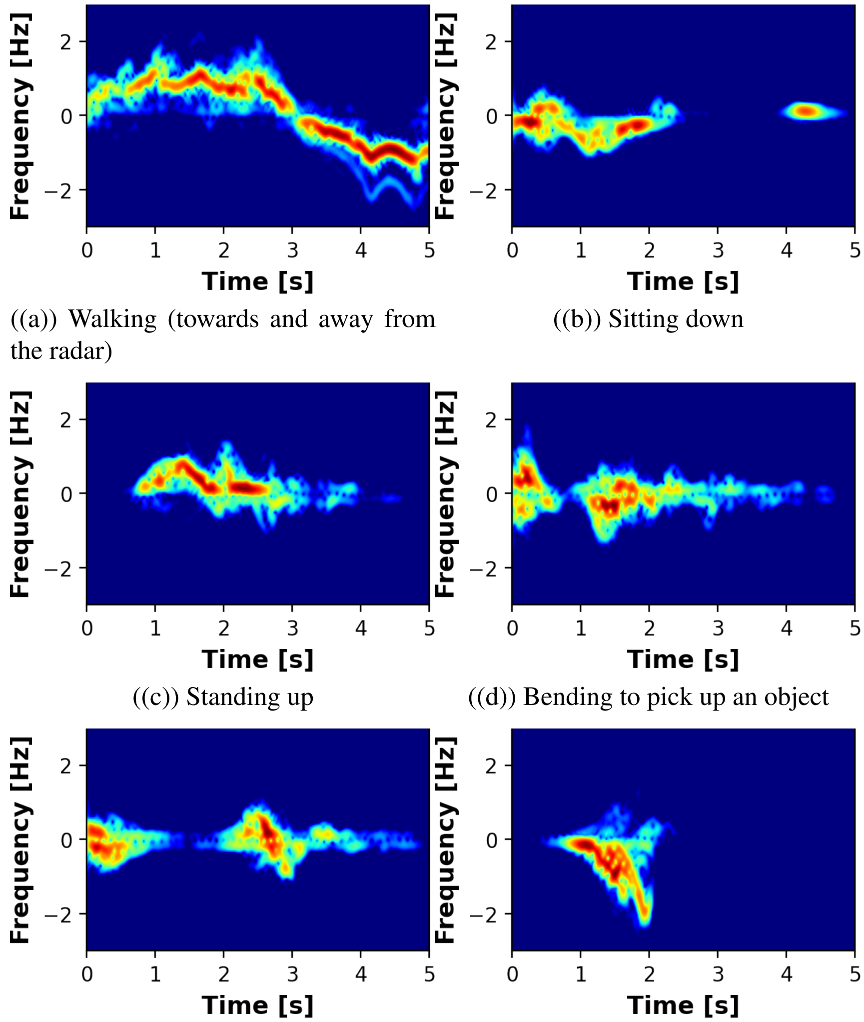
SPWVD-based TF Map Generation

To overcome the resolution limitations of STFT, SPWVD is employed as a higher-resolution TF analysis technique. SPWVD is derived from the Wigner–Ville Distribution (WVD), which offers excellent time–frequency resolution but suffers from cross-term interference. SPWVD mitigates this issue by introducing independent smoothing operations in the time and frequency domains [143].

Applied to the selected range bins $X_n[k^*]$, the SPWVD is defined as

$$\Theta_{\mathrm{SPWVD}}[m,f] = \sum_{n=-\infty}^{+\infty} \sum_{\tau=-\infty}^{+\infty} X\left(k^* + \frac{\tau}{2}, n\right) X^*\left(k^* - \frac{\tau}{2}, n\right) h(\tau)\, w(n-m)\, e^{-j2\pi f \tau}, \tag{3.4}$$

where $h(\tau)$ and $w(n-m)$ denote the frequency- and time-smoothing windows, respectively. In



((a)) Walking (towards and away from the radar)

((b)) Sitting down

((c)) Standing up

((d)) Bending to pick up an object

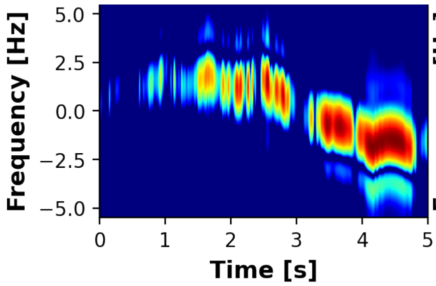
((e)) Drinking water

((f)) Fall

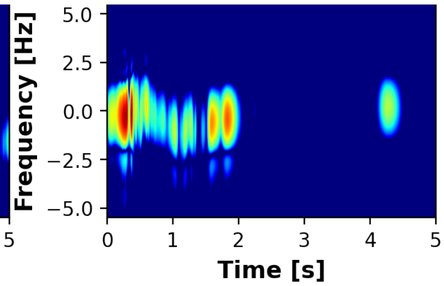
Figure 3.3: Representative STFT maps for the six activity classes from a young adult participant.

this study, a Hann window of length 15 is used for frequency smoothing, and a Kaiser window of length 25 is used for time smoothing. The index m represents the time frame. To further stabilise the representation and avoid intensity spikes, the SPWVD outputs are normalised by the number of selected range bins. This process yields high-resolution TF maps that preserve fine velocity variations while significantly reducing cross-term artefacts. Representative SPWVD-based TF maps for the six activity classes are shown in Figure 3.4.

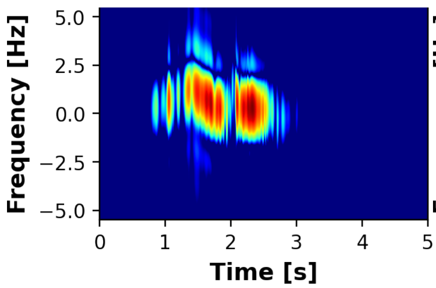
These representative maps illustrate why the choice of radar representation matters for large-scale activity recognition. TR maps preserve range displacement and are therefore useful for activities involving clear body translation, while STFT-based TF maps expose time-varying Doppler patterns that are more directly related to limb and body velocity. SPWVD can provide sharper TF localisation, making transient events such as falls more visually distinguishable; however, this benefit must be weighed against its higher preprocessing cost and the possibility of cross-term artefacts. These differences are important because activities such as sitting, stand-



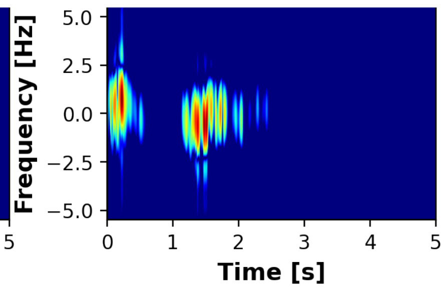
((a)) Walking (towards and away from the radar)



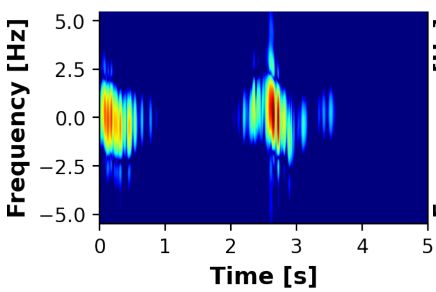
((b)) Sitting down



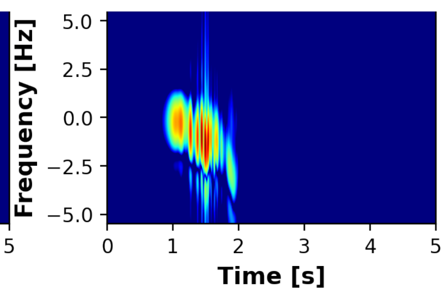
((c)) Standing up



((d)) Bending to pick up an object



((e)) Drinking water



((f)) Fall

Figure 3.4: Representative SPWVD maps for the six activity classes from a young adult participant.

ing, and drinking may exhibit limited range displacement and partially overlapping low-velocity signatures, whereas walking and falling produce broader Doppler spreads and more abrupt high-energy patterns. Therefore, the representations shown in Figures 3.2, 3.3, and 3.4 are not merely illustrative examples, but provide the empirical motivation for comparing representation–model pairs rather than selecting a single radar representation a priori. This comparative analysis is essential for identifying representations that provide an appropriate balance between discriminative capability and computational efficiency for real-time radar-based large-scale HAM.

3.4 Model Training and Optimisation for Large-Scale Human Activity Monitoring

Once the radar signals are preprocessed (Section 3.3), they are integrated into the deep learning models via the training procedures described here. As illustrated in Figure 3.5, the learning pipeline consists of three main stages: data preparation, transfer learning-based model configuration and optimisation, and supervised training with subject-wise evaluation. The adopted methodology is designed to ensure robust generalisation to unseen individuals while enabling a systematic comparison between radar representations and learning architectures.

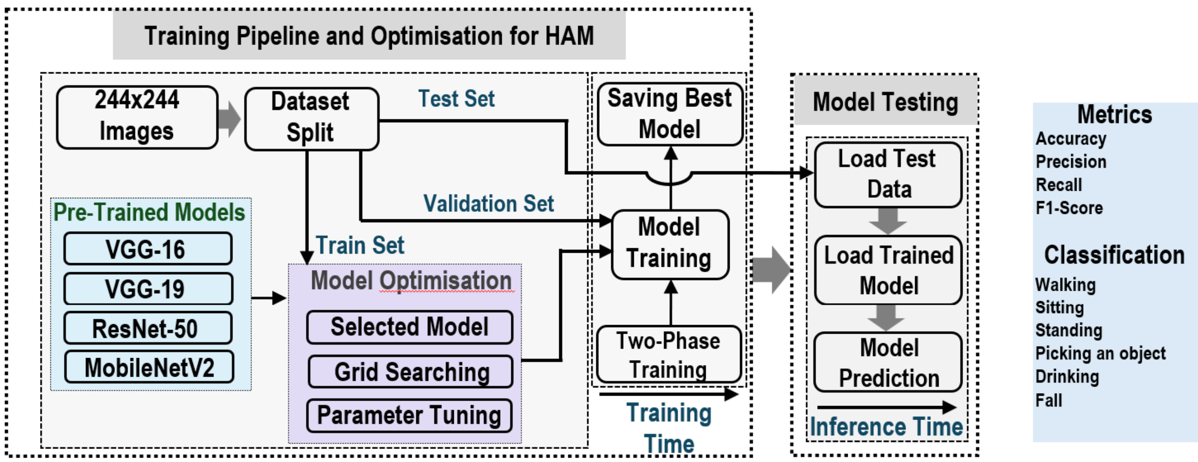


Figure 3.5: Model training and optimisation procedure for evaluating radar-based large-scale human activity monitoring models.

3.4.1 Data Preparation and Dataset Splitting

The first stage of the training pipeline prepares radar representations for input to CNNs. All radar maps generated during preprocessing, including TR, STFT-based TF, and SPWVD-based TF representations, were converted into image form. Each image was resized to a uniform resolution of 224×224 pixels with three channels to satisfy the input requirements of the selected CNN architectures. Pixel intensities were normalised, and the mean RGB value computed from the training data was subtracted to improve optimisation stability and convergence behaviour. To support a realistic and unbiased assessment of cross-subject generalisation, a strict subject-wise partitioning strategy was adopted, ensuring that no participant appeared in more than one subset. The University of Glasgow FMCW radar dataset comprises 1,748 samples collected from 105 subjects, which represents a comparatively broad participant base for radar-based device-free activity recognition. This is particularly relevant because several device-free sensing studies have relied on substantially smaller cohorts, sometimes involving only one or two participants, thereby limiting the strength of generalisation claims [237].

The dataset was divided into three fixed and disjoint subsets. Seventy-nine subjects were assigned to the training set for model learning and hyperparameter optimisation, corresponding to approximately 75% of the participants; 19 subjects were used for validation and model selection, corresponding to approximately 18%; and the remaining 7 subjects were reserved exclusively for final testing, corresponding to approximately 7%. This partitioning resulted in 1,286 training samples, 336 validation samples, and 126 testing samples. Since the test subjects were completely unseen during training, validation, and optimisation, the final evaluation should be interpreted as unseen-subject testing within a 105-subject dataset, rather than as a study based on only a few participants. The split was kept fixed across all preprocessing domains and model architectures to ensure reproducibility and fair comparison, although the limited number of held-out test subjects remains a factor to consider when interpreting the final performance. The resulting class-wise sample distribution across the three partitions is summarised in Table 3.3.

Table 3.3: Summary of activity classes and sample distribution across dataset partitions.

Classes	**Class**	**Training samples**
Walking back and forth	A1	232
Sitting on a chair	A2	234
Standing	A3	234
Picking up an object	A4	232
Drinking	A5	234
Fall	A6	120
Training		1286 samples
Validation		336 samples
Testing		126 samples

Class imbalance, particularly for the fall activity, was addressed using two complementary strategies. First, data augmentation was applied exclusively to the training set using the Keras data generator, introducing controlled variability while preserving the semantic structure of the radar maps. The validation and test sets were left unchanged to ensure a fair evaluation. Second, class weights computed using the scikit-learn library were incorporated into the loss function during training, encouraging the model to assign greater importance to under represented classes.

3.4.2 CNN Pre-Trained Models and Transfer Learning Strategy

The learning framework is based on transfer learning, a deep learning paradigm in which a model trained for one task is reused as a starting point for a related task. This approach is particularly well-suited to radar-based large-scale HAM, where labelled datasets are relatively small and training deep networks from scratch can lead to severe overfitting. All models were initialised with weights pre-trained on the ImageNet dataset [39]. Four representative CNN architectures were investigated:

Visual Geometry Group (VGG)

The VGG family of convolutional neural networks, including VGG-16 and VGG-19, was introduced by the Visual Geometry Group at the University of Oxford and has become a foundational benchmark in deep learning for visual recognition tasks [231]. These architectures are characterised by their uniform and highly regular design, relying exclusively on stacked 3×3 convolutional filters combined with max-pooling layers to progressively increase the receptive field while maintaining manageable computational complexity.

VGG-16 consists of 16 learnable weight layers, while VGG-19 extends this design to 19 layers by adding additional convolutional blocks [232]. The depth of these architectures enables the hierarchical extraction of features, ranging from low-level spatial patterns in early layers to increasingly abstract and discriminative representations in deeper layers. On the ImageNet benchmark, VGG-16 achieved a top-5 classification accuracy of approximately 92.7%, demonstrating its strong representational capacity despite its relatively simple architectural principles.

In the context of radar-based large-scale HAM, the VGG architectures are particularly well-suited to learning structured spatial patterns from two-dimensional radar representations such as TR and TF maps. Their deep, sequential convolutional structure allows the network to capture extended motion trajectories and spatial correlations that arise from large-scale body movements such as walking, sitting, and falling. However, this representational power comes at the cost of increased parameter count and computational load, which motivates a careful evaluation of their suitability for real-time deployment.

ResNet-50

Residual Networks (ResNets) were proposed to address the degradation problem commonly observed when increasing the depth of conventional CNNs, where deeper models suffer from optimisation difficulties and reduced performance [233]. The defining feature of ResNet architectures is the introduction of residual connections, which allow the network to learn residual mappings with respect to the input of a given layer. These skip connections facilitate gradient propagation during backpropagation and enable the effective training of substantially deeper networks.

In this work, ResNet-50 was selected as a representative deep residual architecture, comprising 50 layers organised into residual blocks with identity and projection shortcuts. This depth allows ResNet-50 to model highly complex feature hierarchies and capture subtle temporal–spatial patterns embedded within radar representations. Such capability is particularly relevant for distinguishing between activities with overlapping motion characteristics, where small variations in Doppler signatures or temporal evolution play a critical role.

For large-scale HAM, ResNet-50 provides a strong balance between depth and stability, offering enhanced discriminative power compared to shallower networks while maintaining reliable convergence during training. Nevertheless, its increased architectural complexity introduces higher computational demands, which must be weighed against performance gains when targeting latency-sensitive or resource-constrained applications.

MobileNetV2

MobileNetV2 is a lightweight convolutional neural network architecture specifically designed for efficient inference on mobile and embedded platforms [234]. Unlike traditional CNNs, MobileNetV2 employs depthwise separable convolutions, which decompose standard convolutions into depthwise and pointwise operations. This design significantly reduces the number of parameters and floating-point operations without substantially compromising representational capacity.

A key innovation in MobileNetV2 is the use of inverted residual blocks combined with linear bottlenecks [234]. In this structure, the network expands feature dimensionality before applying depthwise convolutions and then projects the features back to a lower-dimensional space using linear transformations. This approach preserves informative features while avoiding excessive information loss caused by nonlinear activations in low-dimensional spaces.

Within the scope of radar-based large-scale HAM, MobileNetV2 is particularly attractive due to its low inference latency and reduced memory footprint. These characteristics make it well-suited for real-time activity recognition scenarios, including safety-critical applications such as fall detection, where rapid response is essential. Although MobileNetV2 is less expressive than deeper architectures such as VGG or ResNet, its efficiency-oriented design allows it to achieve competitive performance when paired with informative radar representations, making it a strong candidate for edge and embedded deployment.

3.4.3 Model Optimisation and Hyperparameter Selection

Model optimisation was performed using an experimental grid search to determine hyperparameter configurations that achieved high classification accuracy while maintaining stable convergence. The search was conducted independently for each radar representation (TR, STFT, and SPWVD) in order to account for the distinct statistical and structural characteristics of the corresponding radar domains. The grid search explored key hyperparameters known to influence learning dynamics and generalisation, including learning rate, batch size, dropout rate, optimiser selection, and the number of neurons in the dense layers. Candidate configurations were evaluated based on validation performance, and the final settings were selected based on classification accuracy and training stability. The grid search was not performed independently for all twelve model–representation pairs. Instead, hyperparameter tuning was first conducted separately for each radar representation domain (TR, STFT, and SPWVD), and the resulting configurations were then applied consistently across the evaluated CNN architectures, with only minor adjustments introduced when required for stable convergence. This strategy ensured fair comparison between model architectures while accounting for the different statistical properties of the radar representations.

Consistency was maintained across radar representations and model architectures (VGG-16, VGG-19, ResNet-50, and MobileNetV2), with adjustments introduced only where empirically justified. To mitigate overfitting and minimise training loss, batch normalisation and dropout layers were strategically incorporated before the flattening stage and after dense layers. All models employed categorical cross-entropy as the loss function, consistent with the multi-class classification objective.

For VGG-16, a dense layer with 512 neurons was utilised. For VGG-19 and ResNet-50, 1024 neurons were employed. All dense layers used ReLU activation with `he_normal` kernel initialisation. An additional dropout layer was added after the final dense layer to enhance generalisation. To adapt model capacity to the characteristics of each radar representation, the number of neurons in the fully connected layers was adjusted accordingly. The TR representation, which exhibited higher noise levels and lower structural clarity, utilised 2048 neurons. The STFT representation, offering intermediate time–frequency resolution, used 1024 neurons. The SPWVD representation, characterised by sharper feature localisation and higher spectral resolution, used 512 neurons. These adjustments aimed to balance representational capacity and overfitting risk according to domain complexity.

For MobileNetV2, a similar grid search process was applied. However, empirical experimentation indicated that the Adam optimiser provided more stable convergence and improved generalisation compared to SGD for radar-based representation maps. Initial experiments using optimisation settings identical to the VGG and ResNet architectures achieved competitive accuracy but exhibited signs of overfitting. Adjusting the optimiser and learning rate mitigated this behaviour and resulted in improved validation stability.

Table 3.4: Pre-trained CNN hyperparameters.

Hyperparameter	**VGG-16**	**VGG-19**	**ResNet-50**	**MobileNetV2**
Batch size	32	32	32	32
Dropout	0.5	0.2	0.2	0.2
Learning rate	2×10^{-3}	2×10^{-3}	2×10^{-3}	4×10^{-4}
Optimiser	SGD	SGD	SGD	Adam
Weight decay	—	10^{-6}	10^{-6}	—
Momentum	0.9	0.9	0.9	—

Model Training Procedure

After hyperparameter selection, the final model was trained using the optimised configurations. All models were trained end-to-end in a single phase. The ImageNet pre-trained backbones were fully unfrozen, and all layers were set as trainable, allowing complete fine-tuning of the network weights to adapt the learned visual representations to radar representation map data.

Training was conducted for up to 50 epochs. Early stopping was employed with a patience of 8 epochs to prevent overfitting and unnecessary computation. Validation accuracy was monitored during training, and model optimisation was guided by this metric. Models were trained on the 79-subject training set, with performance monitored on the independent 19-subject validation set. The validation set was used exclusively for performance monitoring and model selection. For each model–representation pairing, the checkpoint achieving the highest validation accuracy was retained. Final evaluation was conducted once on the held-out test set, which remained completely unseen during both optimisation and training stages. This strict separation between training, validation, and test subjects ensures an unbiased estimate of inter-subject generalisation performance.

3.4.4 Performance Evaluation

The performance of the proposed radar-based large-scale HAM framework is evaluated using a set of complementary classification metrics that jointly characterise recognition accuracy, class-level reliability, and robustness under class imbalance. All evaluation metrics are computed exclusively on the held-out test set comprising data from seven unseen subjects, ensuring a subject-independent assessment of generalisation performance.

Four standard metrics are employed, namely accuracy, precision, recall, and F1-score. These metrics are derived from the confusion matrix by comparing predicted activity labels against ground-truth annotations, resulting in four possible outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

Accuracy measures the overall correctness of the classifier and is defined as the ratio of correctly classified samples to the total number of samples:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \times 100\%. \tag{3.5}$$

Precision quantifies the reliability of positive predictions by measuring the proportion of correctly identified positive samples among all samples predicted as positive:

$$\text{Precision} = \frac{TP}{TP+FP}. \tag{3.6}$$

Recall evaluates the sensitivity of the classifier and reflects its ability to correctly identify all relevant samples belonging to a given class:

$$\text{Recall} = \frac{TP}{TP+FN}. \tag{3.7}$$

The F1-score represents the harmonic mean of precision and recall, providing a balanced metric that accounts for both false positives and false negatives:

$$\text{F1-score} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}}. \tag{3.8}$$

To account for class imbalance across the six activity categories, precision, recall, and F1-score are reported using macro-averaging. Macro-averaged precision, recall, and F1-score were computed using the scikit-learn evaluation functions under a one-vs-rest formulation for the multi-class setting. This approach assigns equal weight to each class, ensuring that performance on minority classes, such as falls, is not masked by dominant activities.

3.4.5 Computational Efficiency

Beyond recognition performance, runtime behaviour is central to deployment in continuous monitoring and safety-critical scenarios. We therefore report both training time (as an indicator of optimisation cost) and inference time (as an indicator of online latency) for each model–representation pair under a fixed hardware and software configuration.

- **Preprocessing time:** This metric quantifies the computational cost required to convert raw radar signals into CNN-compatible representations. It includes clutter suppression, range processing, representation transformation (TR, STFT, or SPWVD), resizing, and normalisation. Preprocessing time reflects the overhead introduced by signal representation and is measured per input sample to enable fair comparison across radar representations.

- **Training time:** This metric represents the average time required to complete one training epoch for a given model–representation pairing. Training time reflects the computational cost associated with model optimisation and is relevant in scenarios where models require retraining or adaptation to new environments.

- **Inference time:** Inference time captures the latency associated with generating a single activity prediction from a trained model. It is measured as the elapsed time between issuing an inference request and obtaining the predicted activity label:

  $$T_{\text{inference}} = T_{\text{end}} - T_{\text{start}}, \tag{3.9}$$

  where T_{start} and T_{end} denote the timestamps corresponding to the initiation and completion of the inference process, respectively.

Inference time is reported per sample and averaged across multiple runs to mitigate transient computational fluctuations. This metric is particularly important for latency-sensitive applications, such as fall detection, where rapid response is essential.

Proposed Algorithm

The proposed radar-based large-scale HAM framework was designed around three criteria: discriminative capability, data efficiency, and computational efficiency. TR, STFT, and SPWVD representations were selected because they capture complementary aspects of FMCW radar data, namely range displacement, time-varying Doppler content, and higher-resolution TF localisation, respectively. Transfer learning was adopted because the dataset, although relatively broad in subject coverage, remains modest compared with the parameter count of deep CNN architectures. Reusing pre-trained feature extractors, therefore, helps reduce the amount of radar-specific labelled data required for effective model training. MobileNetV2 was included to assess whether a compact architecture could preserve recognition performance while reducing inference latency, whereas VGG-16, VGG-19, and ResNet-50 provide heavier comparative baselines.

The evaluation protocol is therefore intended to expose the accuracy–cost trade-off that governs deployment on constrained or real-time platforms. Rather than proposing a new network architecture, this chapter systematically pairs established ImageNet-pretrained CNN backbones with three radar representations. This is important because many radar-HAR studies evaluate a single representation–model configuration and focus primarily on recognition accuracy [86, 118, 126]. In contrast, the contribution of this chapter is a controlled comparison of twelve representation–model pairs using recognition accuracy, preprocessing time, training time, and inference latency jointly. Algorithm 1 summarises this complete processing pipeline, from raw radar signal acquisition to activity inference, and shows that the framework is not designed solely to maximise accuracy, but to achieve a practical balance between classification performance and computational efficiency for real-time radar-based human monitoring.

Algorithm 1: Algorithm for evaluating radar representation–CNN model pairs, including accuracy and computational cost metrics.

Require: Raw FMCW radar data from multiple subjects

Ensure: Predicted activity labels, with accuracy metrics and computational costs (preprocessing, training, inference) reported per model–representation pair

```
Radar data acquisition
Acquire raw FMCW radar signals and convert to complex baseband (I/Q) data
Radar signal preprocessing and representation construction
Apply clutter suppression and range processing
Identify target range bins of interest
Select a radar representation:
    Option 1: TR maps
    Option 2: STFT-based TF maps
    Option 3: SPWVD-based TF maps
Resize and normalise representations to the CNN input format
Record preprocessing time for each representation pipeline
Subject-wise dataset partitioning
Split the dataset into training, validation, and testing sets with no subject overlap
    for each model–representation pair do
    Hyperparameter optimisation
Perform a grid search on the training set
Select hyperparameters using validation performance
Model initialisation and fine-tuning
Initialise the CNN backbone with ImageNet weights
Train/fine-tune the network using the selected hyperparameters
Record training time (average time per epoch) and retain the best checkpoint
Final evaluation
Evaluate the best checkpoint on the held-out test set
Record accuracy metrics
Record inference time

Activity inference
Apply the selected trained model to unseen radar data and output predicted activity labels
```

3.4.6 Runtime Environment

All experiments were conducted using a GPU-accelerated workstation to ensure reproducible and consistent performance measurements. Radar signal preprocessing and deep learning model training were implemented using Python 3.9.0. The system was equipped with an 11th-generation Intel Core i7-11700 processor featuring eight cores and sixteen threads, operating at a base frequency of 2.50 GHz, along with 16 GB of system memory.

Graphical processing was supported by an NVIDIA GeForce RTX 3060 Ti GPU with 8 GB of dedicated memory. Radar signal processing tasks were implemented using standard scientific Python libraries, including `SciPy` for signal processing operations and `fftpack` for fast Fourier transforms. Time–frequency analysis was performed using dedicated time–frequency processing libraries.

Deep learning model development and training were carried out using TensorFlow 2.10.0 and Keras 2.8.0, leveraging GPU acceleration for efficient optimisation and inference. All reported timing measurements were obtained under this fixed runtime configuration to ensure fair comparison across model–representation pairings.

3.5 Results and Discussion

This section reports subject-wise test performance and runtime measurements for all twelve model–representation pairs under the fixed runtime environment described in the previous Section. The evaluation considers all CNN architectures introduced in Section 3.4.2, using the radar representations described in Section 3.3 as model inputs. The analysis focuses on two complementary aspects, namely classification performance and computational efficiency, with the aim of assessing the suitability of different model–representation pairings for real-world deployment scenarios.

All reported results are obtained under a subject-wise evaluation protocol, where model training and hyperparameter tuning are conducted exclusively on a disjoint set of subjects, and final performance is assessed on unseen test subjects. This protocol ensures that the reported metrics reflect true generalisation across individuals rather than memorisation of subject-specific motion patterns. Performance metrics are reported using macro-averaging across all activity classes to account for class imbalance and to provide an unbiased assessment of recognition performance. Inference times are reported as the average over five independent runs to mitigate the effect of system-level variability.

3.5.1 Performance Comparison of Large-Scale Human Activity Monitoring Models

The following analysis compares the classification performance of the twelve evaluated model–representation pairs (MRPs), denoted as M1 to M12, as defined in Section 3.4.4. Each MRP corresponds to a specific combination of radar preprocessing technique and CNN architecture. The quantitative performance results, including macro-averaged accuracy, precision, recall, and

F1-score, are summarised in Table 3.5.

All classification results were obtained as the mean of five independent training runs using different random initialisations. The variation across runs was small, with standard deviations below 0.5% for all MRPs and, in most cases, corresponding to a difference of no more than 2 test samples. The ranking of the best-performing configurations remained unchanged across all runs, indicating that the reported performance differences are not attributable to random seed effects. For readability, Table 3.5 reports only the mean values, while the consistently low variance across runs confirms that the relative performance differences between MRPs are stable, which was the goal of the extensive hyperparameter selection in this work to ensure a fair comparison.

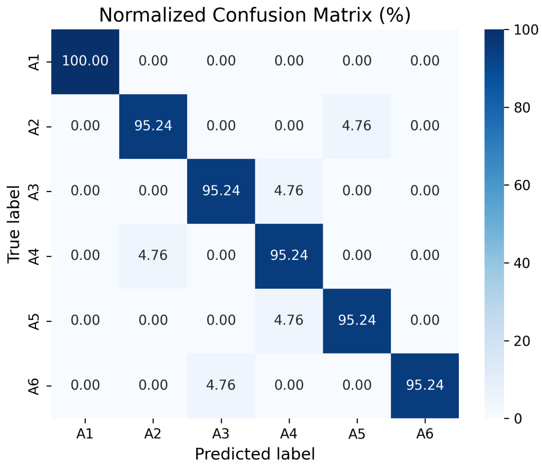
Table 3.5: Comparison of evaluation metrics for model–domain pairs across radar representations.

MRP	Radar representation	Model	Accuracy (%)	Precision	Recall	F1-score
M1	**TR**	**VGG-16**	**95.73**	**0.9576**	**0.9573**	**0.9572**
M2	TR	VGG-19	94.30	0.9436	0.9430	0.9429
M3	TR	ResNet-50	93.73	0.9373	0.9373	0.9368
M4	TR	MobileNetV2	92.88	0.9307	0.9288	0.9284
M5	STFT	VGG-16	96.87	0.9697	0.9687	0.9687
M6	STFT	VGG-19	96.01	0.9639	0.9624	0.9624
M7	**STFT**	**ResNet-50**	**97.15**	**0.9721**	**0.9731**	**0.9721**
M8	STFT	MobileNetV2	96.30	0.9635	0.9651	0.9642
M9	SPWVD	VGG-16	97.44	0.9764	0.9744	0.9745
M10	**SPWVD**	**VGG-19**	**98.01**	**0.9803**	**0.9801**	**0.9801**
M11	SPWVD	ResNet-50	97.15	0.9720	0.9715	0.9715
M12	SPWVD	MobileNetV2	96.01	0.9629	0.9580	0.9600

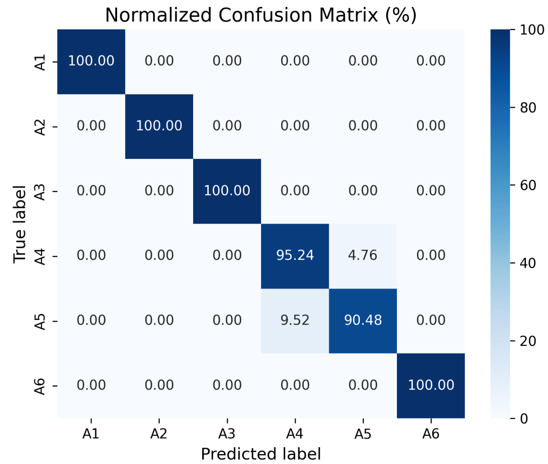
Across the evaluated configurations, three MRPs emerged as the strongest performers in terms of overall recognition accuracy within their respective radar representations. Specifically, M1 (TR with VGG-16), M7 (STFT with ResNet-50), and M10 (SPWVD with VGG-19) achieved the highest accuracies for the TR, STFT, and SPWVD representations, respectively. This observation highlights the strong interaction between the choice of radar representation and network architecture, rather than the dominance of a single model across all representations.

To gain deeper insight into class-wise performance, confusion matrices were analysed for the best-performing MRPs across the three radar maps. Figure 3.6 illustrates the confusion matrices corresponding to the best runs of M1, M7, and M10.

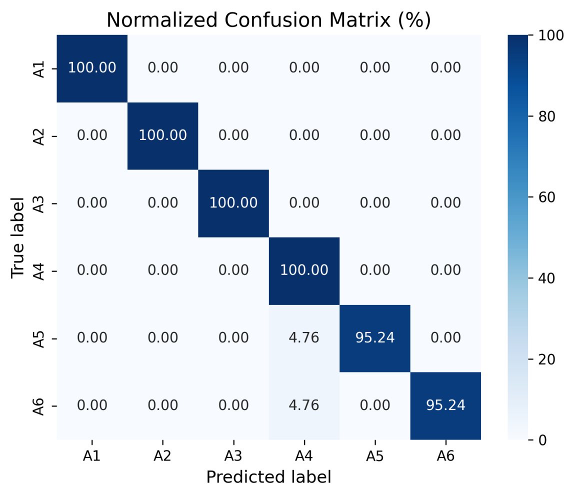
A consistent trend across these configurations is the reliable recognition of large-scale dynamic activities. In particular, fall detection (A6) was identified with perfect accuracy by M7



((a)) RT input with VGG-16v(M1)



((b)) STFT input with ResNet-50 (M7)



((c)) SPWVD input with VGG-19 (M10)

Figure 3.6: Confusion matrices for the best-performing model–representation pairs using pre-trained CNN backbones.

(STFT), while M1 (TR) and M10 (SPWVD) achieved recognition rates of 95.24%. Walking activity (A1) was also robustly detected, achieving 100% accuracy for M1, M7 and M10. These results confirm the effectiveness of radar-based monitoring systems in capturing distinctive motion characteristics; however, performance varies across different radar maps and CNN models.

In contrast, reduced classification performance was observed for activities involving more subtle and localised motion patterns, most notably A4 (bending to pick up an object) and A5 (drinking water). These activities exhibit similar temporal and spatial motion signatures, primarily driven by upper-body and arm movements, which can overlap significantly in time-frequency radar representations. This effect is particularly evident in SPWVD- and STFT-based representations, where enhanced time-frequency resolution does not improve the separation of the two classes compared to the TR map.

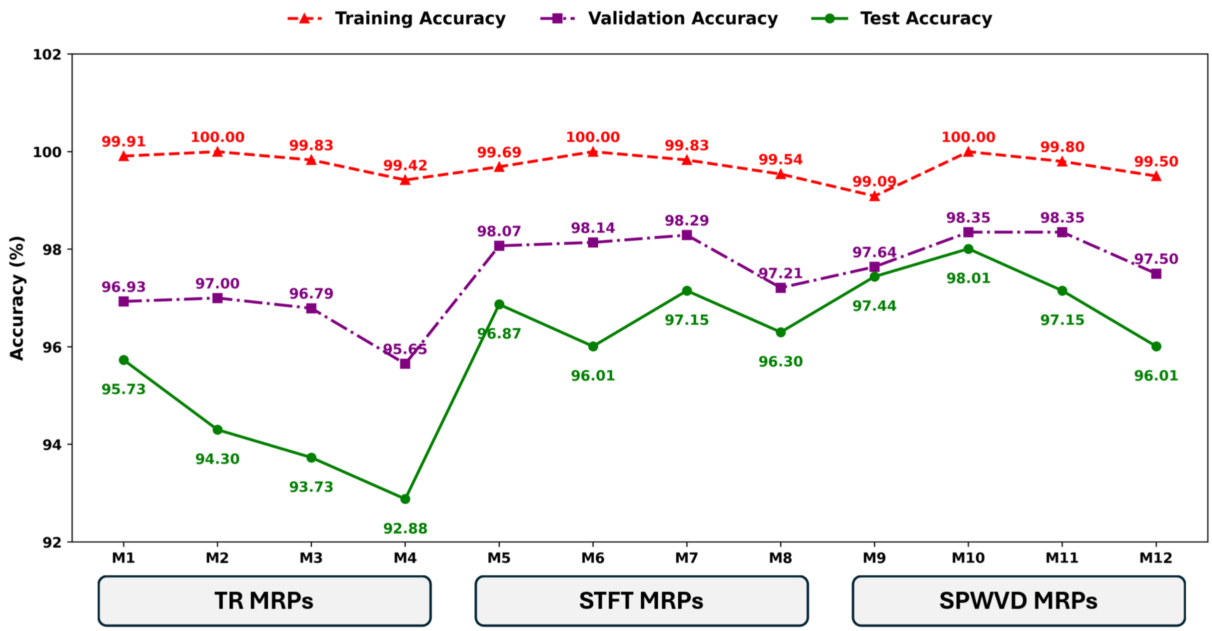


Figure 3.7: Comparison of training, validation, and test accuracies across all model–representation pairs in the proposed large-scale human activity monitoring system.

These findings indicate that classification performance is influenced not only by model capacity but also by the intrinsic separability of activity signatures within a given radar representation. Consequently, the choice of preprocessing technique plays a critical role in determining which motion characteristics are emphasised, directly affecting inter-class separability and overall recognition performance.

3.5.2 Generalisation Performance of Large-Scale Human Activity Monitoring Models

Evaluating the generalisation capability of radar-based large-scale HAM systems is essential, particularly when working with limited datasets where deep learning models are prone to overfitting. In this study, generalisation is assessed under a subject-wise evaluation protocol, in which all test samples are from subjects that were completely unseen during both training and hyperparameter tuning. This setting provides a realistic assessment of how well trained models transfer to new individuals, a critical requirement for practical deployment in real-world monitoring scenarios. The generalisation performance of all twelve MRPs is illustrated in Figure 3.7.

Across all configurations, the test accuracies fall within a narrow range, spanning from 92.88% to 98.01%, as summarised in Table 3.5. This consistency indicates that the evaluated models maintain stable performance when exposed to unseen subjects, despite differences in network architecture and radar preprocessing technique. Among the evaluated configurations, pair M10 achieved the highest test accuracy of 98.01%, reflecting the strong discriminative capability of SPWVD representations when combined with a deeper CNN architecture (VGG-19).

It is also informative to contrast training performance with subject-wise test results. Certain configurations, such as M2 and M4, achieved perfect or near-perfect average training accuracies of 100% and 99.42%, respectively. However, their corresponding test accuracies dropped to 94.30% and 92.88%. This gap highlights the presence of mild overfitting and demonstrates that strong performance on training data alone is not sufficient to guarantee robust generalisation across subjects. These observations reinforce the importance of subject-wise evaluation in radar-based large-scale HAM, where inter-subject variability in motion patterns can significantly influence classification outcomes.

Overall, the results confirm that all proposed MRPs exhibit reliable generalisation behaviour, with no configuration showing a severe degradation in performance on unseen subjects. Consequently, the selection of an optimal MRP should not be based solely on peak accuracy, but should instead consider a balanced trade-off between generalisation performance and computational efficiency, which is examined in the following subsections.

3.5.3 Computational Efficiency and Lightweight Model Selection

Computational efficiency is a critical requirement for real-time radar-based large-scale HAM, particularly when targeting deployment on resource-constrained edge devices. In such settings, models must deliver accurate predictions while maintaining low inference latency and modest computational overhead. Excessive processing time can undermine system responsiveness and limit practical usability, especially for time-sensitive applications such as fall detection or continuous activity monitoring.

In this study, computational efficiency is evaluated using two primary time-based metrics, namely training time and inference time, as defined in Section 3.4.5. The results are based on the average of the 10 samples for inference, with the first sample discarded because it includes time for initialising the code and model. Training time is measured per epoch and is averaged across the five runs for each MRP. These metrics provide complementary insights into both the offline cost associated with model optimisation and the online cost incurred during real-time inference. The comparative results for all twelve model–representation pairs are summarised in Table 3.6.

Among the evaluated configurations, pairs M4, M8, and M12 were identified as the most computationally efficient based on their inference and training times. This criterion is particularly important for edge-based systems, where rapid activity prediction is required under strict latency constraints. All three configurations employ the MobileNetV2 architecture, confirming the suitability of lightweight convolutional models for radar-based activity classification when computational efficiency is prioritised.

Inference time is defined as the elapsed time between initiating a prediction request and receiving the corresponding classification output from the trained model. This metric directly reflects the monitoring system's responsiveness and is therefore a key determinant of its applicability in real-world deployments. For applications involving continuous monitoring or safety-

Table 3.6: Training time per epoch and inference time per sample for the twelve model–representation pairs.

MRP	Training time/epoch (s)	Inference time/sample (ms)
M1	3.40	7.16
M2	3.77	8.11
M3	2.77	3.80
M4	**1.79**	**2.78**
M5	3.38	7.10
M6	4.38	6.90
M7	2.74	3.54
M8	**1.49**	**2.57**
M9	3.50	7.02
M10	3.76	6.88
M11	2.73	3.99
M12	**1.34**	**2.76**

critical events, even small inference delays can negatively impact system effectiveness. Training time is measured per epoch and is important for retraining when the environment changes or new data emerges.

The results demonstrate that lightweight architectures can achieve inference times of a few milliseconds per sample while maintaining competitive recognition performance. This highlights the importance of jointly considering model architecture and radar preprocessing strategy when designing practical large-scale HAM systems. Rather than maximising recognition accuracy alone, an effective deployment-oriented solution must balance performance with computational efficiency, ensuring that the system remains responsive, scalable, and suitable for execution on edge hardware.

3.5.4 Computational Cost Across Radar Representations

What follows is an examination of the computational cost associated with each radar preprocessing representation, accounting for both preprocessing overhead and subsequent model training and inference performance. By jointly examining these factors, the practical suitability of each radar representation for real-time large-scale HAM can be evaluated.

For the TR map, the preprocessing stage is computationally inexpensive. The transformation of raw radar data into TR maps, which represent range variation over time as shown in Figure 3.2, requires only 0.035 s per sample. This makes TR the most efficient representation in terms of preprocessing latency, as further illustrated in Figure 3.8. However, despite this advantage, TR-based representations provide limited discriminative information, as they do not explicitly encode Doppler dynamics. When used as input to CNNs, this limitation results in comparatively lower recognition accuracy and reduced feature separability.

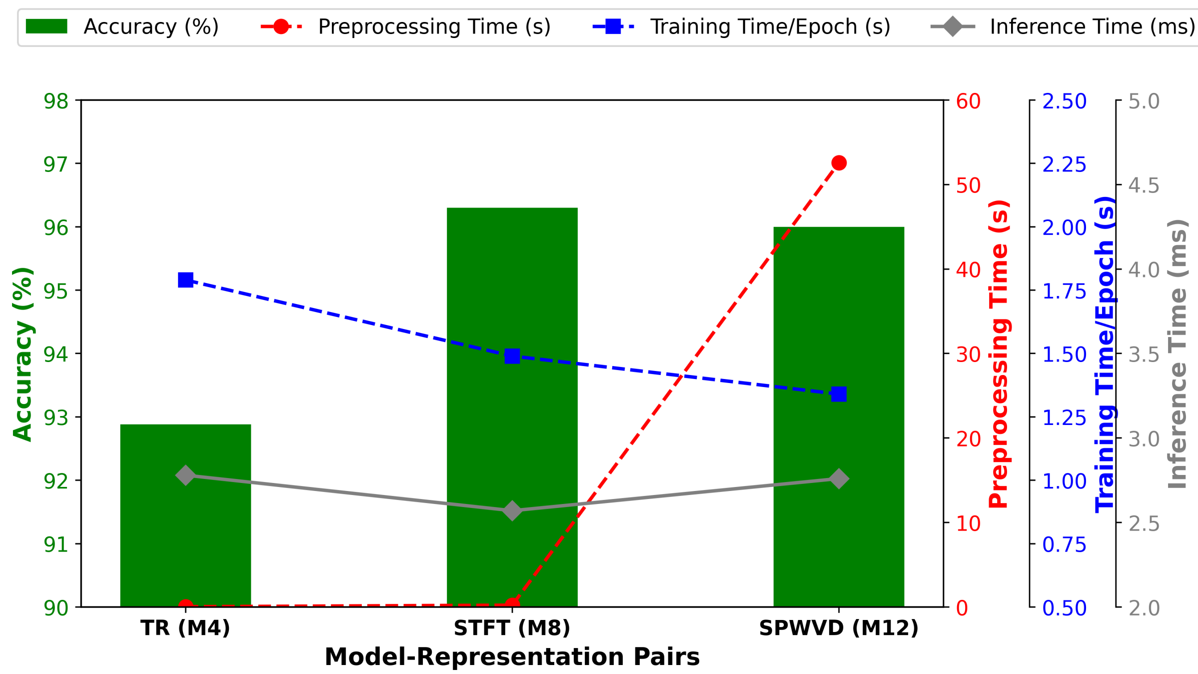


Figure 3.8: Performance and computational cost across radar representations (TR, STFT, and SPWVD) as inputs to MobileNetV2, showing representative maps alongside preprocessing time, training time per epoch, inference time, and test accuracy.

Among all model–representation pairs operating on TR maps, MobileNetV2 (M4) achieves the best balance between efficiency and performance, with a training time of 1.79 s per epoch, an inference time of 2.78 ms per sample, and a recognition accuracy of 92.88%, as shown in Figure 3.8. In contrast, deeper architectures such as VGG-16, VGG-19, and ResNet-50 achieve higher recognition accuracy when paired with TR maps, but at the expense of significantly increased training and inference times. This highlights the limited scalability of TR representations when combined with computationally intensive models.

The STFT-based TF representation captures temporal frequency variations associated with human motion, as illustrated in Figure 3.3. The preprocessing time required to generate an STFT map is modest, approximately 0.22 s per sample, making it substantially faster than higher-resolution time–frequency techniques. When used as input to convolutional networks, STFT-based representations provide a favourable trade-off between recognition performance and computational efficiency.

For example, MobileNetV2 paired with STFT (M8) achieves a training time of 1.49 s per epoch and an inference time of 2.57 ms per sample, while maintaining a high test accuracy of 96.30%. Although VGG-16 and ResNet-50 paired with STFT achieve slightly higher recognition accuracies of 96.87% and 97.15%, respectively, these gains are accompanied by increased computational cost, as evidenced by the longer training and inference times reported in Tables 3.5 and 3.6. VGG-16, in particular, exhibits an inference time of 7.10 ms per sample, rendering it less suitable for latency-sensitive systems.

The SPWVD map provides the highest time–frequency resolution by jointly exploiting independent time and frequency smoothing windows, as detailed in Section 3.3. However, this improvement comes at a substantial preprocessing cost. Generating a single SPWVD map requires 52.58 s, which is several orders of magnitude higher than both TR and STFT preprocessing. Despite this limitation, SPWVD-based representations enable strong classification performance once the map images are generated.

When paired with MobileNetV2 (M12), SPWVD achieves a training time of 1.34 s per epoch, an inference time of 2.76 ms per sample, and a recognition accuracy of 96.01%, as shown in Figure 3.8. These results demonstrate that the high-resolution information encoded by SPWVD maps can be moderately exploited by lightweight architectures compared to STFT, which achieved a slightly higher accuracy of 96.30%. Such a complex preprocessing, however, achieved the highest accuracy of 98.01% when paired with VGG-19. Nevertheless, the excessive preprocessing latency renders SPWVD unsuitable for real-time systems that require rapid end-to-end response, particularly for safety-critical scenarios such as fall detection in elderly care environments. Overall, this analysis highlights a clear trade-off between representational richness and computational feasibility. While TR incurs minimal preprocessing cost, its limited motion discrimination limits recognition performance, as shown in its best confusion matrix in Figure 3.6. SPWVD provides superior time–frequency resolution but incurs prohibitive preprocessing latency. In contrast, STFT-based representations strike a practical balance, delivering strong recognition performance with manageable preprocessing and inference costs, making them well-suited for real-time large-scale HAM on edge platforms.

3.5.5 Comprehensive Evaluation of Model–Representation Pairs

Evaluating model–representation pairs across various constraints reveals which combinations are practically suitable for large-scale HAM. Each radar representation exhibits distinct characteristics that influence both recognition accuracy and computational efficiency, thereby determining its applicability to specific deployment scenarios.

TR representations are the least demanding in terms of preprocessing, requiring only 0.035 s per sample to generate range–time maps following a single FFT operation with a Hamming window and MTI filtering. These representations encode the temporal evolution of target range and retain all FFT-derived range bins, including bins dominated by noise and environmental clutter. While this comprehensive inclusion preserves global motion information, it also introduces redundant and non-discriminative features. In addition, TR representations do not explicitly encode Doppler information, which limits their ability to characterise motion dynamics associated with complex human activities.

Among the evaluated models using TR inputs, VGG-16 achieved the highest recognition accuracy of 95.73% by effectively learning robust features in the presence of noise, as reported in Table 3.5. However, this performance gain is accompanied by increased training and inference times, as shown in Table 3.6. In contrast, MobileNetV2 paired with TR achieved significantly lower computational cost, with faster training and inference times, albeit with reduced recognition accuracy of 92.88%. This combination is therefore better suited to applications with relaxed accuracy requirements and strict latency constraints.

Doppler-aware representations, namely STFT and SPWVD, provide richer motion representations by capturing frequency variations induced by human movement. These representations focus on range bins where target activity is present, as identified from the TR maps, thereby isolating target-specific information and reducing background interference. This selective processing enhances feature discriminability and improves classification performance across all evaluated models.

SPWVD-based representations paired with VGG-19 achieved the highest overall recognition accuracy of 98.01%. This combination is well-suited to scenarios where classification accuracy is prioritised over computational efficiency, such as offline analysis or post-event behavioural assessment. However, the substantial preprocessing overhead associated with SPWVD, requiring 52.58 s per one radar map generation, limits its feasibility for real-time systems that demand rapid end-to-end response.

In contrast, STFT-based representations offer a more balanced trade-off between accuracy and computational efficiency. When paired with MobileNetV2, the STFT map achieved 96.30% recognition accuracy, including perfect classification of fall events, while maintaining the lowest inference time of 2.57 ms per sample. The preprocessing time of 0.22 s further supports its suitability for latency-sensitive applications. This combination, therefore, represents a practical and efficient solution for real-time large-scale HAM on resource-constrained edge platforms. Moreover, the STFT–MobileNetV2 pair (M8) in Figure 3.9 displays a confusion matrix that provides effective classification across all classes and achieves 100% in fall detection, making its selection well justified.

Overall, the comparative evaluation demonstrates that the choice of radar representation and learning architecture should be matched to the intended deployment objective rather than selected on accuracy alone. Pair M8, which combines STFT representations with MobileNetV2, is recommended as the balanced default configuration for the use cases targeted in this thesis because it achieves high recognition accuracy while maintaining the lowest inference latency among the evaluated pairs. This does not imply that M8 is universally optimal. If the priority is sharper TF localisation for offline analysis, post-event review, or applications where computational cost is less restrictive, the SPWVD–VGG-19 configuration, denoted as M10, may be preferred despite its higher preprocessing cost. Conversely, if minimal preprocessing cost is the overriding constraint, TR-based representations may be considered, although they provide less explicit velocity information and may be less discriminative for activities with similar range evolution. Therefore, real-time embedded monitoring favours STFT–MobileNetV2, offline interpretability may justify SPWVD-based configurations, and minimal preprocessing may

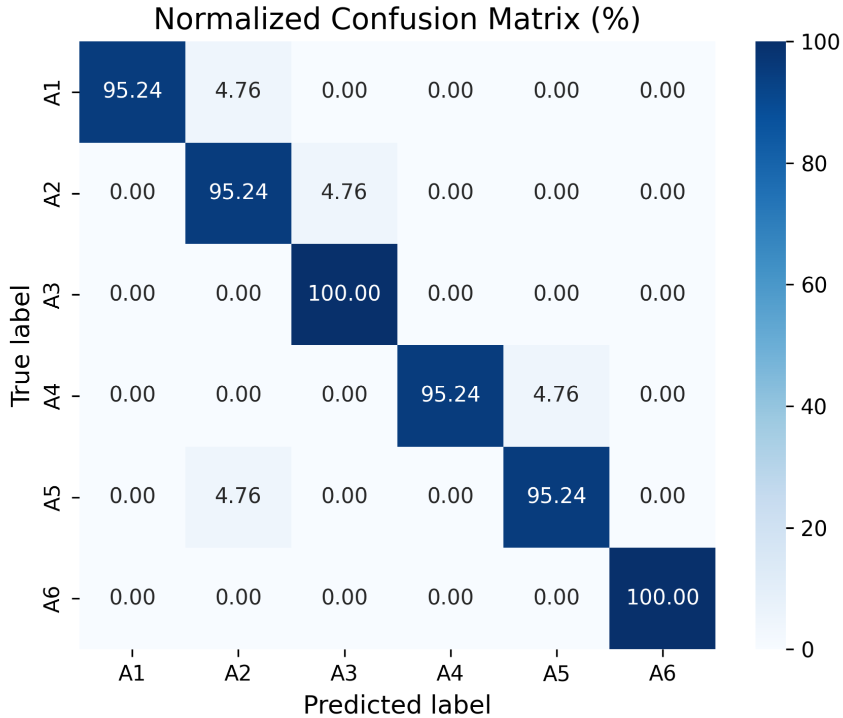


Figure 3.9: Confusion matrix of M8 (STFT–MobileNetV2), demonstrating consistent classification performance across all activities and perfect detection of fall events.

favour TR-based representations. The pairwise results reported in Tables 3.5 and 3.6 enable the representation–model pair to be selected according to these application-specific trade-offs.

3.5.6 Comparison of Pair M8 with State-of-the-Art Models

A comparative evaluation against representative state-of-the-art architectures is summarised in Table 3.7. To ensure a fair comparison, all methods are re-implemented and evaluated under the same hardware and software configuration using STFT-based map inputs resized to 224×224.

The CNN model trained from scratch in [20] achieved a recognition accuracy of 95.44%, but required an inference time of 5.14 ms per sample, nearly twice that of the proposed transfer-learning-based MobileNetV2 model. The hybrid CNN+LSTM architecture reported in [139] achieved the shortest training time per epoch at 1.12 s; however, its recognition accuracy was limited to 84.90%, and the inference latency reached 6.04 ms per sample. The Bi-LSTM model presented in [140] demonstrated competitive accuracy of 95.16% with an inference time of 2.77 ms per sample, yet remained inferior to MobileNetV2 in both training efficiency and run-time latency.

The proposed MobileNetV2 model, corresponding to Pair M8, achieved the highest recognition accuracy of 96.30% while maintaining the lowest inference time of 2.57 ms per sample. This result highlights the effectiveness of combining STFT-based radar representations with a lightweight transfer learning architecture. Compared to models trained from scratch or

Table 3.7: Comparison of recognition accuracy and computational time between the proposed MobileNetV2-based model and representative state-of-the-art approaches using STFT radar representations.

Reference	Model	Training Time/epoch (s)	Inference Time/sample (ms)	Accuracy (%)
[139]	CNN-LSTM	1.12	6.04	84.90
[140]	Bi-LSTM	1.70	2.77	95.16
[20]	CNN (from scratch)	1.88	5.14	95.44
This work	**Pair M8 (STFT + MobileNetV2)**	**1.49**	**2.57**	**96.30**

recurrent-based approaches, the proposed solution offers a superior balance between accuracy and computational efficiency, making it well-suited for real-time large-scale HAM on resource-constrained platforms.

3.6 Summary

This chapter addressed radar-based large-scale HAM through a systematic investigation of radar preprocessing strategies and deep learning architectures under realistic data and computational constraints. In direct response to **C1.1**, which concerns the reliable sensing and interpretation of large-scale human motion such as walking, sitting, standing, bending, and falling, an end-to-end framework was developed that transforms raw FMCW radar signals into discriminative two-dimensional representations suitable for automated classification.

Three radar preprocessing representations were examined, namely TR, STFT, and SPWVD, and were paired with four convolutional neural network architectures, VGG-16, VGG-19, ResNet-50, and MobileNetV2. This resulted in twelve distinct model–representation pairs that were evaluated using a subject-wise experimental protocol to ensure generalisation across unseen individuals. This evaluation strategy strengthens the relevance of the findings for real-world deployment, where inter-subject variability is a dominant challenge.

The comparative analysis demonstrated that the choice of radar representation significantly influences both recognition performance and computational efficiency. TR representations exhibited minimal preprocessing cost but lacked Doppler information, limiting their discriminative power for complex activities. SPWVD representations achieved the highest classification accuracy due to their high time–frequency resolution; however, their substantial preprocessing overhead rendered them unsuitable for latency-sensitive applications. In contrast, STFT representations provided a balanced trade-off between representational richness and computational cost.

Within this design space, the MobileNetV2 and STFT pairing (Pair M8) emerged as the most balanced configuration. Pair M8 achieved high recognition accuracy while maintaining the lowest inference latency among all evaluated model–representation pairs. This directly addresses **C2**, which emphasises the need for data- and computationally efficient learning models capable of operating in real time and under resource-constrained conditions. The results confirm that lightweight transfer learning architectures, when combined with appropriately selected radar representations, can deliver robust large-scale HAM without incurring high computational cost.

Overall, this chapter provided a structured and reproducible evaluation of radar preprocessing and deep learning design choices for large-scale HAM. By explicitly quantifying the trade-offs between accuracy, generalisation, and computational efficiency, it offered practical guidance for selecting model–representation combinations that satisfy both sensing and deployment requirements. The outcomes of this chapter establish a solid foundation for the subsequent investigation of finer-grained motion analysis and multimodal human monitoring in the following chapters.

Chapter 4

Radar-Based Fine-Grained and Vital Signs Monitoring

This chapter presents the second contribution of the thesis and addresses challenges **C1.2**, **C1.3**, and **C2**, as outlined in Section 1.3. Following the large-scale activity monitoring investigated in Chapter 3, the focus shifts to more challenging radar-sensing scenarios in which target motions are smaller, the signal-to-noise ratio is lower, and labelled radar data are more limited. These conditions introduce additional challenges for both signal processing and machine learning, requiring methods that remain robust under weak radar signatures and constrained data availability.

To investigate these challenges, the chapter considers two sensing objectives operating at fine spatial and temporal scales: fine-grained HAM and physiological-scale estimation of vital signs. Although these problems differ in application, both involve small-amplitude radar returns, strong sensitivity to motion artefacts, and limited training data, and therefore represent a common class of radar sensing problems under data scarcity. The first study investigates fine-grained upper-body and head activity monitoring using mmWave radar (Section 4.3), where subtle movements with limited spatial displacement pose a challenge for conventional radar-based activity recognition pipelines. The second study focuses on radar-based estimation of heart rate (HR) and respiration rate (RR) (Section 4.4) and addresses the degradation in estimation accuracy caused by random body movement. Together, these studies demonstrate how radar can support behavioural and physiological monitoring at small spatial and temporal scales under realistic sensing constraints, and their relationship is discussed in Section 4.5.

4.1 Introduction

Radar-based human monitoring introduced in this chapter addresses sensing challenges that extend beyond those encountered in large-scale activity recognition. In this context, two related sensing problems are considered. The first concerns fine-grained HAM (**C1.2**), where subtle upper-body and head movements must be distinguished despite weak radar signatures and limited inter-class separability. The second concerns the physiological-scale of vital signs estimation (**C1.3**), where micro-motions associated with respiration and cardiac activity must be reliably extracted in the presence of body movement. Both problems are characterised by small-amplitude motion, strong susceptibility to interference, and constrained data availability, and therefore require sensing, signal processing, and learning strategies that differ from those used for large-scale activities. Together, these studies examine radar sensing under data scarcity and during body movements, addressing challenges **C1.2**, **C1.3**, and **C2**.

While large-scale human activities such as walking, sitting, or falling generate strong and easily separable radar signatures, fine-grained movements and physiological signals are fundamentally more challenging to interpret. Upper-body gestures and head movements produce low-energy micro-Doppler signatures that often overlap in the time–frequency domain, while respiration- and cardiac-induced chest motions are easily masked by voluntary or involuntary body movement. These characteristics directly reflect the challenges defined in **C1.2** and **C1.3**, where both behavioural interpretation and physiological estimation must be performed under realistic conditions.

Radar sensing, particularly using mmWave FMCW platforms, offers a unified means of addressing these challenges. Owing to its sensitivity to small displacements, radar can capture both behavioural motion and physiological micro-motions without physical contact, while preserving privacy and operating independently of lighting conditions. However, exploiting this capability in practice requires carefully designed signal processing pipelines and data-efficient learning strategies that explicitly account for motion interference and limited labelled data, in line with the efficiency constraints highlighted in **C2**.

Recent research has increasingly explored fine-grained radar-based activity monitoring in applications such as driver monitoring and gesture recognition, where subtle behavioural cues carry meaningful information about human state and intent [84,152,155,225]. In parallel, radar-based vital signs estimation has emerged as a non-contact alternative to wearable sensing in healthcare and paediatric monitoring scenarios [238,239]. Despite promising results under controlled conditions, both research directions consistently report performance degradation in the presence of body movement, highlighting motion interference and data scarcity as persistent barriers to reliable deployment.

Accordingly, this chapter investigates fine-grained behavioural monitoring and physiological-scale vital signs estimation as two closely related yet distinct sensing problems. Although they target different outputs, both are governed by common constraints, including weak signal characteristics, overlap between motion components, and limited training data. By addressing these problems within a single chapter, the work establishes a coherent progression from behavioural monitoring to physiological sensing at fine spatial and temporal scales, providing a foundation for more comprehensive human-centric monitoring frameworks.

4.2 Chapter Contributions

This chapter presents two complementary contributions that advance fine-grained radar-based monitoring under data scarcity and motion interference and fill the gaps identified in Sections 2.3 and Subsection 2.4.3. The main contributions are summarised as follows:

- **Fine-grained upper-body and head movement monitoring:** A non-contact mmWave FMCW radar-based framework is developed to recognise subtle seated upper-body and head movements representative of constrained monitoring scenarios. A dedicated dataset comprising six fine-grained activities is collected using a single radar sensor across multiple participants and two indoor environments, introducing inter-subject variability and a controlled cross-room shift.

- **Signal processing and representation for subtle motion sensing:** A tailored radar processing pipeline is designed to transform raw radar measurements into discriminative time–frequency representations suitable for fine-grained motion analysis.

- **Data-efficient learning via augmentation and transfer learning:** To mitigate limited labelled data, a targeted augmentation strategy inspired by SpecAugment is adapted to radar TF maps and combined with transfer learning using a lightweight model.

- **Movement-compensated radar-based HR and RR estimation:** A motion-robust framework is introduced for non-contact HR and RR estimation using mmWave FMCW radar. The approach targets movement interference through signal-level separation of physiological components from motion artefacts using Variational Mode Decomposition (VMD).

- **Performance enhancement over conventional processing:** The proposed VMD-based motion artefact suppression is systematically evaluated against conventional radar-based signal processing approaches.

- **Unified fine-grained monitoring across behavioural and physiological targets:** By jointly addressing fine-grained movement recognition and motion-robust vital signs estimation, this chapter establishes a coherent progression from behavioural monitoring to physiological sensing under shared constraints, including weak signatures, motion interference, and data limitations.

4.3 Fine-Grained Human Activity Monitoring

This section addresses fine-grained HAM using radar, with a specific focus on subtle upper-body and head movements performed in a seated configuration. In contrast to the large-scale activities investigated in Chapter 3, such movements are characterised by small spatial displacements, low radial velocities, and limited inter-class separability, making them significantly more challenging to sense and analyse reliably.

From a sensing perspective, fine-grained activities often produce weak micro-Doppler signatures that are easily obscured by noise, clutter, and involuntary body motion. As discussed in Section 2.2.2, the use of mmWave FMCW radar offers a distinct advantage in this context due to its shorter wavelength, which increases sensitivity to small-amplitude motions and improves Doppler separability for low-velocity movements. This property makes mmWave radar particularly well suited for capturing fine-grained behavioural cues that are poorly resolved at lower carrier frequencies, which lack the range resolution to distinguish different body parts.

In this work, fine-grained HAM is realised through the classification of six representative upper-body and head movements that reflect natural human behaviour in constrained environments, such as vehicle cabins or seated monitoring scenarios. The proposed framework integrates a dedicated data collection protocol, a tailored signal processing pipeline for TF map extraction, and data-efficient learning strategies designed to operate effectively under limited training data. Consistent with the findings of Chapter 3, STFT-based TF representations and a MobileNetV2 backbone are adopted as the baseline configuration, providing a favourable balance between representational capacity and computational efficiency.

4.3.1 Data Collection Protocol

This subsection describes the protocol adopted for collecting radar data for fine-grained HAM, with an emphasis on consistency, repeatability, and subject-level generalisation. The protocol was designed to capture subtle upper-body and head movements in a seated configuration under controlled yet realistic conditions, while limiting confounding factors such as large body translations and uncontrolled environmental motion.

An overview of the complete data acquisition and processing workflow is illustrated in Figure 4.1. The system comprises a single mmWave FMCW radar sensor, positioned approximately 90 cm in front of the participant, a guided activity execution interface, and a multi-stage signal processing and learning pipeline. This setup reflects practical monitoring scenarios, such as vehicle cabins or seated indoor environments, where fine-grained behavioural cues must be captured unobtrusively using a minimal sensing configuration.

Data were collected from 15 male adult participants aged between 24 and 38 years, with heights ranging from 165 to 186 cm and no reported mobility impairments. The recordings were conducted across two indoor environments to introduce environmental variability and reduce overfitting to a single room configuration. In total, approximately 4,400 seconds of raw radar recordings were acquired, from which 264 labelled samples were extracted. These samples were evenly distributed across six fine-grained activity classes, yielding 44 samples per class.

The selected activities represent behaviourally meaningful seated upper-body and head move-

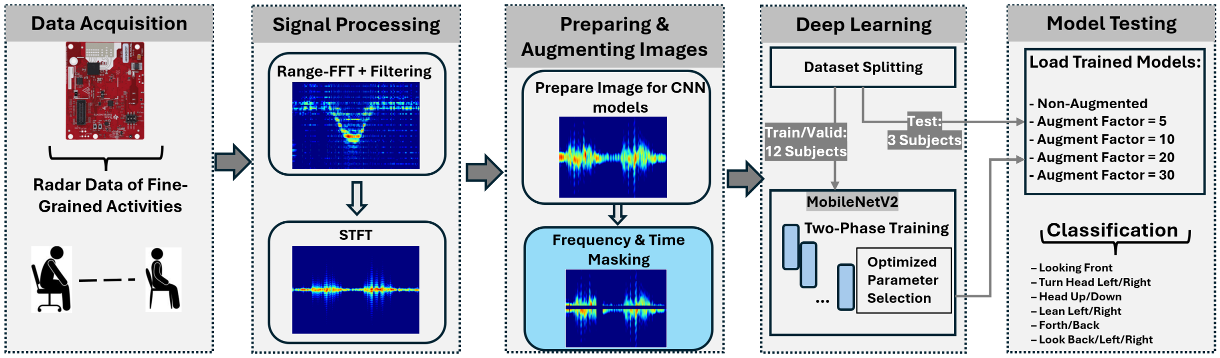


Figure 4.1: System model of the proposed fine-grained activity monitoring pipeline, from radar acquisition through signal processing to activity classification.

ments relevant to constrained monitoring scenarios, including in-cabin driver-state assessment and AAL. Looking front provides a stationary baseline; head turning and nodding capture rotational head motion; lateral leaning and forward/backward movement capture upper-body translation; and looking back over the shoulder represents a larger combined head–torso movement associated with attention shifts. The activity set, therefore, includes both rotational and translational components, as well as kinematically similar movements that are deliberately challenging to distinguish using radar TF representation maps. This design allows the experiment to assess whether mmWave radar can discriminate subtle seated movement differences that may be linked to distraction, attention shifts, or posture changes. The activity classes and their descriptions are summarised in Table 4.1.

Table 4.1: Fine-grained upper-body and head movement classes and descriptions.

ID	Activity	Description
1	Looking Front	Remain seated and still while facing forward towards the radar sensor.
2	Turn Head	Rotate the head left and right before returning to the centre-facing position.
3	Nod Head	Move the head up and down before returning to the neutral position.
4	Lean Body	Lean the upper body laterally to the left and right before returning upright.
5	Move Forwards/Backwards	Translate the upper body forwards and backwards relative to the radar before returning to the neutral position.
6	Look Back	Rotate the head and upper torso to look over the left and right shoulder before facing forward again.

To promote subject-level generalisation and reduce the risk of memorisation, a subject-based data partitioning strategy was adopted. Ten subjects were used for training, producing 180 samples, while two subjects were reserved for validation, yielding 36 samples. The remaining three subjects were held out exclusively for testing, resulting in 48 unseen samples. In addition to subject-wise splitting, a controlled cross-environment protocol was introduced. Recordings for training subjects were collected in one room, while validation and test recordings were obtained in a different room. This design introduces a realistic cross-room domain shift, reflecting deployment scenarios in which both the monitored individual and the surrounding environment may differ from training conditions.

All recording sessions followed a structured procedure. During each session, the participant was seated on a chair facing the radar sensor at a fixed distance. A visual display positioned behind the radar guided the participant through a predefined sequence of activities, ensuring consistent execution across subjects while preserving natural motion patterns. A short familiarisation phase was conducted prior to data recording to reduce hesitation and inter-trial variability. Each activity was repeated multiple times, with rest intervals introduced between trials to minimise fatigue-related effects.

Each participant performed a limited number of repetitions per activity, prioritising subject diversity over repeated trials. This design choice contrasts with many existing studies that rely on extensive repetitions from a small number of subjects, which can artificially inflate performance through intra-subject memorisation. By emphasising inter-subject variability under constrained repetition, the dataset presents a more challenging and realistic generalisation scenario. This setting directly motivates the use of data-efficient learning strategies, including targeted data augmentation and transfer learning, to mitigate data scarcity (**C2**) while preserving subject-level generalisation.

4.3.2 Data Acquisition and Experimental Setup

Radar data were acquired using a Texas Instruments AWR1843BOOST [117] mmWave FMCW radar operating in the 77 GHz band. The choice of a mmWave platform is motivated by its enhanced sensitivity to small-amplitude motions, which is critical for fine-grained monitoring of upper-body and head activity. As discussed in Section 2.2.2, the shorter wavelength at mmWave frequencies improves Doppler separability for low-velocity movements and enables distinguishing a narrower range of body parts compared to lower-frequency FMCW radars.

The radar was configured to transmit linear FMCW chirps with a starting frequency of 77 GHz and a chirp slope of 50 MHz/μs, resulting in a total bandwidth of 3 GHz, spanning the 77–80 GHz sub-band. Each chirp had a duration of 60 μs, providing high range resolution suitable for isolating subtle motion components from background clutter. The intermediate-frequency signal was sampled at 3×10^6 samples/s, yielding $N_{\mathrm{ADC}} = 128$ samples per chirp. A pulse repetition frequency of 2,000 Hz was selected to ensure sufficient temporal resolution for capturing rapid micro-Doppler variations associated with head and upper-body movements.

The transmitted equivalent isotropically radiated power (EIRP) of the radar complies with the applicable UK (Ofcom) and ETSI short-range-device limits for operation in the 77–81 GHz band [240, 241].

During data acquisition, the radar sensor was positioned directly in front of the seated participant, aligned with the frontal body axis, and mounted at approximately chest height. The system operated in a single-input single-output configuration, reflecting a low-complexity and deployment-oriented sensing setup appropriate for constrained environments such as vehicle cabins or indoor monitoring spaces.

All radar data were recorded as raw intermediate-frequency samples for offline processing. This design choice enables full control over subsequent signal processing stages, including clutter suppression, micro-Doppler extraction, and time–frequency analysis. Computational processing and model training were performed on the same GPU-accelerated workstation used in Chapter 3, ensuring consistency across experimental evaluations. Implementation details and runtime performance are briefly revisited in the results section for completeness. Representative examples of the experimental setup are shown in Figure 4.2. The radar sensor was mounted at approximately chest height relative to the seated participant and aligned with the frontal body axis. This placement ensured that upper-body and head movements dominated the radar field of view, while interference from lower-body motion was minimised, thereby supporting consistent extraction of fine-grained radar TF representation maps across sessions.

Figure 4.2: Experimental setup for fine-grained motion capture, showing participant seating, mmWave radar placement, and the guidance interface.

4.3.3 Signal Processing Pipeline for Fine-Grained Movements

The signal processing pipeline builds directly upon the formulation introduced in Section 2.2.2, with adaptations tailored to the characteristics of fine-grained upper-body and head movements. After dechirping and analogue-to-digital conversion, the received FMCW radar signal is represented by the complex intermediate-frequency (IF) samples $x[n,m]$, where n denotes the fast-time index within a chirp and m denotes the slow-time index across successive chirps. The discrete measurements are given by

$$x[n,m] = \sum_{p=1}^{P} A_p \exp\left\{ j2\pi \left(f_{\mathrm{b},p} nT_{\mathrm{s}} + f_{\mathrm{D},p} mT_{\mathrm{r}} \right) \right\} + w[n,m], \tag{4.1}$$

which corresponds to the multi-target signal model in (2.22), with each reflector p associated with a moving body segment contributing to the radar return.

Range Processing and Motion Isolation Range processing is performed by applying the FFT along the fast-time dimension of $x[n,m]$ for each chirp, yielding range-resolved measurements $X[k,m]$. To suppress static reflections from the environment and seating structure, clutter removal is performed along the slow-time dimension for each range bin using a fourth-order Butterworth high-pass filter with a cut-off frequency of 0.75 Hz. This choice, in particular for mmWave radars as explained in Section 2.2.2, effectively attenuates near-zero Doppler components while preserving low-velocity micro-Doppler signatures associated with fine-grained upper-body and head movements.

Under the adopted experimental configuration, in which the seated participant is positioned directly in front of the radar and dominates its field of view, the remaining dynamic components primarily reflect voluntary upper-body and head motion. The motion-related radar response is obtained by coherently aggregating the filtered signal across the subject-related range bins,

$$s[m] = \sum_{k \in \mathscr{K}} X[k,m], \tag{4.2}$$

where $\mathscr{K}$ denotes the set of range bins associated with the detected subject. This aggregation yields a single complex slow-time sequence that captures fine-grained velocity variations while reducing residual noise and clutter contributions.

STFT-Based TF Representation Fine-grained upper-body and head movements are inherently non-stationary and exhibit time-varying Doppler characteristics. To characterise these dynamics, the STFT is applied to the slow-time signal $s[m]$, following the formulation presented in Section 2.2.2. Using a sliding analysis window, the STFT produces a TF representation,

$$S(m,\nu) = \sum_{\ell} s[\ell]\, w[\ell - m]\, e^{-j2\pi\nu\ell}, \tag{4.3}$$

where $w[\cdot]$ denotes the analysis window and ν is the Doppler frequency index.

In this work, a Hann window of length $L = 256$ samples is used with an overlap of 200 samples, corresponding to a hop size of $H = 56$ samples. For each windowed segment, a 4096-point FFT is computed to obtain the Doppler spectrum, yielding a high-resolution TF representation suitable for capturing subtle micro-Doppler variations from fine-grained upper-body and head movements.

The magnitude of $S(m,\nu)$ forms a TF representation map that highlights subtle micro-Doppler modulations associated with fine-grained upper-body and head movements. Compared with large-scale activities, these signatures occur at lower Doppler frequencies and exhibit weaker energy concentration, making their extraction more susceptible to noise and interference. At mmWave carrier frequencies, the shorter wavelength enhances Doppler sensitivity for low-velocity motion components, improving separability in this fine-grained regime, as illustrated with examples in Section 2.2.2.

STFT Map Normalisation and Smoothing To improve robustness against noise, background clutter, and inter-subject variability, the resulting TF representations are normalised and subsequently smoothed using a light Gaussian filter with standard deviation $\sigma = 1.0$. This operation attenuates isolated spectral artefacts and small-amplitude fluctuations while preserving the dominant micro-Doppler structures that characterise fine-grained upper-body and head movements. Representative STFT-based TF maps for all six activity classes after normalisation and smoothing are shown in Figure 4.3. These maps indicate that class separability is not uniform across the six movements. Activities involving larger torso displacement produce broader and stronger Doppler structures, whereas nodding and small head turns generate weaker signatures that can overlap with residual body motion and environmental clutter. The fine-grained signatures also exhibit substantially smaller Doppler excursions than the large-scale activities examined in Chapter 3, confirming the greater difficulty of this classification task. Consequently, the main sources of confusion are expected to arise between kinematically similar movements with subtle temporal and spectral differences. At the same time, the residual separability visible in these TF maps indicates that informative motion patterns remain, motivating the use of data augmentation and transfer learning to improve generalisation. These processed radar maps, therefore, constitute the final input to the augmentation and learning stages described in the following subsections.

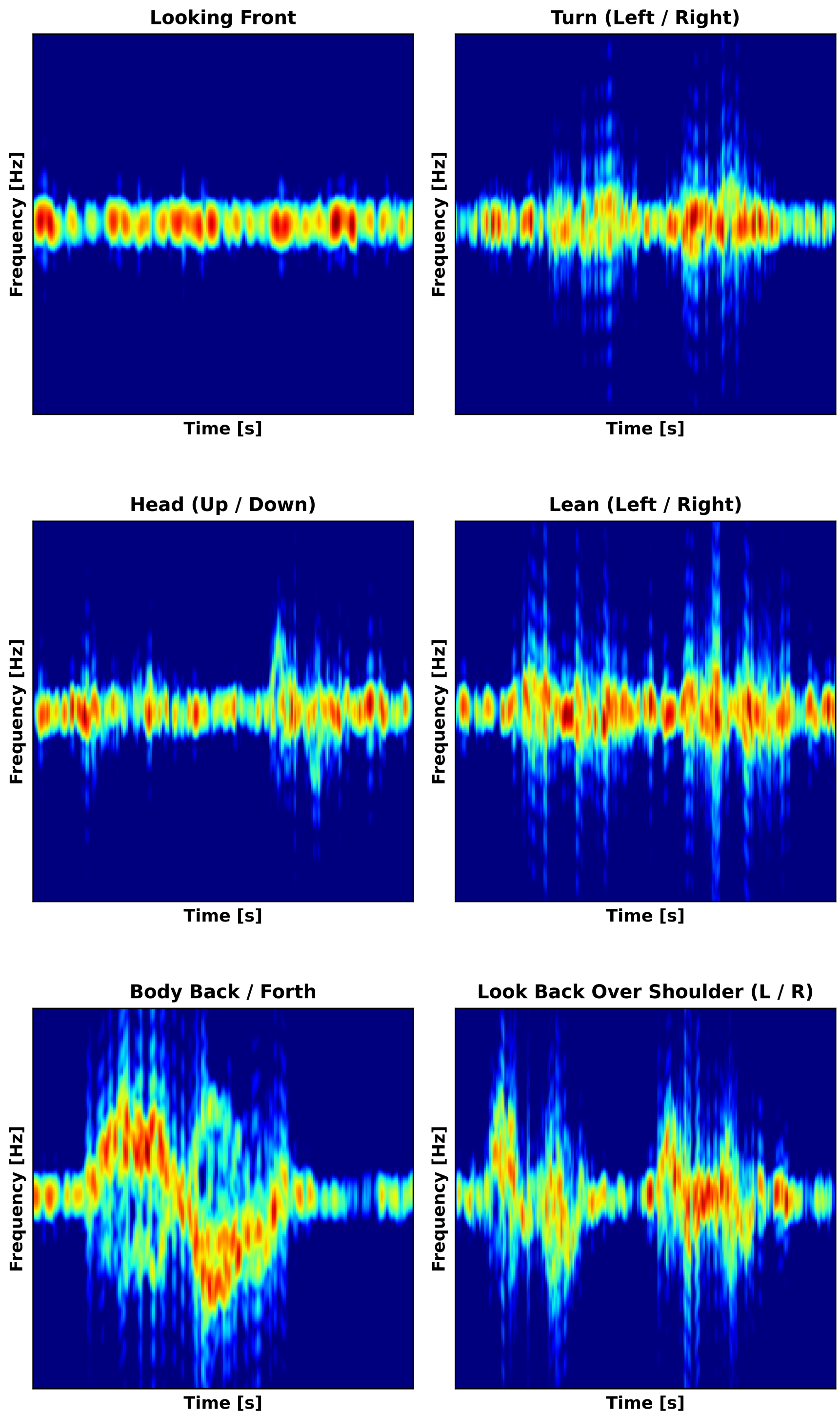


Figure 4.3: Normalised and smoothed STFT-based TF representations corresponding to the six fine-grained upper-body and head movement classes.

4.3.4 Data Augmentation Strategies

A key challenge in radar-based fine-grained HAM is the limited availability of labelled training data. This issue is particularly pronounced for subtle upper-body and head movements, where collecting large datasets is time-consuming and subject-dependent, and where inter-class variability is inherently small. When combined with the high capacity of deep learning models, this data scarcity substantially increases the risk of overfitting and degrades generalisation to unseen subjects and environments.

To address this limitation, a multi-stage data augmentation pipeline was designed to artificially expand the training set while preserving the physical and semantic characteristics of radar micro-Doppler signatures. The adopted strategy is inspired by *SpecAugment*, originally proposed for speech recognition [40], which has demonstrated effectiveness in improving robustness under limited training data by enforcing invariance to partial time–frequency occlusions. In this work, the SpecAugment principle is adapted to radar TF maps and complemented with controlled geometric perturbations to reflect realistic variability in fine-grained human motion.

All augmentation operations are applied exclusively to the training set, while validation and test sets remain unchanged to ensure a fair and unbiased performance evaluation. For each STFT-based map, the following augmentation steps are performed sequentially.

Geometric Rotation To account for slight variations in posture, seating position, and sensor alignment, a random in-plane rotation is applied to each STFT map. The rotation angle is uniformly sampled within $\pm 5^\circ$, which introduces controlled geometric variability without distorting the underlying time–frequency structure of the micro-Doppler signatures. This operation encourages the model to learn rotation-invariant representations while maintaining physical plausibility.

Time and Frequency Masking Following rotation, structured occlusions are applied along the time and frequency axes of the STFT map to improve robustness to partial signal corruption and missing information. Two masking operations are employed:

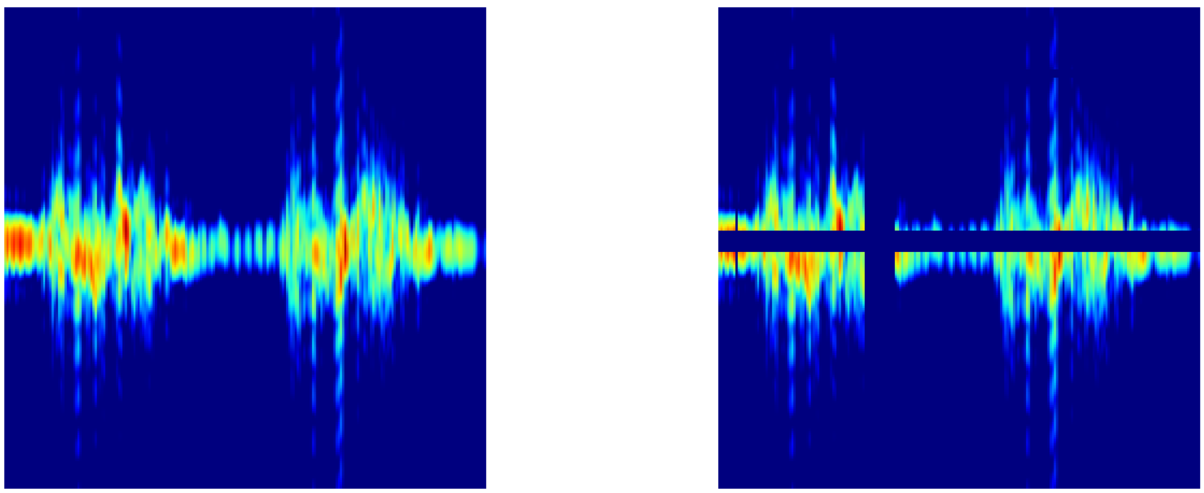
- *Frequency masking:* Two independent frequency masks are applied, each covering a contiguous set of frequency bins with a maximum height of 22 pixels for 224×224 TF maps. This operation simulates variations in Doppler energy distribution caused by changes in motion amplitude or partial self-occlusion.

- *Time masking:* Two independent time masks are applied along the temporal axis, each with a maximum width of 22 pixels. This operation emulates brief interruptions or inconsistencies in motion execution, such as pauses or irregular movement timing.

The locations and extents of all masks are randomly sampled for each augmented instance, ensuring that the network is exposed to diverse perturbation patterns during training.

Augmentation Factor and Dataset Expansion To systematically analyse the effect of training data volume on classification performance, the augmentation pipeline is applied multiple times to each original training data. Augmentation factors of 5, 10, 20, and 30 are considered, where a factor of N indicates that N independently augmented samples are generated from each original STFT map using stochastic rotation and time–frequency masking.

This controlled dataset expansion enables a systematic investigation of the relationship between augmentation strength and model generalisation under data-scarce conditions. An example of the applied augmentation process is illustrated in Figure 4.4, which contrasts an original STFT-based map with its augmented counterpart after time and frequency masking. Importantly, all augmented samples remain physically consistent with plausible fine-grained human motion, ensuring that the learning process is driven by realistic signal variations rather than artificially introduced artefacts.



((a)) Original STFT-based TF map ((b)) Augmented Version

Figure 4.4: Example STFT-based TF map before and after SpecAugment time and frequency masking, illustrating the structured modifications used for data augmentation.

4.3.5 Classification Models for Fine-Grained Human Activity Monitoring

Building on the systematic evaluation presented in Chapter 3, which established the STFT–MobileNetV2 combination as an effective balance between recognition performance and computational efficiency for radar-based human monitoring, this work adopts the same representational and architectural principles for fine-grained activity analysis. Extending this configuration to fine-grained monitoring enables a consistent methodological progression from large-scale to subtle motion sensing within the thesis framework.

Fine-grained HAM introduces additional learning challenges that are not dominant in large-scale activity recognition. Upper-body and head movements produce weak and closely spaced micro-Doppler signatures with limited inter-class separability, increasing susceptibility to noise, subject variability, and overfitting. These characteristics reduce the effectiveness of large, highly parameterised models when training data are limited, and place stronger emphasis on generalisation rather than raw representational capacity [57, 103].

To address these constraints, the classification framework adopted in this work prioritises data efficiency, robustness to subject variability, and computational practicality. Consistent with the learning principles established in Chapter 3, a transfer learning strategy based on a lightweight convolutional neural network is employed. MobileNetV2 is selected as the backbone architecture due to its favourable balance between expressive power and inference efficiency. Its use of depthwise separable convolutions and inverted residual blocks enables effective feature extraction from TF radar representations while maintaining low latency, which is particularly suitable for fine-grained monitoring in embedded and real-time deployment scenarios.

The MobileNetV2 backbone is initialised with weights pre-trained on the ImageNet [39] dataset. Although natural images differ significantly from radar representation maps, early convolutional layers capture generic spatial patterns, such as edges and local textures, which remain beneficial when transferring to TF representations. This reuse of learned representations mitigates the limited size of the fine-grained radar dataset and improves optimisation stability.

A two-phase transfer learning strategy is employed. In the first phase, the pre-trained backbone is frozen and used solely as a feature extractor, while a task-specific classification head is trained. This head consists of a global average pooling layer followed by a fully connected layer with 128 neurons and ReLU activation, a dropout layer for regularisation, and a final softmax layer corresponding to the six fine-grained activity classes. Training only the classification head allows the network to adapt decision boundaries to radar-specific features without perturbing the general representations learned during pre-training.

In the second phase, selected layers of the backbone are unfrozen and fine-tuned using a reduced learning rate. This controlled fine-tuning enables the convolutional filters to adapt to the statistical structure of radar TF representation maps, capturing subtle temporal and frequency variations associated with fine-grained upper-body and head movements. Batch normalisation layers are also updated during this phase to accommodate the distribution shift between natural images and radar-based inputs. Early stopping is applied to prevent overfitting, reflecting the limited size of the available training data. The key hyperparameters used during both phases of the fine-tuning process are summarised in Table 4.2. These parameters were selected to ensure stable optimisation while preserving sensitivity to small-amplitude micro-Doppler variations characteristic of fine-grained upper-body and head movements.

This lightweight transfer learning framework ensures that the classification model remains robust under data scarcity while retaining sufficient representational capacity to resolve fine-grained motion patterns. By aligning model complexity with sensing characteristics and deployment constraints, the adopted approach supports accurate and efficient fine-grained HAM using mmWave radar.

Table 4.2: Optimisation and two-stage fine-tuning settings for the MobileNetV2-based fine-grained activity classifier.

Parameter	**Phase 1 (Classifier-head warm-up)**	**Phase 2 (Full-network fine-tuning)**
Trainable layers	Classification Head Only	Entire Network
Optimiser	AdamW	AdamW
Initial Learning Rate	3×10^{-4}	1×10^{-4}
LR Schedule	Reduce on Plateau	Reduce on Plateau
LR Factor	0.3	0.2
Patience	3 epochs	2 epochs
Max Epochs	8	25
Early Stopping Patience	6 epochs	10 epochs

4.3.6 Experimental Results and Performance Evaluation

This subsection evaluates the performance of the proposed fine-grained HAM framework under subject-wise and environment-wise generalisation conditions. All experiments were conducted on the same GPU-accelerated workstation used in Chapter 3, ensuring consistency across large-scale and fine-grained evaluations. Model training and inference were implemented in Python using standard scientific computing and deep learning libraries, as detailed previously.

Dataset Composition and Evaluation Protocol Following the protocol described in Section 4.3.1, a strict subject-wise data split was employed to assess generalisation performance and prevent subject-specific memorisation. Data from 10 participants were used for training, 2 participants for validation, and 3 participants were reserved for testing.

In addition to subject-wise separation, a controlled cross-environment evaluation was introduced. Training data were collected exclusively in one indoor environment (Room 1), while validation and test data were collected in a different environment (Room 2). This setup introduces a mild domain shift arising from differences in room geometry and background reflections, enabling a more realistic assessment of model performance under unseen environmental conditions.

Impact of Data Augmentation To quantify the effect of data augmentation on classification performance, experiments were conducted using different augmentation factors, as described in Section 4.3.4. The resulting classification accuracy as a function of augmentation strength is shown in Figure 4.5.

When trained without any augmentation, the transfer learning model achieved a mean classification accuracy of $60.00 \pm 1.57\%$, indicating that transfer learning alone was insufficient to overcome the limited size and fine-grained nature of the dataset. Introducing data augmentation

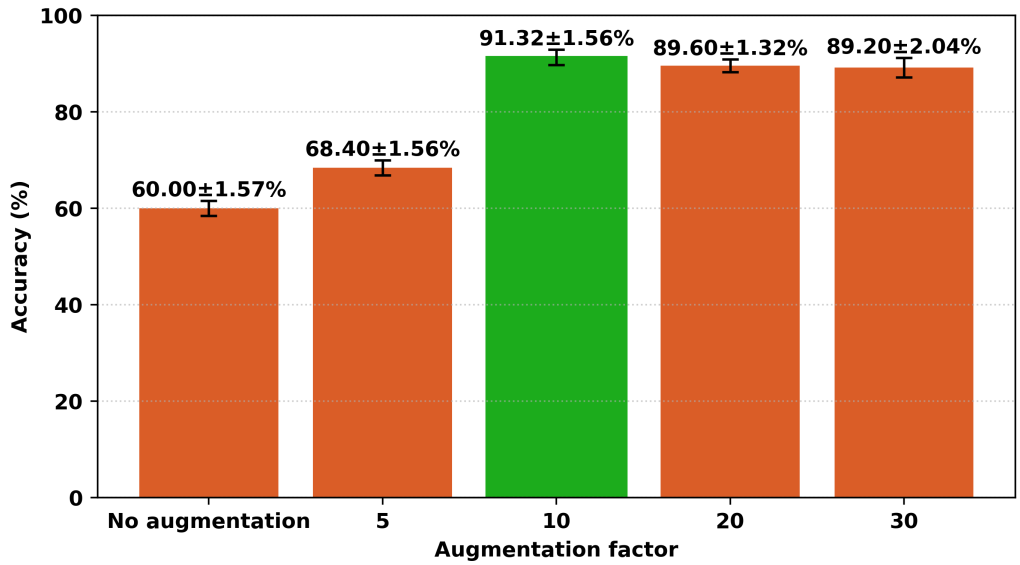


Figure 4.5: Classification accuracy as a function of augmentation factor. Performance improves substantially by a factor of 10, beyond which further augmentation yields diminishing returns.

led to a substantial improvement in performance. An augmentation factor of 5 increased the accuracy to $68.40 \pm 1.56\%$, while a factor of 10 yielded the highest performance of $91.32 \pm 1.56\%$, corresponding to 1,800 training samples.

Further increasing the augmentation factor to 20 and 30 did not provide additional benefits. Instead, a slight degradation in accuracy was observed, with mean accuracies of $89.60 \pm 1.32\%$ and $89.20 \pm 2.04\%$, respectively. This behaviour suggests diminishing returns beyond an augmentation factor of 10, indicating a saturation point in the diversity that can be generated from the available data using SpecAugment-style masking. At higher augmentation levels, the additional samples become increasingly redundant or overly distorted, providing limited new discriminative information and leading to marginal performance degradation. All reported results are averaged over six independent training runs, with the mean and standard deviation reported to reflect robustness.

Per-Class Performance Analysis Detailed per-class precision, recall, and F1-score metrics for the best-performing configuration (augmentation factor of 10) are reported in Table 4.3.

The results indicate strong class-wise performance across most activities, with perfect or near-perfect scores achieved for movements with distinct kinematic signatures, such as *Looking Front* and *Move Forth/Back*. Lower recall values were observed for *Nod Head* and *Turn Head*, reflecting their greater similarity in micro-Doppler structure when viewed from a frontal radar perspective.

Confusion Matrix Analysis The confusion matrix for the baseline model without augmentation is shown in Figure 4.6. Significant confusion is observed between kinematically similar

Table 4.3: Per-class precision, recall, and F1-score for the best-performing fine-grained activity monitoring model (augmentation factor of 10).

Class	Precision	Recall	F1-score
Looking Front	100%	100%	100%
Turn Head	73%	100%	84%
Nod Head	100%	63%	77%
Lean Body	100%	88%	93%
Forth/Back	100%	100%	100%
Look Back	89%	100%	94%

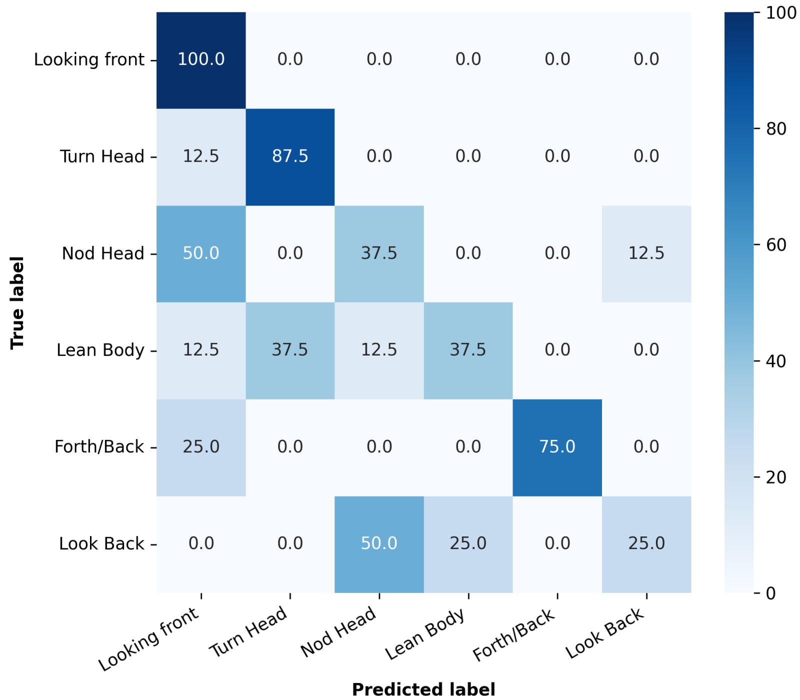


Figure 4.6: Confusion matrix for the fine-grained activity monitoring model without data augmentation, yielding an average accuracy of 60%.

activities. Recall values for *Nod Head*, *Lean Body*, and *Look Back* drop to 37.5%, 37.5%, and 25.0%, respectively, indicating that the model struggles to learn robust decision boundaries from the limited original dataset.

After applying data augmentation with a factor of 10, the confusion matrix becomes strongly diagonal, as shown in Figure 4.7. Recall reaches 100% for four of the six classes, and misclassification rates are substantially reduced across all activities.

To verify that the improvement obtained with data augmentation is statistically significant, McNemar's test was performed by comparing the predictions of the model trained without augmentation and the model trained with an augmentation factor of 10. Because both models were

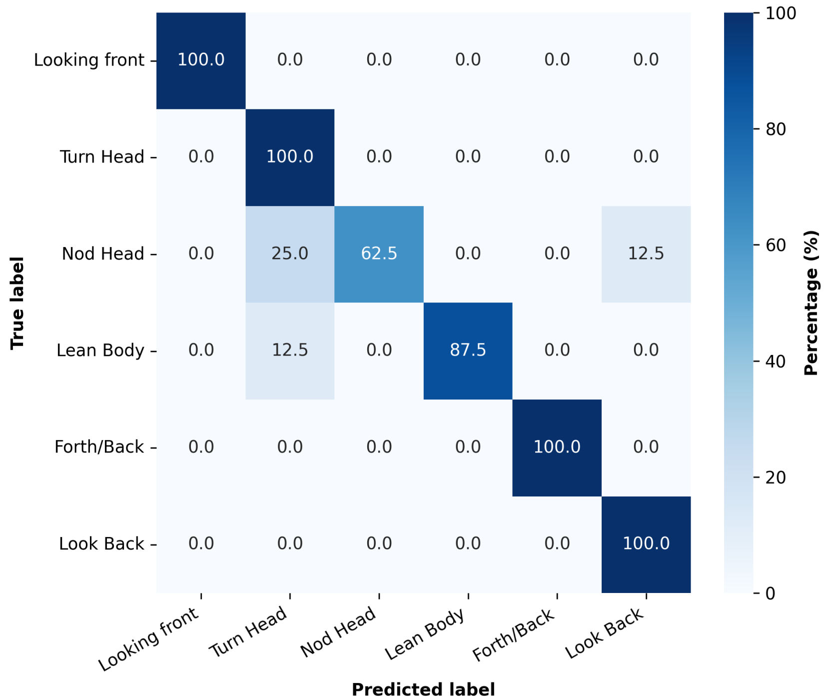


Figure 4.7: Confusion matrix for the fine-grained activity monitoring model with an augmentation factor of 10, achieving an average accuracy of 91.32%.

evaluated on identical data, McNemar's test provides an appropriate paired statistical comparison. The test yielded a highly significant result ($p = 6.10 \times 10^{-5}$), confirming that the increase in accuracy from approximately 60% to over 91% is not due to random variation but reflects a consistent improvement introduced by data augmentation.

Despite these improvements, residual confusion persists between *Look Back* and *Lean Body*, which share similar frontal micro-Doppler characteristics. This limitation reflects an inherent ambiguity when using a single frontal radar perspective. Incorporating angular diversity through multiple antennas or elevation and azimuth information could further mitigate this ambiguity, an aspect revisited in Section 4.5.

4.4 Movement Compensation for Radar-Based Vital Signs Estimation

Radar-based non-contact estimation of physiological vital signs has emerged as a promising alternative to conventional contact-based monitoring, particularly in scenarios where comfort, compliance, and hygiene are critical, such as paediatric and long-term care settings [172, 242]. Using FMCW radar, subtle chest wall micro-motions associated with HR and RR can be sensed without physical contact, enabling continuous monitoring through clothing and bedding [238, 239].

Despite these advantages, reliable radar-based vital signs estimation remains challenging in the presence of body movement. Random and involuntary motion introduces strong interference components that overlap with the physiological frequency bands, masking the comparatively weak signatures of respiration and cardiac activity [167, 172]. This degradation has been explicitly observed in recent public mmWave vital signs datasets, where estimation accuracy drops markedly during subject movement [173].

Addressing motion-induced artefacts is therefore essential for extending radar-based vital signs monitoring beyond controlled laboratory conditions towards realistic clinical and in-home environments. In this section, the focus is placed on movement-compensated estimation of HR and RR using FMCW radar. Rather than constraining subject behaviour or introducing additional sensing modalities, the proposed approach targets signal-level separation of physiological components from motion interference through advanced decomposition techniques, directly addressing the physiological-scale sensing challenge **C1.3**.

4.4.1 Public mmWave Vital Signs Dataset

The vital signs analysis in this chapter is conducted using a publicly available mmWave radar dataset introduced in [173]. The dataset was collected using a Texas Instruments IWR6843 FMCW radar operating in the 60–64 GHz band, with a sweep bandwidth of 3.75 GHz and a starting frequency of 60.25 GHz. The radar configuration employed a single transmit antenna and four receive antennas, transmitting two chirps per frame at a frame rate of 20 frames/s. Each recording session lasted 300 s, resulting in a total of 6,000 frames per subject.

The dataset comprises recordings from 50 children under the age of 13, with a balanced gender distribution of 24 males and 26 females. To ensure safety and postural stability for participants under 6 years of age, additional seating support was provided using a child car seat. Radar measurements were synchronised with a clinically approved reference device, the Nihon Kohden BSM6501K, which provided ground-truth HR and RR annotations.

This publicly available dataset was selected because, to the author's knowledge, it is one of the few open mmWave vital-signs datasets that includes movement conditions. It therefore enables reproducible evaluation of the proposed motion-compensation method against an external benchmark, rather than relying only on self-collected data. The cohort of 50 children provides a challenging test case because paediatric subjects typically exhibit higher and more variable HR and RR than adults and are less able to remain still during measurement. As a result, the dataset places strong demands on the motion-artefact-suppression capability targeted by the proposed method.

Importantly, the dataset includes periods of spontaneous body movement, reflecting realistic paediatric behaviour rather than highly constrained laboratory conditions. While accurate HR and RR estimation was reported under minimal-motion conditions, the original study observed clear degradation during movement phases [173]. This characteristic makes the dataset particularly suitable for evaluating the robustness of the proposed motion compensation under physiologically realistic and motion-affected conditions. The proposed method itself is not age-specific; rather, the paediatric cohort is used as a demanding validation scenario for motion robustness. Nevertheless, the demographic mismatch between this dataset and the adult AAL and driver-monitoring use cases targeted in this thesis is acknowledged and evaluation on adult vital-signs datasets should motivate future work.

4.4.2 Radar-Based Vital Signs Processing Chain

The extraction of HR and RR from FMCW radar measurements in this work follows the principles established in Section 2.3, where small periodic chest wall displacements induced by respiration and cardiac activity are encoded as phase modulations in the radar slow-time signal. These phase variations contain low-frequency components associated with respiration and higher-frequency components associated with cardiac motion.

Figure 4.8 illustrates a conventional FMCW radar vital signs processing chain. Following range processing and target selection, a single complex slow-time signal corresponding to chest motion is extracted. Phase unwrapping is then applied to obtain a continuous phase signal that captures cumulative displacement over time, forming the basis for subsequent physiological analysis.

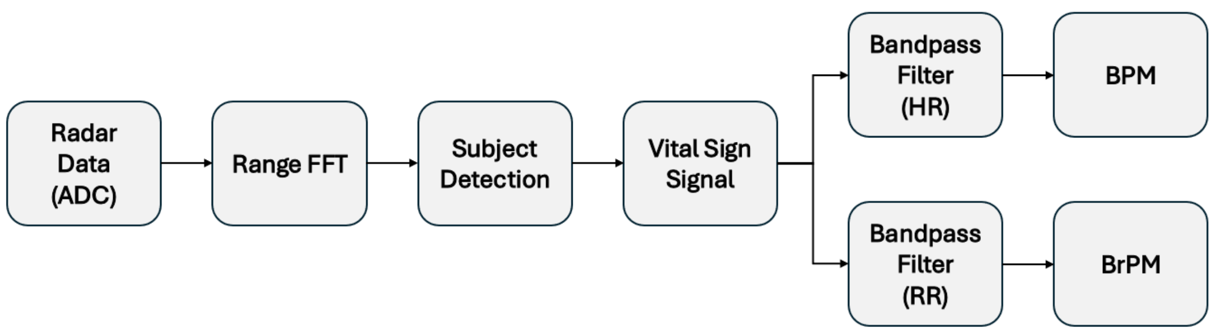


Figure 4.8: Conventional FMCW radar vital signs processing chain, illustrating phase-based extraction of respiration and cardiac components using band-pass filtering.

In this chapter, the baseline processing chain follows the methodology reported in the original public dataset study [173], including phase-based extraction and band-pass filtering of respiration and cardiac components. The band-pass filters were set to 0.1–0.6 Hz for RR and 0.8–4.0 Hz for HR.

Under minimal motion conditions, this baseline pipeline enables reliable estimation of HR and RR through spectral analysis and peak detection. However, in the presence of voluntary or involuntary body movement, motion-induced phase variations dominate the slow-time signal. These components overlap spectrally with the physiological bands, leading to distorted frequency estimates and degraded accuracy. This limitation was explicitly observed in the original dataset, where estimation performance dropped significantly during periods of subject movement [173].

In this work, the baseline HR and RR extraction pipeline is retained as a reference stage. Motion compensation is introduced after band-limited physiological signal extraction, rather than modifying the initial preprocessing steps. Specifically, VMD is applied to the respiration- and cardiac-band signals to separate physiological oscillations from motion-related interference at the signal level. This design enables principled enhancement of robustness without imposing additional constraints on subject behaviour or sensing configuration.

4.4.3 VMD for Motion Artefact Suppression

Motion artefacts constitute the primary source of performance degradation in radar-based HR and RR estimation under realistic conditions. Voluntary and involuntary body movements introduce non-stationary components whose amplitudes often exceed those of the physiological signals, leading to spectral overlap and masking effects that cannot be adequately resolved using fixed band-pass filtering alone. To address this limitation, this work adopts VMD as a post-filtering signal decomposition technique for motion artefact suppression.

VMD is an adaptive, non-recursive decomposition method that represents a signal as a sum of band-limited intrinsic mode functions (IMFs), each characterised by compact frequency support and slowly varying instantaneous frequency [243]. Unlike empirical mode decomposition, VMD formulates the decomposition as a constrained variational optimisation problem, improving robustness to noise and reducing mode mixing. These properties make VMD well-suited for physiological radar signals acquired under motion, where frequency content is non-stationary and partially overlapping.

In the proposed framework, VMD is applied to the respiration- and cardiac-band signals obtained after baseline preprocessing, as described in Section 4.4.2. Let $s_{\mathrm{RR}}(t)$ and $s_{\mathrm{HR}}(t)$ denote the band-limited phase signals corresponding to respiration and cardiac activity, respectively. Each signal is decomposed using VMD into a finite set of IMFs,

$$s(t) = \sum_{k=1}^{K} u_k(t), \tag{4.4}$$

where each mode $u_k(t)$ represents an amplitude-modulated and frequency-modulated oscillatory component.

In the implemented pipeline, VMD was applied separately to the RR- and HR-band signals obtained after conventional band-pass filtering described earlier. The decomposition parameters were fixed for all recordings, with $K = 6$, $\alpha = 10{,}000$, $\tau = 0$, and convergence tolerance 10^{-9}.

Because the signals are already restricted to the physiological frequency bands by the preceding band-pass filters, the dominant oscillatory component was observed to be concentrated in the first mode of the VMD decomposition in all recordings analysed. Therefore, the first mode produced by VMD was selected as the denoised physiological signal for both RR and HR estimation,

$$s_{\mathrm{RR,denoised}}(t) = u_1^{\mathrm{RR}}(t), \qquad s_{\mathrm{HR,denoised}}(t) = u_1^{\mathrm{HR}}(t), \tag{4.5}$$

where $u_1^{\mathrm{RR}}(t)$ and $u_1^{\mathrm{HR}}(t)$ denote the first intrinsic mode function obtained from the VMD decomposition of the respiration- and cardiac-band signals, respectively.

This strategy reduces the influence of residual motion components distributed across the remaining modes while preserving the dominant physiological oscillation. The resulting motion-compensated processing chain is illustrated in Figure 4.9, where VMD is introduced as a post-filtering refinement stage within the baseline vital signs extraction pipeline.

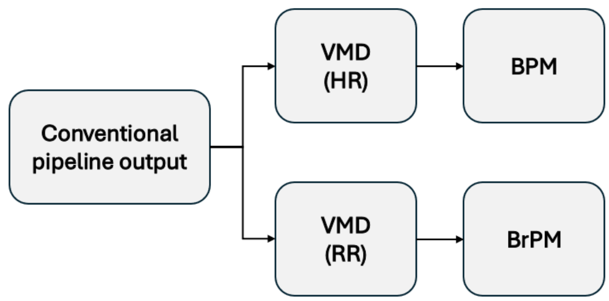


Figure 4.9: Proposed motion-compensated vital signs processing chain, illustrating VMD-based decomposition and reconstruction of HR and RR signals under movement conditions.

By separating physiological oscillations from motion-induced interference at the signal level, the proposed VMD-based stage improves the robustness of HR and RR estimation without introducing additional sensing modalities or behavioural constraints. The quantitative impact of this motion compensation strategy is evaluated in the following subsection.

4.4.4 Performance Evaluation under Movement Conditions

This subsection evaluates the effectiveness of the proposed VMD-based motion compensation framework for radar-based HR and RR estimation under varying movement conditions. The evaluation is conducted using a subset of the public mmWave vital signs dataset introduced in [173], with a particular focus on scenarios where subject movement significantly degrades estimation accuracy.

The analysis is performed on data from eight paediatric subjects selected from the fifty children in the dataset. The subset was chosen to include recordings that exhibit the highest levels of body movement, as the objective of the proposed VMD-based processing is to improve HR and RR estimation under motion. Evaluating the method on recordings with minimal movement would not yield a meaningful assessment of its effectiveness, since conventional band-pass-based processing already achieves reliable results in low-mobility conditions, as shown in

Figure 4.11.

To ensure objective selection, the movement magnitude for each recording was quantified using the MTI-based motion measure, in which frame-to-frame changes in the MTI signal across all receive channels were accumulated to obtain a total movement index. Subjects with larger movement indices were prioritised, while still maintaining a representative age range across the selected recordings. Following the dataset preparation protocol described in Section 4.4.1, invalid entries and missing values were removed prior to processing. All signal processing and performance analysis were carried out using Python. Figure 4.10 presents representative examples of low- and high-mobility patterns extracted from the dataset, illustrating both chest motion and overall body movement for two children of different ages. These examples highlight the substantial variability in movement magnitude encountered in realistic paediatric monitoring scenarios. This variability motivates the focus on higher-mobility recordings in the following evaluation, since these conditions represent the most challenging scenario for radar-based vital signs estimation.

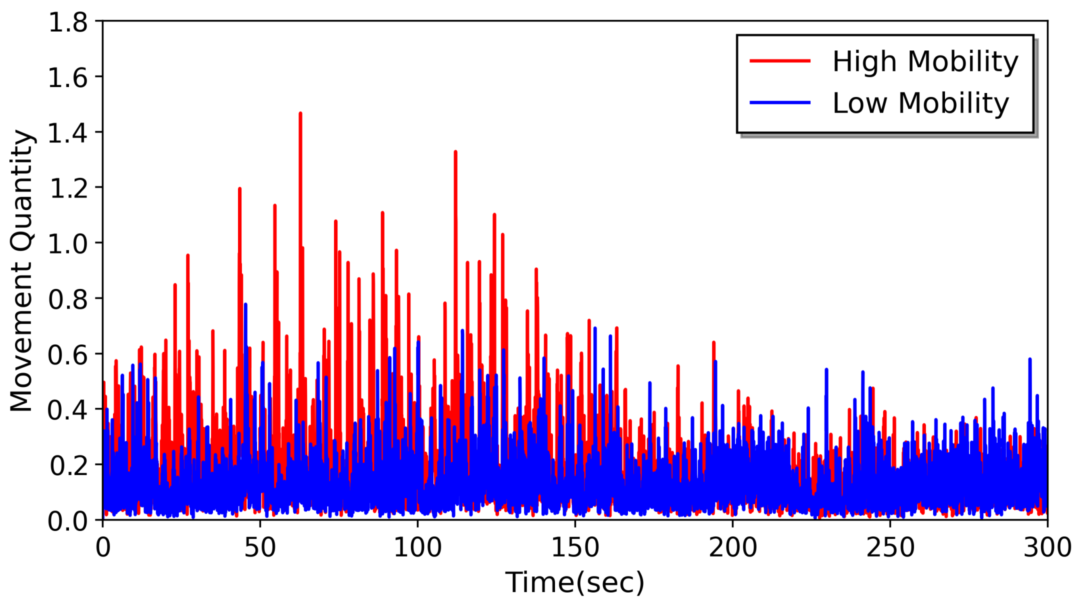


Figure 4.10: Examples of low- and high-mobility patterns in paediatric subjects, illustrating chest and body movement magnitudes extracted from the public dataset.

Estimated Vital Signs under Low-Mobility Conditions

Figure 4.11 shows representative RR and HR estimates over a 30 s window for subject #49 during a low-mobility period. The RR estimates obtained using conventional radar signal processing, as illustrated in Figure 4.8, closely align with the reference measurements, indicating reliable performance when movement is minimal. A similar level of agreement is observed for HR estimation, where the radar-derived signal closely follows the reference trace.

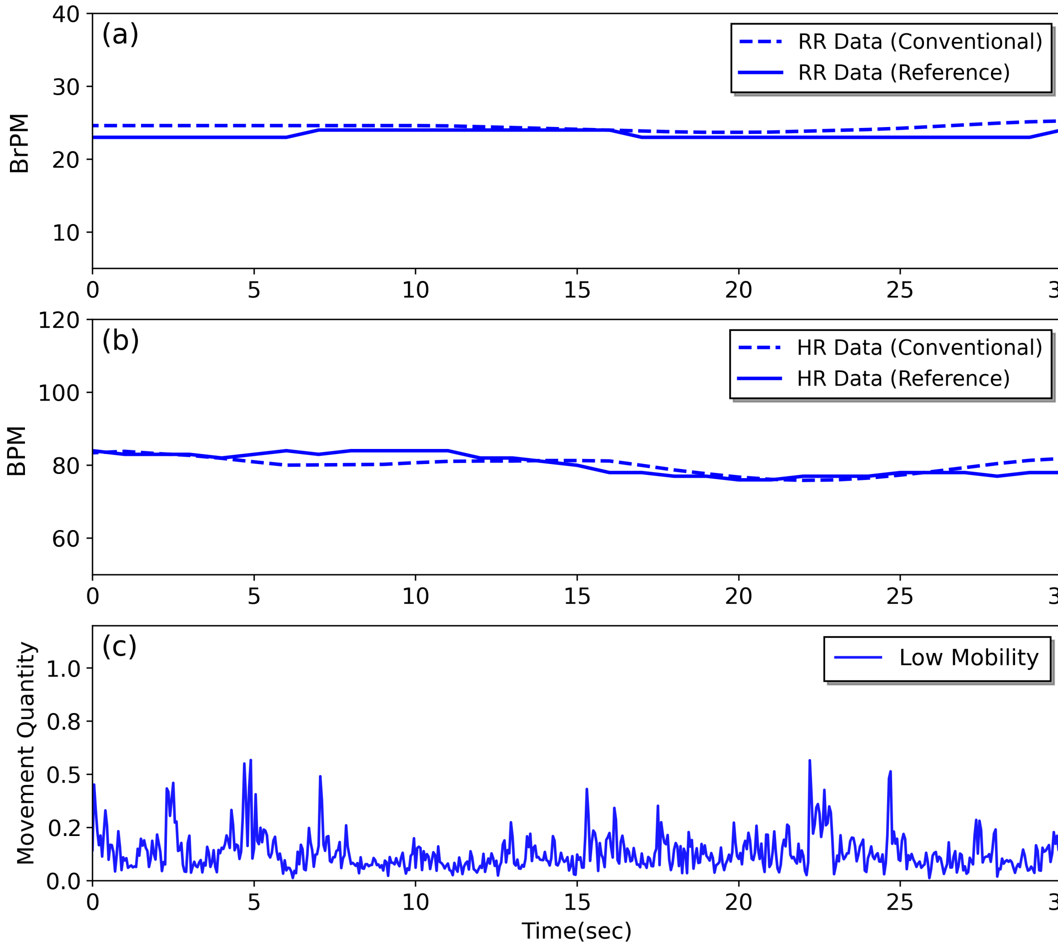


Figure 4.11: Radar-based RR and HR estimation under low-mobility conditions: (a) RR, (b) HR, and (c) corresponding movement magnitude.

These results confirm that, under low-mobility conditions, conventional phase-based radar processing is sufficient for accurate HR and RR estimation. However, as will be demonstrated in the following, this performance degrades markedly as movement intensity increases.

VMD-Processed Vital Signs under High-Mobility Conditions

Figure 4.12 presents RR and HR estimates for subject #39 over a 30 s interval characterised by pronounced movement. Under these conditions, conventional radar processing exhibits substantial deviations from the reference measurements for both RR and HR. In contrast, the VMD-processed signals closely track the reference traces, demonstrating improved robustness against movement-induced interference.

The corresponding movement magnitude is shown in Figure 4.12, highlighting the challenging motion conditions under which the VMD-based approach maintains reliable estimation. These results demonstrate that VMD effectively suppresses motion-induced artefacts that otherwise dominate the physiological frequency bands.

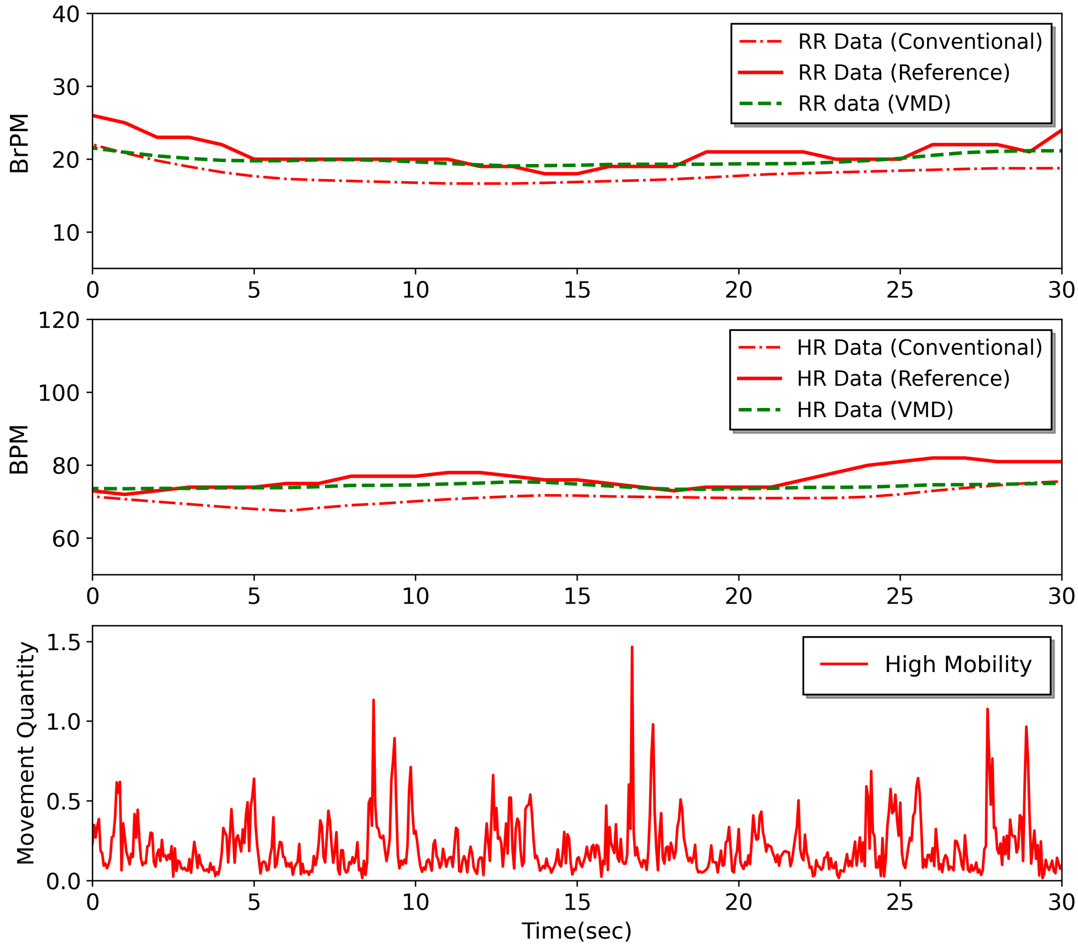


Figure 4.12: Radar-based RR and HR estimation under high-mobility conditions, comparing conventional processing and VMD-enhanced estimation.

Error Analysis of HR and RR Estimation

To quantitatively assess performance, mean absolute error (MAE) and root mean square error (RMSE) are computed for both RR and HR across the selected eight subjects who exhibit high mobility. Figures 4.13 and 4.14 compare the error metrics obtained using conventional processing and the proposed VMD-based approach.

For RR estimation, conventional processing yields MAE values ranging from 1.3 to 3.5 BrPM, whereas the VMD-enhanced approach reduces the MAE range to 0.7–2.4 BrPM. A similar trend is observed for RMSE, which decreases from 1.5–3.9 to 0.4–2.9 BrPM when VMD is applied. For HR estimation, the VMD-based approach consistently outperforms the conventional method. MAE values are reduced from 3.0–7.7 BPM to 0.9–4.7 BPM, while RMSE values decrease from 3.7–8.7 BPM to 1.1–6.3 BPM.

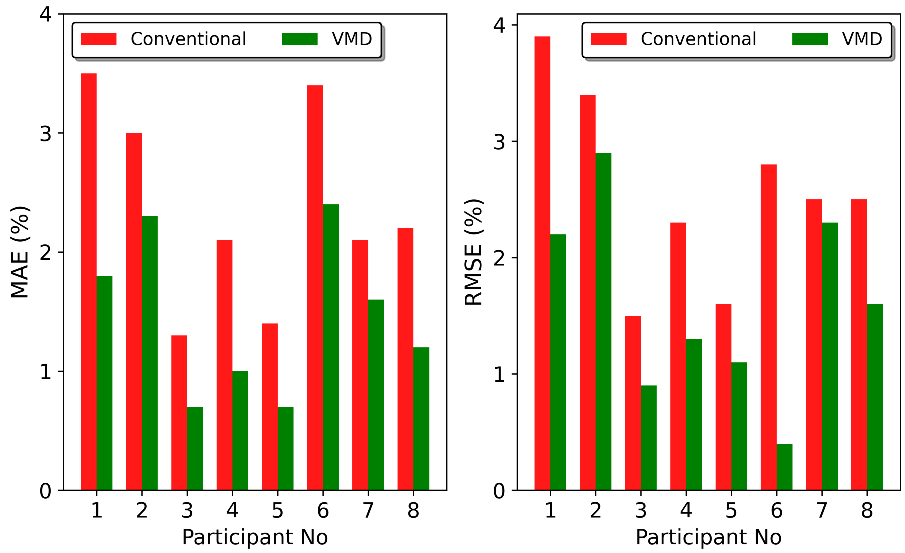


Figure 4.13: Comparison of MAE and RMSE for RR estimation using conventional processing and the proposed VMD-based approach.

Average Error Reduction across Subjects

Figure 4.15 summarises the average MAE and RMSE across the evaluated subjects for both RR and HR. The VMD-based approach achieves an average error reduction of approximately 40.59% compared to conventional processing for the selected subjects under high mobility. This improvement is consistently observed for both physiological metrics, indicating enhanced robustness under movement conditions.

To assess whether the observed improvement was statistically significant, a paired Wilcoxon signed-rank test was performed comparing the estimation errors obtained with the conventional pipeline and the proposed VMD-based method across the eight subjects. The test was applied to both MAE and RMSE values, treating each subject as a paired observation. For MAE, the Wilcoxon test yielded $W = 0$ with $p = 0.0078$, and the same result was obtained for RMSE ($W = 0$, $p = 0.0078$). Since $p < 0.05$, the null hypothesis that both methods produce equivalent errors can be rejected, indicating that the VMD-based processing provides a statistically significant reduction in estimation error compared with the conventional band-pass filtering approach.

This result supports the earlier quantitative improvement and confirms that the performance gain is not attributable to random variation across subjects but reflects a consistent advantage of the proposed motion-compensated processing pipeline.

4.5 Integration and Discussion

This chapter addressed fine-grained human monitoring through two complementary perspectives: the analysis of subtle upper-body and head movements, and the estimation of physiological vital signs under motion. Although these investigations target different outputs, they are

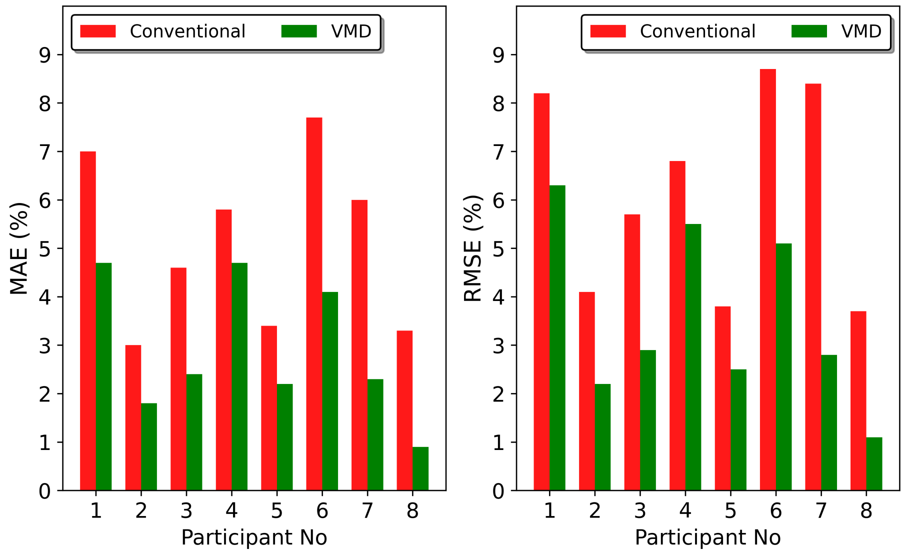


Figure 4.14: Comparison of MAE and RMSE for HR estimation using conventional processing and the proposed VMD-based approach.

unified by shared challenges inherent to fine-grained radar sensing, including low-amplitude signal characteristics, strong sensitivity to motion interference, and limited availability of labelled data.

In the first part of the chapter, fine-grained upper-body and head movements were analysed as indicators of human behaviour in constrained environments. These movements generate micro-Doppler signatures that are substantially weaker and more overlapping than those associated with large-scale activities. The results demonstrate that, under such conditions, data-efficient learning strategies are critical. In particular, the combination of TF representations with targeted augmentation and lightweight transfer learning enabled robust discrimination between subtle movements, even with a limited number of subjects and repetitions.

The second part of the chapter focused on physiological monitoring, where the challenge of motion interference is even more pronounced. Random body movements introduce phase variations that overlap spectrally with respiration- and cardiac-related components, leading to significant degradation in estimation accuracy. By introducing VMD as a post-filtering refinement stage, the proposed framework demonstrated improved robustness in separating physiological rhythms from motion-induced artefacts, without constraining subject behaviour or modifying the sensing setup.

Across both studies, a common limitation emerges from the use of a single frontal radar perspective. Ambiguities arise when different movements or motion components produce similar radar signatures, as observed in fine-grained movement classification and in vital signs estimation during periods of high mobility. These findings highlight the potential benefit of incorporating additional spatial diversity, such as angular information or multi-antenna configurations, to further disambiguate motion sources. While such extensions fall outside the scope of this

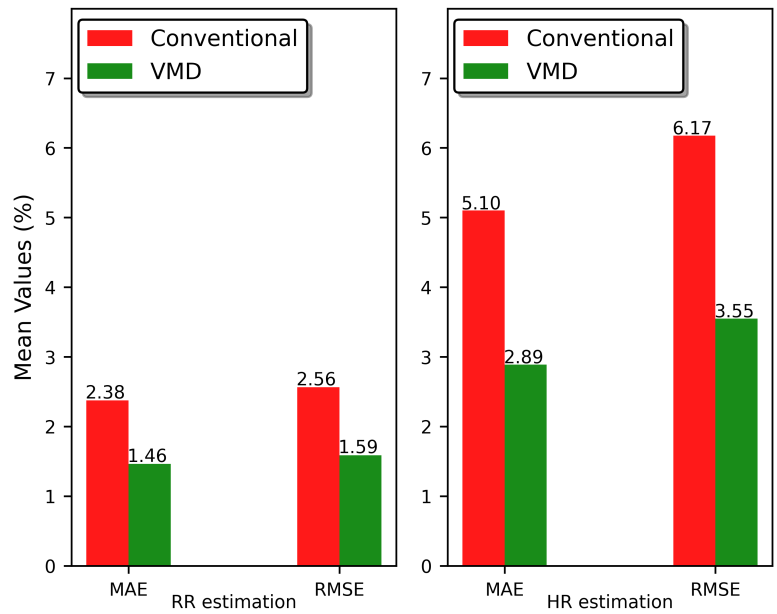


Figure 4.15: Average MAE and RMSE comparison between conventional and VMD-based processing: (a) RR estimation, (b) HR estimation.

chapter, they represent a natural direction for advancing fine-grained radar-based monitoring.

Taken together, the contributions of this chapter demonstrate that both fine-grained activity cues and physiological signals can be extracted at fine temporal and spatial resolutions using single radar, provided that signal processing and learning strategies are explicitly designed to address data scarcity and motion interference. This integrated perspective establishes a foundation for broader human-centric monitoring, where behavioural and physiological information are treated as complementary rather than isolated sensing targets.

4.6 Summary

This chapter addressed radar-based human monitoring beyond large-scale activities by focusing on fine-grained behavioural analysis and physiological-scale vital signs estimation. By targeting small-amplitude motions that are poorly captured by conventional activity recognition approaches, the chapter directly addressed challenges in sensing, interpretation, and robustness at fine spatial and temporal scales.

For fine-grained HAM, a non-contact mmWave radar framework was developed to recognise subtle upper-body and head movements in seated scenarios. The results demonstrated that data-efficient learning strategies, particularly targeted STFT radar representation map augmentation combined with lightweight transfer learning, are essential for achieving robust subject-level generalisation when training data are limited. For physiological-scale monitoring, the chapter investigated radar-based estimation of HR and RR under movement conditions. Motion artefacts were identified as a dominant source of performance degradation, and integrating VMD into the signal processing pipeline was shown to improve robustness by separating physiological oscillations from movement-induced interference.

In the next chapter, these concepts are extended beyond radar-only sensing by considering multimodal monitoring scenarios. In particular, behavioural and physiological information are examined jointly using heterogeneous sensors, enabling a broader exploration of human state monitoring in real-world conditions.

Chapter 5

Multimodal Fusion for Human-State Monitoring

The preceding chapters of this thesis examined human monitoring at multiple sensing resolutions, addressing large-scale activities, fine-grained movements, and physiological signals in isolation. While these studies demonstrated that individual sensing modalities can reliably capture specific aspects of human behaviour or physiology, they also revealed a fundamental limitation: unimodal observations are often insufficient for inferring a person's underlying cognitive or functional state. Similar observable patterns may arise from distinct internal conditions, leading to ambiguity when behavioural or physiological cues are interpreted independently.

This limitation motivates the final contribution of the thesis, which addresses the challenge of comprehensive human-state assessment (**C3**) through multimodal fusion. Rather than treating behavioural and physiological cues independently, this chapter investigates how complementary metrics can be integrated to provide a more robust and interpretable representation of human state. Specifically, the chapter focuses on the fusion of gaze-derived behavioural indicators and HR as a physiological measure of internal response.

A key strength of this study is the use of a naturalistic multimodal dataset collected during approximately 480 minutes of real-world driving. In contrast to many existing studies that rely on driving simulators or controlled laboratory conditions, the data in this chapter were recorded in real vehicles on public roads under varying traffic and environmental conditions. This setting provides a more realistic evaluation of multimodal human monitoring and allows behavioural and physiological responses to be analysed under authentic workload and attention demands.

This chapter addresses challenge **C3** by investigating multimodal fusion for human-state monitoring in realistic driving conditions. Section 5.3 describes the multimodal data collection protocol, Section 5.4 presents the proposed fusion framework, and Section 5.5 reports the experimental results and analysis.

5.1 Introduction

Accurately inferring human state in real-world environments remains a central challenge in human-centric monitoring systems. Behavioural observations alone may fail to distinguish between intentional actions and degraded performance, while physiological signals often lack contextual grounding when interpreted in isolation [38]. A holistic understanding of the human state, therefore, requires the integration of multiple complementary data streams that jointly reflect outward behaviour and internal physiological response.

Driving provides a representative and safety-critical context in which this challenge becomes evident. In naturalistic driving, changes in behaviour and physiology may arise from diverse factors, including cognitive workload, stress, fatigue, or environmental demands. Road traffic accidents remain a major global public health concern, accounting for more than 1.35 million fatalities annually [244], with driver-related factors such as distraction and reduced alertness consistently identified as dominant contributors [245, 246]. These observations highlight the importance of monitoring driver state, but more critically, they underscore the need for state inference beyond any single cue.

Physiological monitoring has emerged as a valuable means of capturing internal driver state. HR, in particular, has been shown to reflect variations in stress, cognitive workload, and fatigue during driving [208, 214, 247]. However, HR alone remains inherently ambiguous, as similar cardiac responses may be elicited by distinct cognitive or emotional processes. In parallel, gaze-based monitoring has become a cornerstone of behavioural analysis in driver monitoring systems. Metrics such as blink rate, fixation duration, gaze dispersion, and saccadic behaviour provide direct insight into visual attention and task engagement [248–250]. Despite their effectiveness, gaze-only systems struggle to disambiguate internal states such as stress and fatigue, particularly under dynamically changing task demands [209].

These complementary limitations motivate multimodal approaches that integrate behavioural and physiological metrics. By jointly analysing gaze behaviour and HR, it becomes possible to contextualise internal physiological responses with outward attentional patterns, reducing ambiguity and improving the reliability of human-state inference. Prior studies have demonstrated that such integration enhances the detection of fatigue, distraction, and cognitive load compared with unimodal systems [199,202,207]. However, as discussed in Section 2.5 and Table 2.6, many existing multimodal frameworks remain confined to simulated or controlled environments.

This chapter addresses these limitations by presenting a multimodal human-state monitoring framework based on real-world driving data, integrating eye-tracking and HR measurements. Driving is used as a representative application domain through which the broader challenge of holistic human-state assessment (**C3**) is examined.

5.2 Chapter Contributions

This chapter advances multimodal human-state assessment by integrating behavioural and physiological metrics in real-world driving. The main contributions are summarised as follows:

1. **Real-world multimodal data collection:** A synchronised dataset comprising HR and eye-gaze measurements is collected during real-world driving. Unlike simulator-based or single-modality studies, the dataset captures naturalistic behaviour across multiple scenarios, enabling analysis under realistic and dynamic conditions.

2. **Scenario- and subject-based statistical analysis:** A systematic analysis examines how different driving environments and individual differences influence the interaction between physiological responses and gaze behaviour. Statistical characterisation of these relationships is used to identify consistent patterns and to inform threshold selection for multimodal state assessment.

3. **Multimodal driver monitoring algorithm:** A real-time multimodal algorithm is proposed that integrates physiological information (HR) with behavioural gaze metrics, including blink, fixation, and saccade characteristics. A weighted scoring framework is developed to classify driving segments as normal or abnormal.

5.3 Data Collection Protocol

This section describes the end-to-end protocol used to collect and curate the real-world multimodal dataset analysed in this chapter, and it establishes the procedural foundations for the scenario-based analysis and the proposed multimodal algorithm. The protocol is organised around three stages: (i) naturalistic driving data acquisition using synchronised wearable sensors, (ii) extraction and structuring of HR and gaze-derived events for statistical analysis, and (iii) preparation of aligned multimodal segments for algorithm development and real-time scoring. An overview of the workflow is shown in Figure 5.1.

5.3.1 Experimental Setup and Participants

Data were collected in real-world driving sessions to capture naturalistic behavioural and physiological variability under realistic operating conditions. Although the study is presented as a pilot dataset, the resulting corpus is substantial, comprising 480 minutes of driving across multiple sessions and days. This quantity provides a sufficiently large number of samples and gaze events to support preliminary statistical characterisation and to motivate the development of a multimodal scoring algorithm.

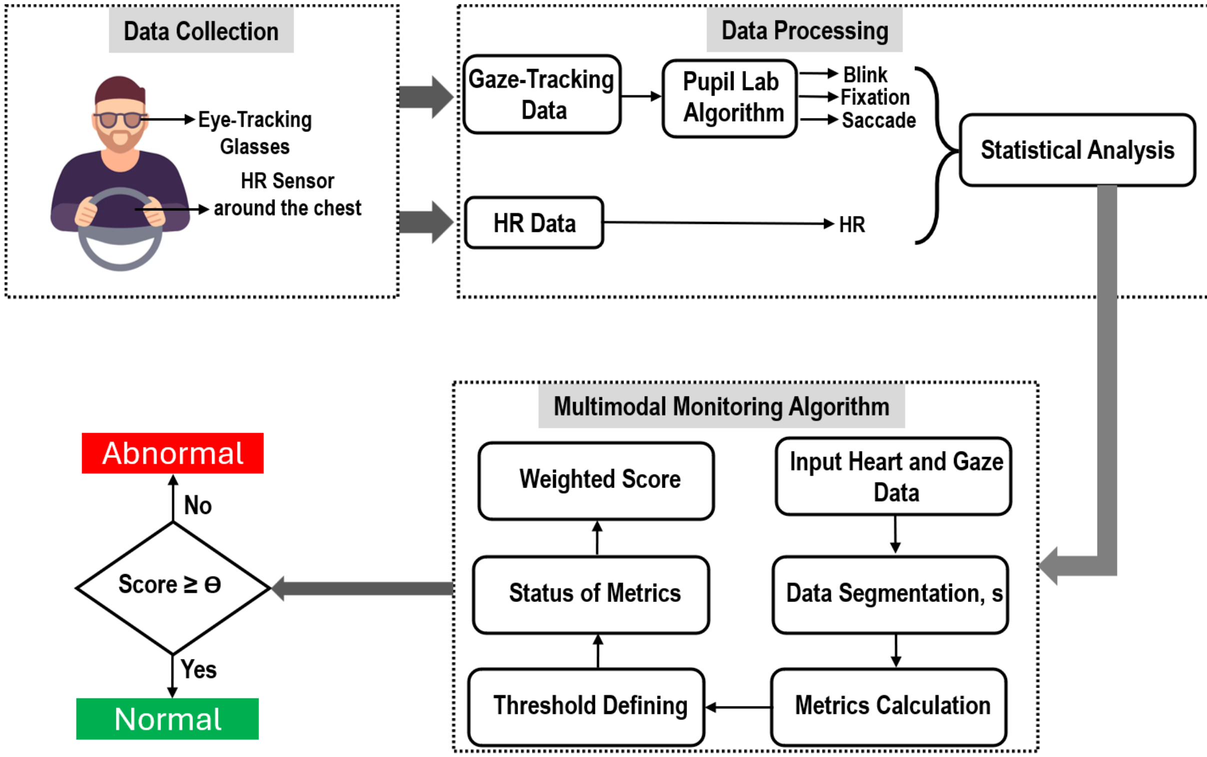


Figure 5.1: End-to-end workflow for the driver monitoring study, from real-world data collection to gaze–HR feature extraction and fusion for driver state monitoring.

Five licensed drivers participated in the study. To reduce confounding effects associated with unfamiliar vehicle dynamics, each participant drove their own vehicle. Participant demographics are summarised in Table 5.1. The cohort reflects experienced drivers with varied ages and driving experience, enabling an initial examination of between-subject variability in gaze behaviour and HR response.

This cohort size is consistent with the exploratory, hypothesis-generating aim of a pilot naturalistic-driving study, where the objective is to characterise within- and between-subject variability and to validate the end-to-end multimodal pipeline under real driving conditions, rather than to establish population-level effect sizes. Its value lies in combining approximately 480 minutes of real-world driving with synchronised HR and gaze measurements, enabling preliminary scenario- and subject-level analysis under realistic conditions. Accordingly, the results are presented as an initial demonstration of feasibility, and no population-level generalisation is claimed regarding driver behaviour, gender differences, age effects, or general threshold values.

All participants in this pilot dataset were male, limiting the study's demographic diversity and potentially influencing the behavioural and physiological patterns observed during driving.

Before each recording session, a structured familiarisation stage was conducted in a parking area. Participants wore all sensors for at least 20 minutes while a copilot verified device operation, sensor placement, and participant comfort. Participants also performed short, low-risk driving loops in the parking area to acclimate to the equipment. The main road session began

Table 5.1: Participant demographics and driving experience.

Participant	Age	Gender	Driving experience (Years)
1	35–40	Male	15–20
2	25–30	Male	5–10
3	30–35	Male	5–10
4	30–35	Male	10–15
5	40–45	Male	20–25

only after the participant confirmed that the setup did not interfere with driving safety.

Sensing configuration and roles.

Two wearable sensing systems were used. Eye gaze was recorded using Pupil Labs Invisible glasses [42] connected to a companion smartphone for continuous acquisition. HR was recorded using the Polar H10 chest strap [41]. A forward-facing dashcam recorded the road scene to support reference checking and potential manual inspection of notable segments. Throughout each drive, the copilot handled practical aspects of the sensing system, including initiating and terminating recordings, monitoring signal quality, and managing device synchronisation procedures. The full setup is illustrated in Figure 5.2.

Environmental diversity and driving scenarios.

Data collection was conducted primarily during daytime and under naturally varying weather conditions, including clear sky, overcast conditions, and light rain. Extreme weather (for example, heavy rain or dense fog) was excluded to avoid systematic degradation of visual sensing and to maintain more consistent measurement conditions across sessions. Traffic conditions were not controlled, and sessions therefore include natural transitions between low-density flow and congested stop-and-go traffic. This uncontrolled variability is treated as a feature of the dataset, since the objective is to evaluate multimodal monitoring under realistic conditions, while recognising that future studies will benefit from more systematic control or explicit annotation of environmental covariates.

Driving sessions contained mixed road types, with participants alternating between urban and motorway segments across days. The resulting dataset comprises approximately 66% urban driving and 33% motorway driving. Across all sessions, the dataset includes 28,963 HR samples, 10,669 blink events, 63,300 fixation events, and 63,050 saccade events. These event counts yield a dense behavioural stream that supports both distribution-level analysis and segment-level multimodal scoring. All recordings were conducted with participant agreement and without collecting sensitive identifiers.

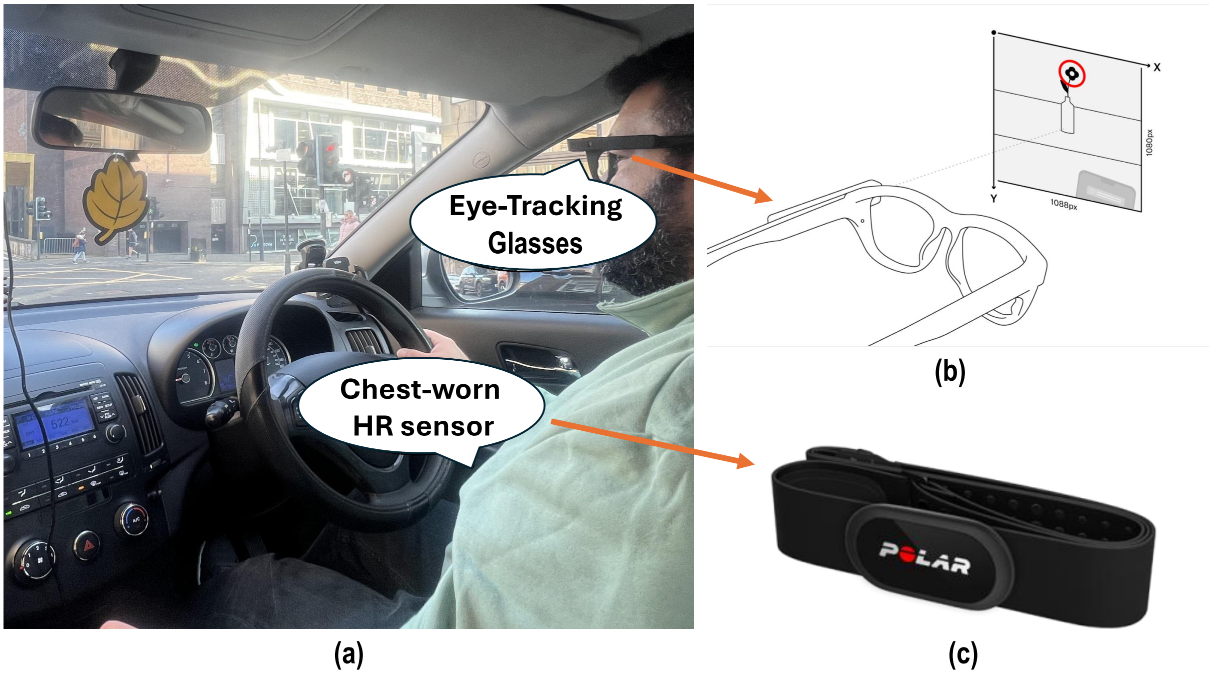


Figure 5.2: Real-world data collection setup: (a) in-vehicle recording configuration, (b) Pupil Labs Invisible eye-tracking glasses, and (c) Polar H10 chest strap for HR acquisition.

5.3.2 HR and Gaze Data Acquisition

This subsection details the acquisition characteristics of the physiological (HR) and behavioural (gaze) streams, including device configuration, sampling, and event extraction.

HR Acquisition

HR was recorded using the Polar H10 chest strap [41], which is widely used in research and is reported to provide reliable HR measurement across diverse use cases [251]. The device records ECG at 1000 Hz [251] and outputs HR as a time series. In this study, HR values were exported at 1 Hz, which is sufficient to capture temporal variations over the time scales relevant to naturalistic driving segments and scenario-level comparisons. The strap was worn around the lower chest with firm contact to stabilise electrode coupling and reduce motion-induced artefacts during prolonged sessions.

Data collection and session management were performed using the Polar Beat smartphone application. After each session, recordings were uploaded to the Polar cloud for storage and then downloaded for offline processing and analysis.

This chapter focuses on HR rather than HRV. HR was selected because it provides a direct and interpretable indicator of physiological response that can be integrated in real time with low computational overhead, and it has demonstrated sensitivity to driving-related demands in prior work [208, 214, 247].

Gaze Acquisition and Event Extraction

Gaze data were acquired using Pupil Labs Invisible glasses [42], designed for dynamic real-world recording. The device includes a forward-facing scene camera and two infrared eye cameras. The eye cameras capture eye images at a resolution of 192×192 pixels at up to 200 Hz. A companion smartphone performs on-device gaze estimation using the Pupil Labs processing pipeline [252]. During recording, the effective gaze output rate depends on the companion device. In this work, using a OnePlus 8 smartphone, the observed gaze output exceeded 120 Hz. After upload to the Pupil Labs cloud, gaze signals can be reprocessed at the device's maximum rate of 200 Hz, enabling consistent offline event reconstruction.

Blink events were detected using the Pupil Labs blink detection method, which employs an XGBoost-based classifier trained on device-specific data [253]. This approach uses patterns associated with eyelid closure and reopening to detect blink events, and it is reported to achieve high recall with low false detections in real-world settings [253].

Fixations and saccades were identified using the Pupil Labs event detection pipeline based on an extended I-VT approach (identification by velocity threshold) [254]. The method compensates for head and body motion using optical flow correction and applies adaptive velocity thresholds to improve robustness under dynamic scenes. Fixation candidates below physiologically plausible duration thresholds were removed to reduce spurious detections. Saccades were detected as high-velocity transitions between fixation periods under standard constraints on event velocity, amplitude, and duration [254]. Gaze velocity is reported in pixels per second, as it is computed directly from the eye tracker's scene-camera coordinate frame; since all participants used the same Pupil Labs Invisible eye tracker [42] with identical scene-camera resolution and field of view, pixel-based velocities are directly comparable across participants. The resulting blink, fixation, and saccade event streams serve as the behavioural inputs for the scenario-based analysis and the multimodal scoring framework in the following section.

5.4 Multimodal Fusion Framework

This section presents the proposed multimodal fusion framework for human-state monitoring, integrating physiological and behavioural metrics derived from HR and gaze signals. The framework is designed to satisfy three key requirements that reflect the multimodal research gap identified in Section 2.5.3: (i) robustness under real-world driving conditions, (ii) computational simplicity suitable for real-time deployment, and (iii) interpretability of the resulting driver state decision. The framework, therefore, adopts a decision-level fusion strategy, in which statistically grounded metrics are first extracted independently from each modality and then combined via a weighted scoring mechanism.

5.4.1 Preprocessing of HR and Gaze Signals

The collected HR and gaze data were processed through a structured preprocessing pipeline prior to fusion. This pipeline was designed to ensure temporal alignment, remove physiologically implausible artefacts, and extract comparable metrics across subjects and driving scenarios. The preprocessing steps consist of data synchronisation, outlier removal, metric normalisation, and statistical characterisation.

Temporal synchronisation.

HR and gaze data streams were recorded asynchronously by independent devices and therefore required explicit temporal alignment. All timestamps were first converted to a common reference by normalising each signal to the start of its corresponding recording segment. For a given signal, the normalised timestamp t was computed as

$$t = t_{\text{original}} - t_{\min}, \tag{5.1}$$

where t_{original} denotes the original device timestamp and $t_{\min}$ is the minimum timestamp within the recording session. This operation ensured consistent temporal correspondence between HR samples and gaze events.

Outlier removal and physiological plausibility.

To improve data quality and reduce the influence of sensor artefacts, outlier removal was applied to all metrics using physiologically and behaviourally plausible bounds. For a given metric m, samples lying outside predefined lower and upper limits (L_m, U_m) were excluded,

$$\text{Cleaned}_m = \{x \in m \mid L_m \leq x \leq U_m\}. \tag{5.2}$$

The limits were selected based on established ranges reported in the eye-tracking and physiological monitoring literature and were intentionally conservative to avoid excessive data rejection. The adopted limits are summarised in Table 5.2.

Distributional analysis and normalised counts.

Following cleaning, frequency distributions were computed for all key metrics, including blink duration, fixation duration, saccade duration, saccade amplitude, saccade velocity, and HR. Each metric was discretised into bins, and the relative frequency of each bin was calculated as

$$\text{Frequency} = \frac{\text{Count in bin}}{\text{Total count}} \times 100\%. \tag{5.3}$$

Table 5.2: Quality-control bounds applied to HR and gaze metrics based on physiologically and behaviourally plausible limits.

Metric	Lower Limit	Upper Limit
Blink duration (ms)	50	2000
Fixation duration (ms)	50	5000
Saccade duration (ms)	10	1000
Saccade velocity (px/s)	0	10000
Saccade amplitude (deg)	0	50
HR (BPM)	40	200

These distributions provided insight into typical operating ranges and informed the selection of thresholds distinguishing normal from abnormal behaviour.

To enable fair comparison across driving sessions of unequal duration, normalised event rates were also computed. For each segment, blink, fixation, and saccade counts were expressed as rates per unit time (per minute for blinks, per second for fixations and saccades),

$$\text{Event rate} = \frac{\text{Total event count}}{\text{Segment duration}}. \tag{5.4}$$

This normalisation ensured that differences in metric values reflected behavioural variation rather than recording length.

Scenario-based statistical validation.

To verify that selected metrics exhibited meaningful variation across driving scenarios, pairwise statistical comparisons were conducted using Welch's t-test [255], which does not assume equal variances. For two scenario-specific samples X_1 and X_2, the test statistic was computed as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{5.5}$$

where $\bar{X}_i$, s_i^2, and n_i denote the sample mean, variance, and size for scenario i, respectively. Metrics with statistically significant differences ($p < 0.05$) and sufficient sample support were retained to inform threshold selection and weight assignment in the fusion algorithm.

5.4.2 Proposed Multimodal Algorithm: Scoring and Thresholding

Building on the preprocessing and statistical analysis, a decision-level multimodal algorithm is proposed to integrate HR and gaze metrics for real-time driver state monitoring. The design prioritises computational efficiency, interpretability, and adaptability, in line with the requirements identified in Sections 5.1 and Subsection 2.5.3. Rather than fusing raw signals, the algorithm combines modality-specific decisions through a weighted scoring mechanism.

Temporal segmentation.

The continuous data streams are divided into fixed-length temporal segments of duration T. In this study, T was set to 10 s, providing a balance between responsiveness to transient changes and stability against short-term noise. Although the framework is agnostic to segment length, this choice aligns with prior work on short-term cognitive and behavioural state estimation [256].

Metric extraction and status assignment.

For each segment, s, summary statistics are computed, including mean HR (HR_{mean}), mean blink duration (BD_{mean}), mean fixation duration (FD_{mean}), and mean saccade duration, amplitude, and velocity (SD_{mean}, SA_{mean}, SV_{mean}). Each metric is then evaluated against its corresponding thresholds and assigned a binary status, *Normal* or *Abnormal*.

Weighted decision-level fusion.

To reflect the relative contribution of each modality, a weighted scoring scheme is employed. Let w_{HR}, w_{Blink}, w_{Fixation}, and w_{Saccade} denote the weights associated with HR, blink, fixation, and saccade metrics, respectively. The total segment score is computed as

$$\begin{aligned} S_{\text{total}} = {} & w_{\text{HR}} \cdot I\{\text{HR status} = \text{Normal}\} + w_{\text{Blink}} \cdot I\{\text{Blink status} = \text{Normal}\} \\ & + w_{\text{Fixation}} \cdot I\{\text{Fixation status} = \text{Normal}\} + w_{\text{Saccade}} \cdot I\{\text{Saccade status} = \text{Normal}\}, \end{aligned} \tag{5.6}$$

where $I\{\cdot\}$ is an indicator function that returns 1 if the condition is satisfied and 0 otherwise. Initial weights were selected based on the statistical sensitivity of each metric and informed by prior multimodal workload and attention fusion studies [257–259]. These values serve as a baseline and can be refined as larger annotated datasets become available.

Decision rule.

A binary decision is produced by comparing S_{total} with a decision threshold θ. In this pilot study, θ was set to 0.5, yielding

$$\text{Decision} = \begin{cases} \text{Normal}, & S_{\text{total}} \geq \theta, \\ \text{Abnormal}, & S_{\text{total}} < \theta. \end{cases} \tag{5.7}$$

This binary formulation was adopted to support rapid and interpretable intervention in safety-critical contexts [260].

Algorithmic summary.

Algorithm 2 summarises the complete multimodal fusion process, from segment-wise metric extraction to decision-level fusion.

Algorithm 2: Decision-level multimodal driver state monitoring algorithm.

Input : Gaze data (Blinks, Fixations, Saccades), HR data; Segment length T;
Thresholds: $\mathrm{HR}_{\mathrm{lower}}$, $\mathrm{HR}_{\mathrm{upper}}$, $\mathrm{BD}_{\mathrm{lower}}$, $\mathrm{BD}_{\mathrm{upper}}$, etc.;
Weights: $w_{\mathrm{HR}}, w_{\mathrm{Blink}}, w_{\mathrm{Fixation}}, w_{\mathrm{Saccade}}$;
Decision threshold: θ
Output: Final decision per segment (Normal or Abnormal)
foreach *segment, s, of length* T **do**
    `// Calculate segment-wise mean metrics`
    $\mathrm{HR}_{\mathrm{mean}} \leftarrow \mathrm{Mean}(\mathrm{HR\ in}\ BPM)$
    $\mathrm{BD}_{\mathrm{mean}} \leftarrow \mathrm{Mean}(\mathrm{Blink\ duration\ in}\ ms)$
    $\mathrm{FD}_{\mathrm{mean}} \leftarrow \mathrm{Mean}(\mathrm{Fixation\ duration\ in}\ ms)$
    $\mathrm{SD}_{\mathrm{mean}}, \mathrm{SA}_{\mathrm{mean}}, \mathrm{SV}_{\mathrm{mean}} \leftarrow$ Mean saccade metrics in $ms, deg, px/s$
    `// Evaluate thresholds`
    $\mathrm{HR}_{\mathrm{status}} \leftarrow$ Normal if $\mathrm{HR}_{\mathrm{lower}} \leq \mathrm{HR}_{\mathrm{mean}} \leq \mathrm{HR}_{\mathrm{upper}}$; else Abnormal
    $\mathrm{Blink}_{\mathrm{status}} \leftarrow$ Normal if $\mathrm{BD}_{\mathrm{lower}} \leq \mathrm{BD}_{\mathrm{mean}} \leq \mathrm{BD}_{\mathrm{upper}}$; else Abnormal
    $\mathrm{Fixation}_{\mathrm{status}} \leftarrow$ Normal if $\mathrm{FD}_{\mathrm{lower}} \leq \mathrm{FD}_{\mathrm{mean}} \leq \mathrm{FD}_{\mathrm{upper}}$; else Abnormal
    if $SD_{lower} \leq SD_{mean} \leq SD_{upper}$ ***and***
    $SA_{lower} \leq SA_{mean} \leq SA_{upper}$ ***and***
    $SV_{lower} \leq SV_{mean} \leq SV_{upper}$ **then**
        $\mathrm{Saccade}_{\mathrm{status}} \leftarrow$ Normal
    else
        $\mathrm{Saccade}_{\mathrm{status}} \leftarrow$ Abnormal
    end
    `// Calculate total score`
    $S_{\mathrm{total}} \leftarrow w_{\mathrm{HR}} \cdot \mathbf{I}(\mathrm{HR}_{\mathrm{status}} = \mathrm{Normal})$
    $S_{\mathrm{total}} \leftarrow S_{\mathrm{total}} + w_{\mathrm{Blink}} \cdot \mathbf{I}(\mathrm{Blink}_{\mathrm{status}} = \mathrm{Normal})$
    $S_{\mathrm{total}} \leftarrow S_{\mathrm{total}} + w_{\mathrm{Fixation}} \cdot \mathbf{I}(\mathrm{Fixation}_{\mathrm{status}} = \mathrm{Normal})$
    $S_{\mathrm{total}} \leftarrow S_{\mathrm{total}} + w_{\mathrm{Saccade}} \cdot \mathbf{I}(\mathrm{Saccade}_{\mathrm{status}} = \mathrm{Normal})$
    if $S_{total} \geq \theta$ **then**
        Final Decision $\leftarrow$ Normal
    else
        Final Decision $\leftarrow$ Abnormal
    end
end

5.5 Results and Discussion

This section presents the experimental results obtained from the multimodal analysis of approximately 480 minutes of real-world driving data collected from five participants. The dataset spans both city and motorway driving conditions and captures substantial inter-subject and inter-scenario variability in physiological and gaze-derived behavioural metrics.

The results are structured in two stages. First, Subsection 5.5.1 presents descriptive and statistical observations of HR and gaze behaviour across subjects and driving scenarios, establishing the empirical basis for multimodal fusion. Second, Subsection 5.5.2 evaluates the proposed real-time multimodal driver monitoring algorithm using statistically informed thresholds and weights.

5.5.1 Scenario- and Subject-Level Observations

HR Characteristics

Figure 5.3 illustrates the distribution of HR for each participant under city and motorway driving conditions. Across all subjects, HR was consistently higher during city driving, reflecting the increased cognitive and environmental demands associated with urban traffic.

Subject 3 exhibited the highest mean HR in both scenarios (97.35 BPM in the city and 96.54 BPM on the motorway), whereas Subject 5 showed the lowest averages (71.30 BPM and 69.89 BPM, respectively). These differences highlight pronounced inter-subject variability, likely influenced by factors such as driving experience, environmental conditions, and vehicle characteristics. For example, Subject 3 drove in warm conditions without air conditioning, while Subject 5, who had over 20 years of driving experience and used an automatic vehicle, exhibited consistently lower physiological arousal. The subject-level HR distributions show that scenario effects are superimposed on large individual baselines: higher HR in city driving is consistent with increased manoeuvring, traffic density, and attentional demand, but the magnitude of the difference varies by participant. This variability indicates that fixed population thresholds should be used cautiously and motivates the use of subject-aware or adaptive baselines in future work.

Aggregated HR distributions across all subjects are shown in Figure 5.4. A statistically significant difference was observed between scenarios, with higher mean HR during city driving (91.63 BPM) compared with motorway driving (89.42 BPM), confirmed by Welch's t-test ($t = -21.093$, $p < 0.001$). These findings support HR as a sensitive physiological indicator of scenario-dependent driving demand.

Blink and Fixation Behaviour

Blink and fixation rates were analysed using normalised counts to enable fair comparison across driving sessions of varying duration. Figure 5.5 presents blink counts per minute and fixation counts per second for all subjects under both driving scenarios.

Blink rates were consistently higher on the motorway, whereas fixation rates were higher in city driving. For instance, Subject 1 exhibited an average blink rate of 31.03 ± 6.89 blinks/min on the motorway compared with 26.61 ± 5.91 blinks/min in the city. Conversely, fixation rates for the same subject were higher in the city (2.18 ± 0.29 fixations/s) than on the motorway (1.77 ± 0.34 fixations/s).

When aggregated across subjects, these differences were statistically significant. Motorway driving yielded higher blink rates (24.29 ± 8.02 blinks/min) than city driving (21.20 ± 7.51 blinks/min; $t = 4.037$, $p = 6.8 \times 10^{-5}$). Fixation rates showed the opposite trend, with higher values in

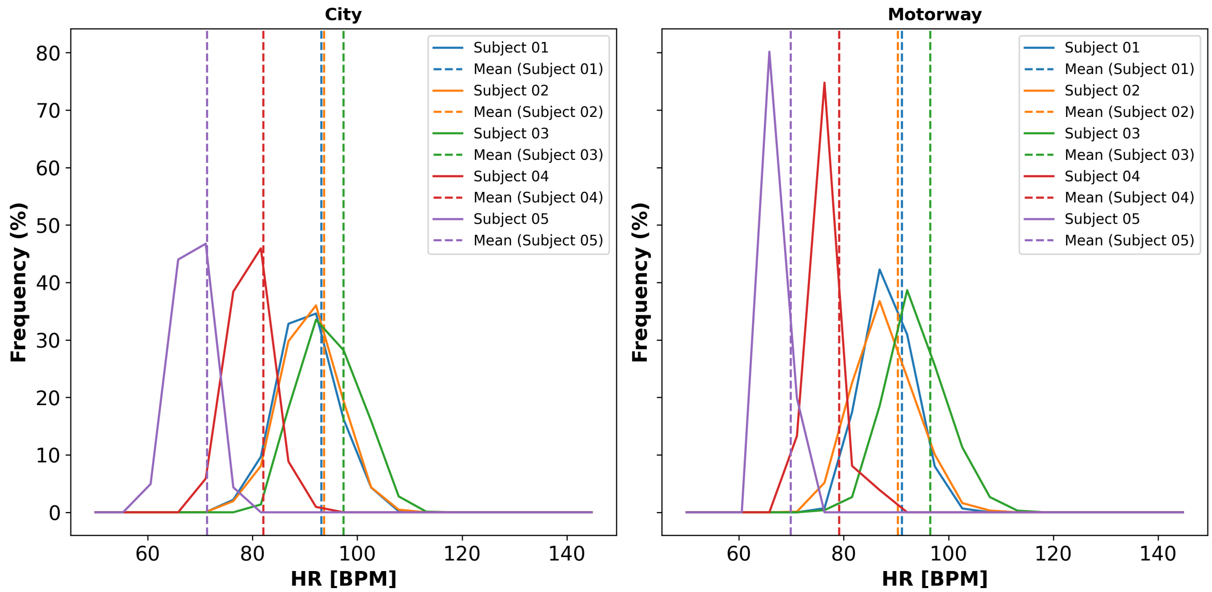


Figure 5.3: HR distribution of each participant during motorway and city driving.

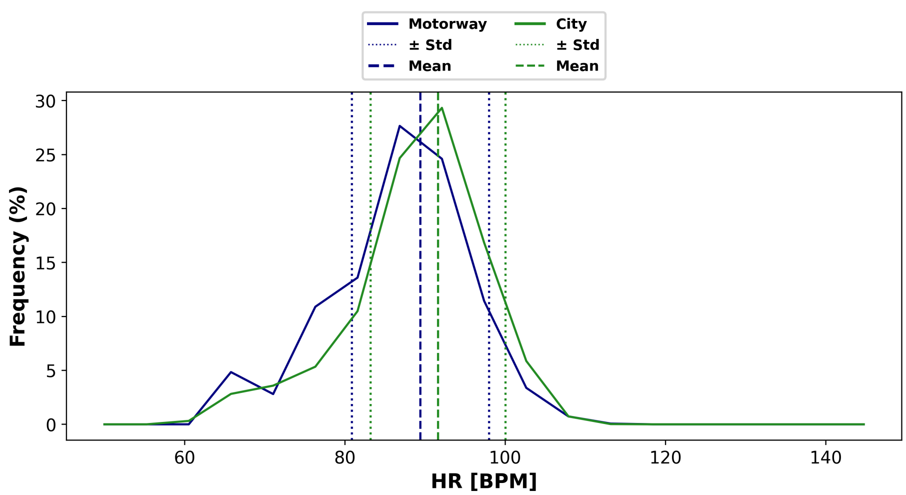


Figure 5.4: Aggregated HR distribution across all subjects for motorway and city driving.

city driving (2.29±0.40 fixations/s) compared with motorway driving (2.02±0.41 fixations/s; $t = -6.847$, $p < 10^{-6}$). These results align with prior findings on increased visual sampling and cognitive load in urban environments [261–263].

Figure 5.6 further illustrate blink and fixation duration distributions. Blink durations were marginally longer on the motorway (367.38±82.68 ms) than in the city (359.99±82.46 ms; $t = 4.494$, $p = 7 \times 10^{-6}$). Fixation durations exhibited a more pronounced difference, with longer fixations on the motorway (390.84±385.23 ms) compared with city driving (343.57±331.27 ms; $t = 15.124$, $p < 10^{-6}$). These patterns reflect sustained forward gaze during motorway driving and more frequent visual scanning in urban environments.

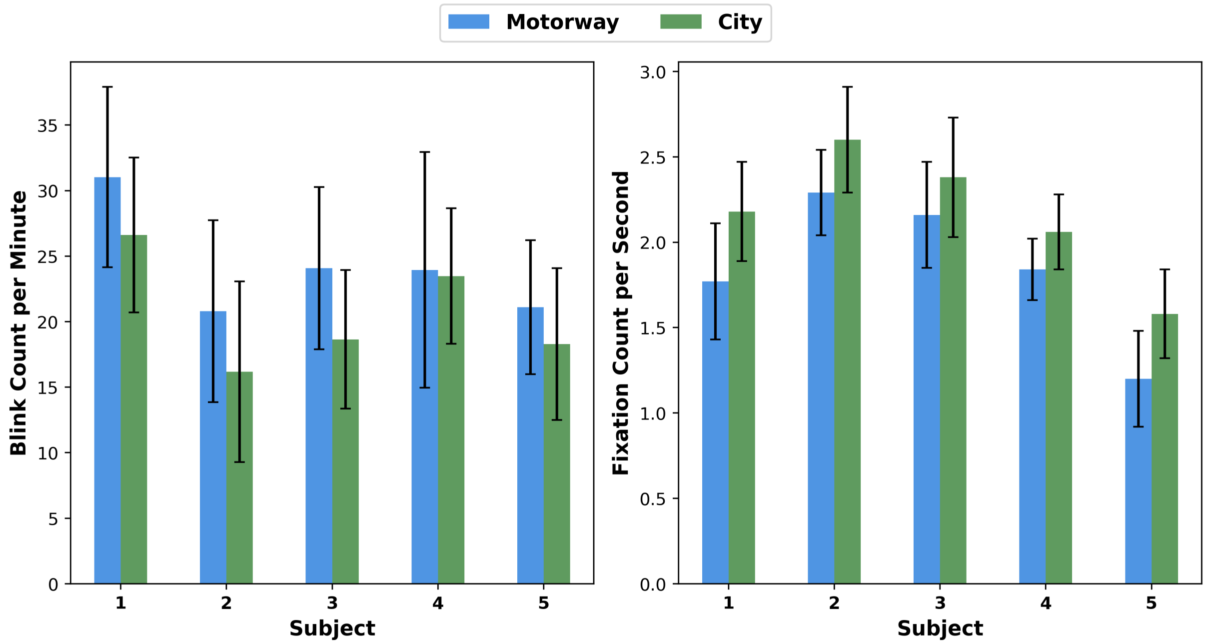


Figure 5.5: Blink and Fixation rates for all subjects across motorway and city driving.

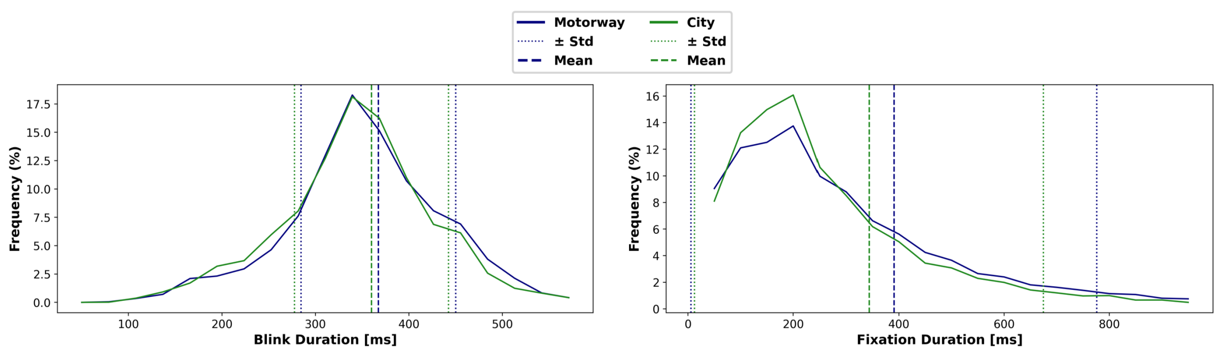


Figure 5.6: Blink and Fixation duration distributions for motorway and city driving.

Saccadic Movement Characteristics Across Scenarios

Saccade metrics are summarised in Figure 5.7. Motorway driving was associated with longer saccade durations (99.60±108.93 ms) and larger amplitudes (11.18±11.34°) compared with city driving (87.14±108.92 ms and 10.46±9.71°, respectively). In contrast, saccade velocity was higher in city environments (2403.99±1016.49 px/s) than on the motorway (2335.76±1064.29 px/s). All observed differences were statistically significant ($p < 10^{-6}$), indicating distinct gaze dynamics driven by the visual and attentional requirements of each scenario.

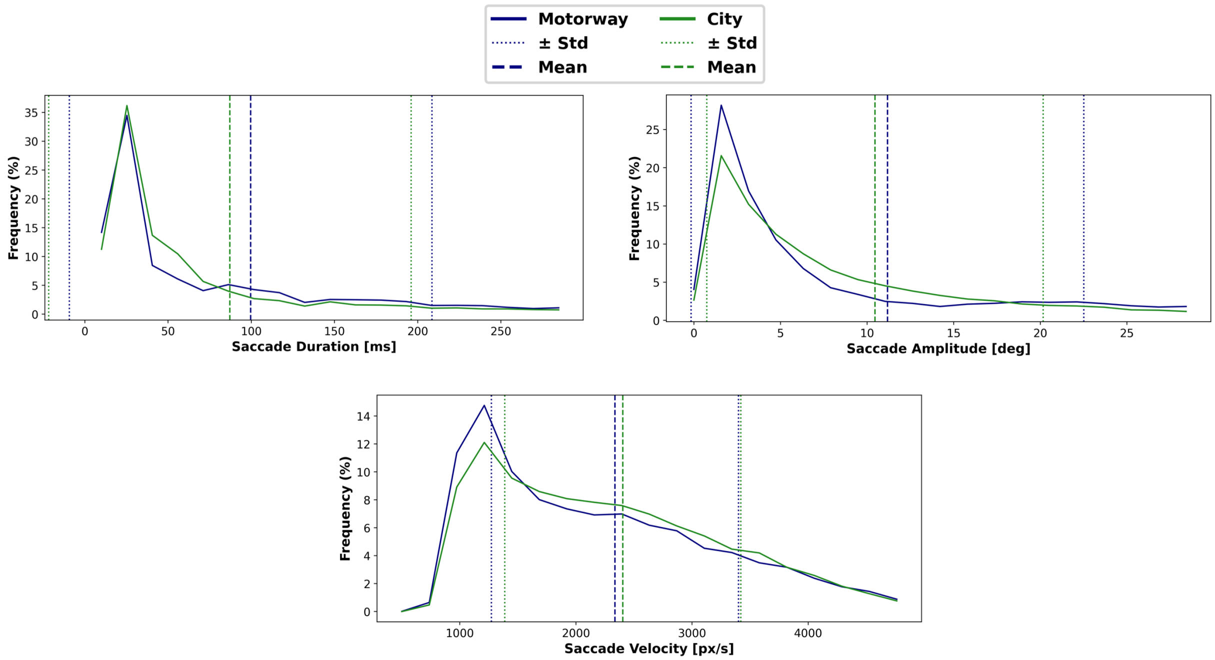


Figure 5.7: Saccade duration, amplitude, and velocity distributions for motorway and city driving.

Interpretation of Scenario Effects and Subject Variability

Taken together, the results demonstrate that both physiological and gaze-derived behavioural metrics exhibit consistent scenario-dependent structure. City driving is associated with higher HR and higher fixation rates, alongside faster saccades, indicating increased physiological arousal and intensified visual sampling under dynamic traffic and complex scene demands. Conversely, motorway driving is associated with higher blink rates and longer fixation and saccade durations, reflecting more sustained monitoring of the forward roadway with less frequent but longer visual transitions.

The gaze results should be interpreted as behavioural indicators rather than direct ground-truth labels of driver state. Higher fixation and saccade rates in city driving may reflect increased visual scanning, whereas longer or less frequent gaze events on the motorway may reflect more stable visual demand. These patterns support the use of multimodal fusion, since HR and gaze features respond to different aspects of driving demand. The co-movement of the HR and gaze indicators across scenarios, as shown in Figures 5.4, 5.5, 5.6, and 5.7, further supports the premise that fusing the two modalities yields a more reliable state estimate than either modality alone.

The subject-level patterns further indicate that individual differences, including driving experience and environmental conditions, can modulate both HR and gaze behaviour. These observations establish a baseline of multimodal variability across scenarios and participants and directly motivate the need for decision-level integration of HR and gaze metrics to improve driver-state assessment.

5.5.2 Algorithm Behaviour and Sensitivity Analysis

Algorithm Configuration and Parameter Selection

The multimodal driver monitoring algorithm introduced in Section 5.4.2 was evaluated using the collected real-world dataset to examine its internal consistency and practical behaviour under realistic driving conditions. Because segment labels were generated using heuristic thresholds rather than independent ground-truth annotations, the purpose of this evaluation is not to measure classification accuracy but to verify that the algorithm produces stable, interpretable outputs when integrating physiological (HR) and behavioural (gaze) metrics. The analysis, therefore, focuses on the consistency of decisions, sensitivity to parameter choices, and qualitative agreement with observed driving conditions, while maintaining computational simplicity suitable for real-time deployment.

Threshold values for each metric were not arbitrarily selected, but instead derived through a combination of (i) scenario- and subject-level statistical analysis presented in Section 5.5.1, and (ii) established normative ranges reported in the eye-tracking and driving literature. This dual grounding ensures that the selected thresholds are both empirically supported by the collected dataset and consistent with previously validated physiological and behavioural limits.

Normal driving segments were defined as those in which HR and gaze-derived metrics fell within well-established ranges reported in prior studies [264, 265]. For example, blink duration thresholds were set to 100–400 ms, a range widely reported as physiologically plausible during attentive driving [264]. This choice was further supported by the empirical distribution observed in this study, where blink durations exhibited a mean of 367.38 ms with a standard deviation of 82.68 ms during motorway driving (Figure 5.6), indicating that the selected bounds adequately capture typical behaviour while excluding extreme outliers.

Similarly, the remaining thresholds were defined by jointly considering literature-based norms and the statistical distributions observed in the dataset. Mean HR values between 65–95 BPM were classified as normal, reflecting typical resting-to-moderate workload levels during driving [208, 214]. Fixation durations between 150–900 ms were adopted based on established gaze behaviour during driving tasks [264]. Saccade-related thresholds, including duration (0–100 ms), amplitude ($0°$–$15°$), and velocity (0–3000 px/s), were selected to encompass typical oculomotor behaviour while excluding extreme gaze shifts unlikely to reflect normal scanning [265]. Segments in which one or more metrics exceeded these bounds were classified as abnormal.

Beyond thresholding, the relative contribution of each modality to the final decision was controlled through a weighted scoring scheme. HR was assigned the highest weight ($w_{\mathrm{HR}} = 0.5$) due to its strong and consistent differentiation between city and motorway driving, as evidenced by the statistically significant scenario effect reported in Section 5.5.1 and the distinct HR distributions shown in Figures. 5.3 and 5.4. These findings indicate that HR is a sensitive indicator of physiological response to varying driving demands.

Gaze metrics were assigned lower, yet complementary, weights (blink: 0.2, fixation: 0.2, saccade: 0.1). This reflects the more subtle and partially overlapping scenario-dependent differences observed in blink, fixation, and saccade characteristics (Figures. 5.5, 5.6, and 5.7). While gaze metrics alone may not uniquely identify internal driver state, their integration with HR provides valuable contextual information that improves robustness.

In addition to statistical analysis, approximately 5% of the segments were manually reviewed using synchronised dashcam footage to qualitatively examine the consistency of the algorithm outputs with observable driving conditions. This manual inspection was not used as ground-truth labelling, but rather as a limited sanity check to verify that segments classified as abnormal generally corresponded to situations involving increased driving workload, such as traffic stops, congestion, or complex urban manoeuvres, as illustrated in Figure 5.9. These observations supported the interpretability of the algorithm decisions, but do not constitute a formal accuracy evaluation.

Table 5.3 summarises the threshold ranges and metric weights used in the multimodal algorithm. While these values provide a statistically justified and interpretable baseline, they are not intended to be optimal or universal. Instead, they establish a transparent reference configuration that can be refined in future work using larger datasets, personalised calibration strategies, or learning-based optimisation methods.

Table 5.3: Threshold ranges and weights used in the multimodal driver monitoring algorithm.

Threshold Values		
Metric	**Lower Bound**	**Upper Bound**
HR	65 BPM	95 BPM
Blink duration [264]	100 ms	400 ms
Fixation duration [264]	150 ms	900 ms
Saccade duration	0 ms	100 ms
Saccade amplitude [265]	0°	15°
Saccade velocity	0 px/s	3000 px/s
Metric Weights		
HR	0.5	
Blink	0.2	
Fixation	0.2	
Saccade	0.1	

Segment-level labelling as normal or abnormal was performed automatically using the proposed multimodal algorithm (Algorithm 2) with the thresholds and weights described above. The limited manual review served only as a qualitative consistency check and did not influence algorithmic decisions.

Multimodal Classification Outcomes

The outputs produced by the proposed multimodal algorithm across all subjects and driving scenarios are summarised in Figure 5.8. Because segment labels were generated using heuristic thresholds without independent ground-truth annotations, these results should be interpreted as an internal consistency analysis rather than a performance evaluation. From a total of 2,937 non-overlapping segments of 10 s duration, 89.38% were classified as *normal*, while 10.62% were classified as *abnormal*. A clear scenario-dependent pattern is evident, with a higher proportion of abnormal segments occurring during city driving. This observation is consistent with the elevated HR levels and increased visual and cognitive demands identified in the scenario-level analysis presented in Section 5.5.1.

Subject-specific outcomes further illustrate the algorithm's sensitivity to individual differences. Subject 3 exhibited the highest proportion of abnormal segments, with 24.73% classified as abnormal during city driving and 20.16% on the motorway. This subject also showed the highest mean HR values and drove under warm conditions without air conditioning, factors that plausibly contributed to heightened physiological responses. In contrast, Subject 5, who had more than 20 years of driving experience and drove an automatic vehicle, showed minimal abnormal classifications, with only 1.65% of city segments flagged and almost none on the motorway. These subject-level trends closely mirror the physiological patterns discussed in Section 5.5.1, reinforcing that the algorithm responds to meaningful variations in HR and gaze behaviour rather than random fluctuations.

To further illustrate algorithm behaviour at the temporal level, representative real-time outputs are shown in Figures 5.9 and 5.10. Figure 5.9 depicts a typical city driving sequence in which HR decreases during traffic stops and gaze metrics stabilise, leading to segments being classified as normal. Conversely, transient HR increases during complex traffic situations result in short periods of abnormal classification. Figure 5.10 shows a motorway driving example characterised by generally stable HR and gaze behaviour, with brief HR elevations producing isolated abnormal classifications. Together, these examples demonstrate that the algorithm is responsive to short-term physiological changes while maintaining stability during steady driving conditions.

Sensitivity Analysis and Parameter Optimisation

A sensitivity analysis was conducted to quantify the impact of key algorithm parameters on classification behaviour and to assess the robustness of the proposed multimodal fusion strategy. In particular, the analysis examined the influence of (i) the relative weighting assigned to

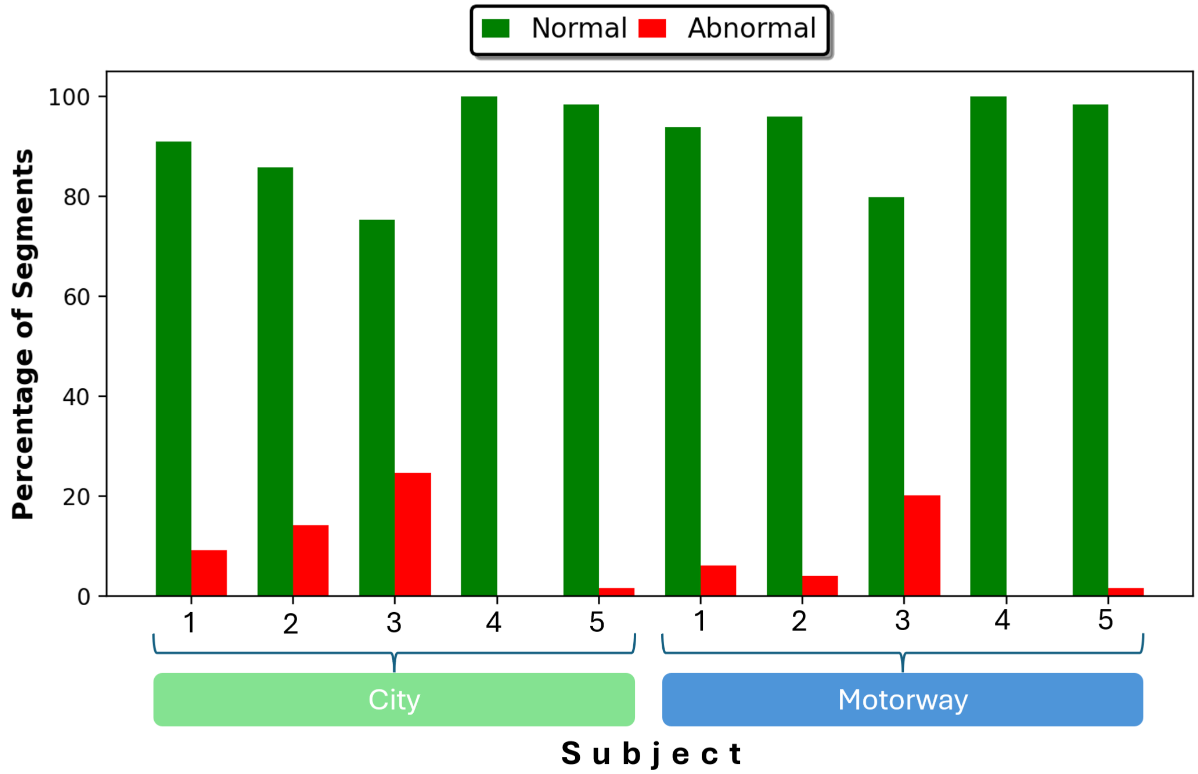


Figure 5.8: Percentage of segments classified as normal and abnormal for each subject across driving scenarios, illustrating inter-subject and inter-scenario variability under the defined thresholds and weights.

HR and gaze-derived metrics and (ii) the decision threshold θ used to distinguish between normal and abnormal segments. Because segment labels were generated using heuristic thresholds rather than independent ground-truth annotations, this analysis does not aim to determine optimal parameter values, but instead to examine how sensitive the algorithm outputs are to different parameter choices. The results are summarised in Tables 5.4 and 5.5.

Table 5.4 shows the proportion of segments classified as normal for different allocations of HR and gaze weights while keeping the decision threshold fixed at $\theta = 0.5$. The results demonstrate that classification outcomes are highly sensitive to the balance between physiological and behavioural contributions. When gaze metrics dominate the fusion process (e.g., HR weight = 0.1, gaze weight = 0.9), more than 92% of segments are classified as normal. As HR increases, the proportion of normal segments decreases progressively. Notably, increasing the HR weight to 0.6 or above results in a sharp reduction in normal classifications to approximately 64.79%. This abrupt change is caused by the interaction between the weight allocation and the fixed decision threshold $\theta = 0.5$. When the HR weight exceeds the threshold, the HR component alone can dominate the decision score, meaning that segments classified as abnormal based on HR are no longer compensated by gaze metrics. As a result, the classification behaviour changes nonlinearly rather than gradually as the weight increases. This behaviour is consistent with the strong scenario-dependent separation observed in HR distributions (Section 5.5.1), and indicates

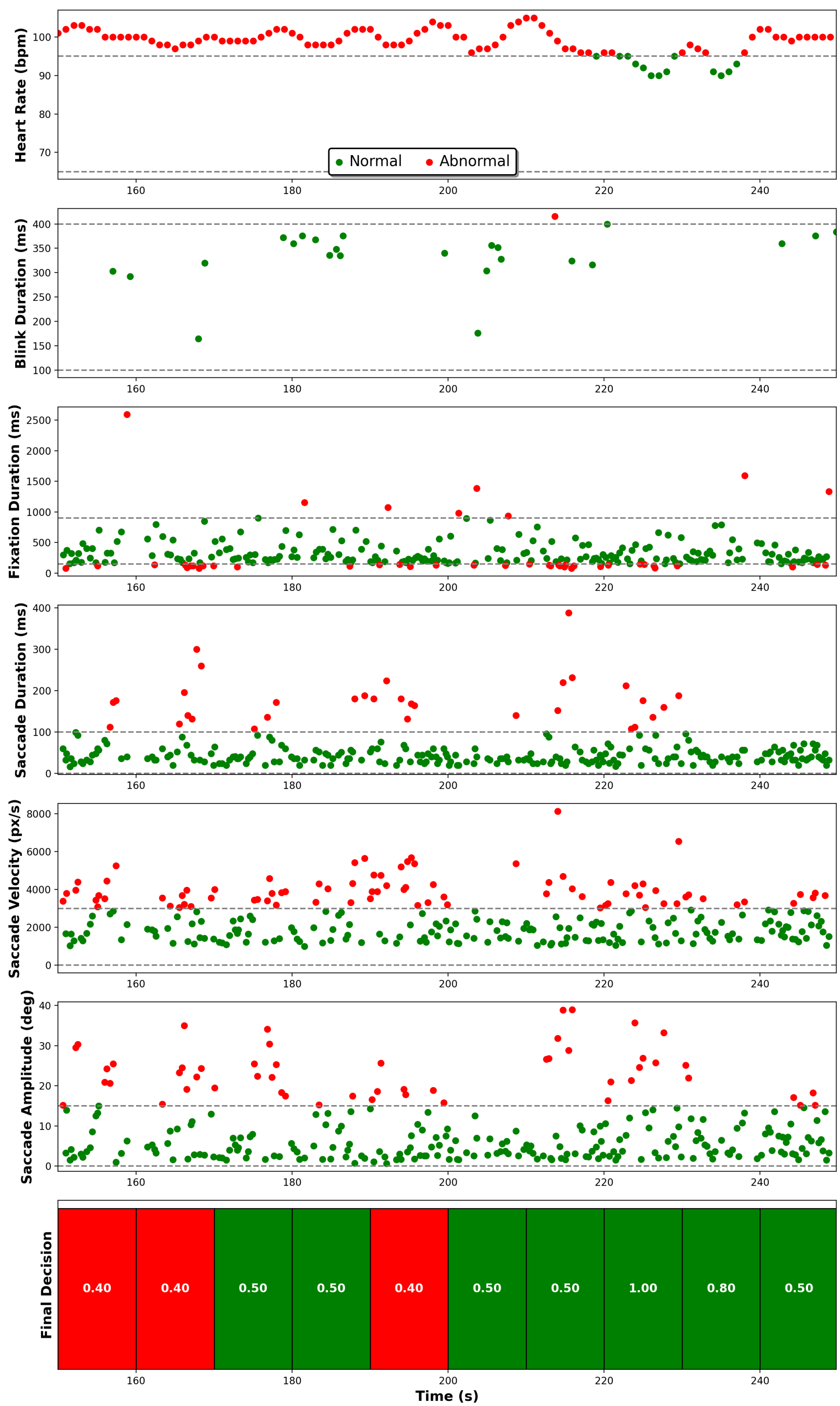


Figure 5.9: Real-time multimodal classification during city driving, illustrating HR stabilisation during traffic stops and corresponding segment-level decisions based on the threshold θ.

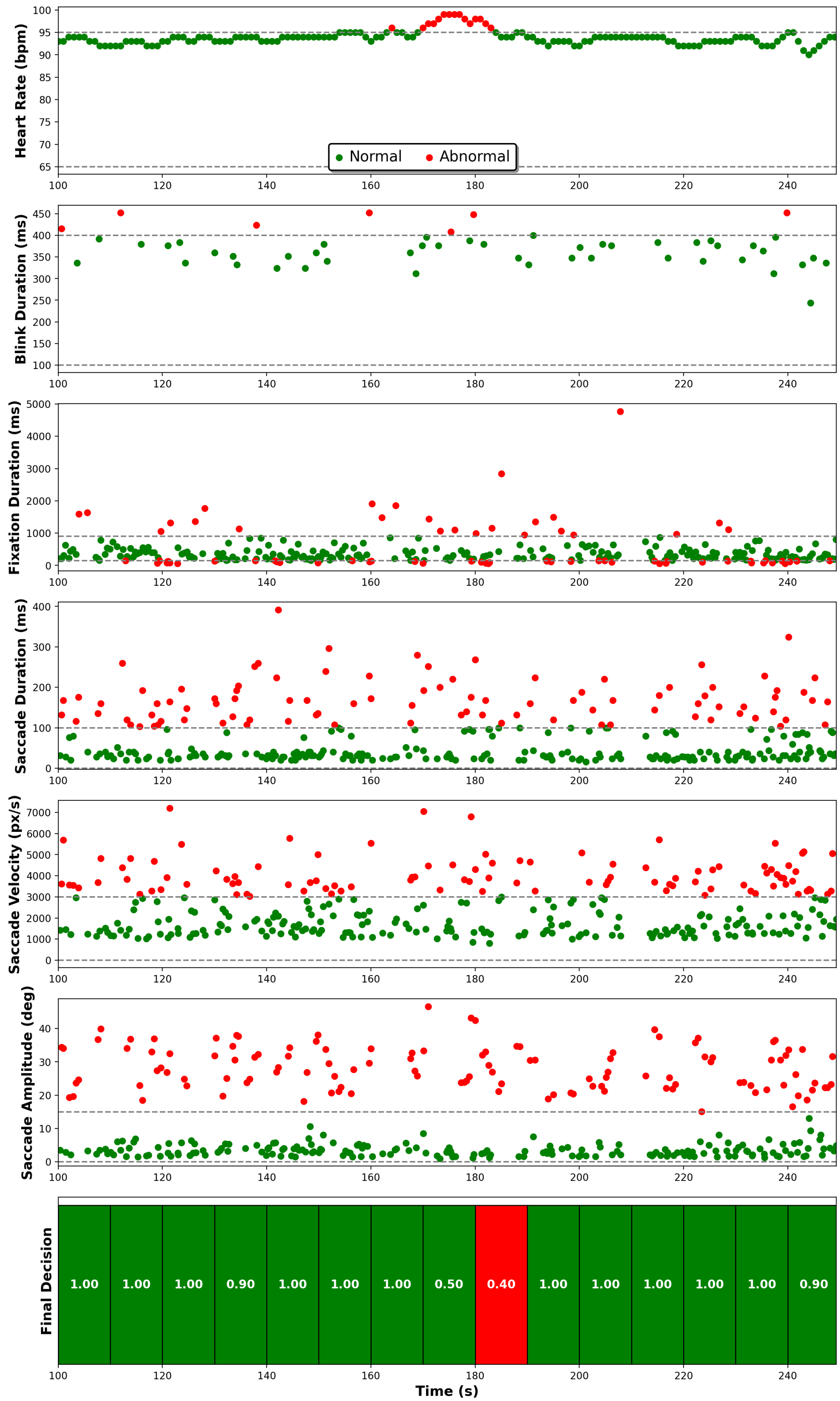


Figure 5.10: Real-time multimodal classification during motorway driving, showing predominantly stable HR with brief increases leading to transient abnormal classifications.

that HR has a strong influence on the final decision within the current fusion configuration.

Table 5.4: Effect of HR and gaze weight allocation on the percentage of segments classified as normal ($\theta = 0.5$).

HR Weight	Gaze Weight	Normal Segments (%)
0.1	0.9	92.32
0.2	0.8	91.75
0.3	0.7	91.75
0.4	0.6	86.32
0.5	0.5	87.38
0.6	0.4	64.79
0.7	0.3	64.79
0.8	0.2	64.79
0.9	0.1	64.79

The sensitivity of the algorithm to the decision threshold θ is illustrated in Table 5.5. As θ increases from 0.30 to 0.80, the proportion of segments classified as normal decreases from 97.34% to 59.76%. Lower thresholds favour normal classifications by requiring fewer modalities to agree, whereas higher thresholds enforce stricter agreement across HR and gaze metrics before a segment is labelled as normal. This trade-off illustrates how the algorithm's behaviour changes under different parameter settings, rather than indicating that any single threshold value is optimal.

Table 5.5: Effect of decision threshold θ on the percentage of segments classified as normal.

Threshold (θ)	Normal Segments (%)
0.30	97.34
0.35	93.26
0.40	93.26
0.45	87.38
0.50	87.38
0.55	63.73
0.60	63.73
0.65	63.28
0.70	63.28
0.75	59.76
0.80	59.76

The sensitivity analysis confirms that both weight assignment and threshold selection have a substantial impact on algorithm behaviour. Since the dataset lacks independent ground-truth annotations for driver state, the results cannot be used to identify an optimal parameter configuration. Instead, the analysis highlights parameter regions where the algorithm's behaviour is relatively stable and less sensitive to threshold variation, which is desirable for real-time monitoring applications.

The threshold θ should be interpreted as an application-dependent operating parameter rather than a fixed or universal value. Lower θ values generate more normal classifications and may be appropriate for exploratory monitoring or applications where false alarms must be minimised, whereas higher θ values require stronger agreement across modalities before a segment is considered normal, increasing sensitivity to abnormal states at the expense of a higher alert rate; this is preferable in safety-critical screening where missing an abnormal-state segment is costly. The flat regions identified in Table 5.5 (for example, $\theta \in [0.35, 0.50]$ and $[0.55, 0.70]$) indicate parameter bands over which behaviour is stable, which is desirable for real-time operation.

The configuration ($w_{\text{HR}} = 0.5$, $\theta = 0.5$) was therefore selected as a balanced, stable operating point for this pilot study rather than a globally optimal solution. In the absence of independent ground-truth labels, $\theta = 0.5$ is adopted only as a balanced demonstration point and is the value used to generate the reported results. These findings motivate future extensions toward adaptive or data-driven parameter tuning strategies, where θ could be tuned per application by optimising against a target sensitivity/false-alarm operating point using externally labelled, event-annotated datasets.

5.6 Summary

This chapter investigated multimodal human-state monitoring by integrating physiological and behavioural metrics, focusing on HR and gaze-derived indicators collected during real-world driving. A key strength of this study was the use of naturalistic on-road recordings rather than simulator-based experiments. The dataset comprised approximately 480 minutes of real driving across multiple sessions, vehicles, traffic conditions, and environmental settings, providing a rare corpus of synchronised physiological and behavioural measurements captured under realistic operating conditions. This allowed the analysis to reflect genuine driver responses rather than simulator-induced behaviour, which was common in prior studies.

Statistical analysis revealed clear and consistent differences in both HR and gaze behaviour across driving scenarios and subjects. City driving was associated with elevated HR, higher fixation rates, and faster saccadic movements, reflecting increased cognitive and visual demands, while motorway driving exhibited higher blink rates and longer fixation and saccade durations, indicative of sustained attention and reduced environmental complexity. Substantial inter-subject variability was also observed, demonstrating that driver state cannot be reliably characterised using a single modality alone and motivating the need for multimodal fusion.

Building on these observations, a lightweight real-time multimodal monitoring algorithm was proposed and analysed. The algorithm integrates HR and gaze metrics using statistically informed thresholds and a weighted decision-level fusion strategy. The resulting outputs were consistent with the scenario-level statistical trends and with qualitative inspection of selected driving segments, indicating that the multimodal approach can produce stable and interpretable state estimates under realistic conditions. Because segment labels were generated using heuristic thresholds rather than independent ground-truth annotations, the evaluation was presented as an internal consistency validation rather than a formal accuracy assessment.

These results demonstrate that combining behavioural gaze indicators with physiological HR measurements provides a more informative representation of driver state than unimodal sensing, while maintaining computational simplicity suitable for real-time deployment. The study establishes a transparent baseline for multimodal monitoring in naturalistic environments and highlights the need for larger cohorts and external ground-truth annotations to enable fully generalisable human-state monitoring systems.

Chapter 6

Conclusion and Future Work

6.1 Summary of Contributions

This thesis explored how FMCW radar sensing, combined with modern AI and signal processing, can support non-invasive human monitoring across behavioural and physiological scales. Section 1.3 defined three challenges: **C1**, multi-scale sensing from large-scale activities to physiological micro-motions; **C2**, limited and heterogeneous radar datasets for learning-based generalisation; and **C3**, the need for interpretable multimodal frameworks that integrate behavioural and physiological cues in realistic environments. Chapters 3–5 addressed these challenges at complementary levels.

Addressing C1: For large-scale monitoring (**C1.1**), Chapter 3 developed and evaluated a complete FMCW radar-based pipeline for six everyday activities (walking, sitting, standing, bending to pick an object, drinking, and falling). Radar returns were represented using three complementary two-dimensional inputs: TR maps (range FFT), TF maps derived using STFT, and higher-resolution TF maps derived using SPWVD. These representations were systematically paired with four well-known CNN architectures (VGG-16, VGG-19, ResNet-50, and MobileNetV2), yielding twelve model–representation pairs. Under subject-wise evaluation, the resulting analysis quantified recognition performance alongside computational cost, including preprocessing time, training time, and inference latency. A key finding was that SPWVD maps, despite achieving high recognition accuracy, incurred an impractical preprocessing cost for real-time use: generating SPWVD TF maps required 52.58 s per sample, compared with 0.22 s for STFT maps and 0.035 s for TR maps. Consequently, the final design choice prioritised balanced end-to-end efficiency rather than accuracy alone. The MobileNetV2 + STFT configuration (pair M8) provided the most favourable trade-off across preprocessing, training time per epoch, and inference time, achieving around 96.30% accuracy with inference latency of only a few milliseconds per sample, supporting practical real-time deployment constraints for **C1.1**.

For fine-grained monitoring (**C1.2**), the first part of Chapter 4 targeted driver-relevant seated upper-body and head motions. Using a single mmWave FMCW radar observing the upper body, a dataset was collected from 15 subjects performing six kinematically similar head-and-torso movements. A tailored pipeline transformed raw radar data into TF maps using STFT. Given the limited dataset size, the learning component adopted transfer learning by fine-tuning a lightweight MobileNetV2 backbone; however, transfer learning alone was insufficient to achieve robust subject-wise generalisation for this fine-grained task under pronounced data scarcity. To reduce overfitting and improve separability, Chapter 4 therefore introduced a radar-specific adaptation of SpecAugment, applying time- and frequency-domain masking to TF maps to generate physically plausible training variability. The augmentation factor was systematically varied, with the best performance achieved at a factor of 10 (effectively expanding the training set tenfold); beyond this point, additional augmentation produced diminishing returns and slightly reduced accuracy. Under subject-wise evaluation, accuracy increased from 60% without augmentation to 91.32% with the proposed augmentation strategy, indicating that TF-consistent augmentation, combined with transfer learning, can materially improve discrimination of subtle movements central to **C1.2**.

For physiological monitoring (**C1.3**), the second part of Chapter 4 addressed motion artefacts in radar-based vital sign estimation. Using an existing paediatric FMCW dataset, VMD was integrated into the processing chain as an adaptive decomposition tool. After conventional FMCW processing produced range-gated respiratory and cardiac signals, VMD decomposed the signals into intrinsic mode functions, enabling suppression of motion-dominated components while retaining HR- and RR-related components. Compared with a reference sensor, the VMD-enhanced pipeline reduced MAE and RMSE by an average of more than 40% across HR and RR in recordings containing motion. This demonstrates that decomposition-based processing can improve robustness beyond band-limited filtering alone, directly supporting **C1.3**.

Addressing C2: To mitigate radar data scarcity (**C2**), this thesis combined transfer learning and targeted augmentation. Chapter 3 leveraged transfer learning by fine-tuning pre-trained CNN backbones on a public 5.8 GHz FMCW dataset and comparing performance across TR and TF representations, while reporting recognition metrics together with preprocessing time, training time, and inference latency. Chapter 4 complemented this with SpecAugment-inspired TF masking for fine-grained monitoring, expanding the effective training distribution without additional data collection. Together, these strategies provide practical, data-efficient mechanisms for learning-based radar monitoring under limited labelled data.

Addressing C3: Chapter 5 addressed **C3** by evaluating multimodal fusion of behavioural and physiological cues in real driving, as an application of human monitoring. A synchronised dataset was collected from five licensed drivers over approximately 480 minutes of mixed urban and motorway routes, using a Polar H10 chest strap for HR and Pupil Labs Invisible eye-tracking glasses for gaze features. Scenario-based and subject-based analyses quantified how HR and gaze metrics (blink, fixation, and saccade measures) vary across driving contexts. Building on these findings, a weighted fusion algorithm combined HR and gaze features to produce a continuous alertness score, with tunable decision thresholds and weights. This provides a proof-of-concept pathway for real-time, multimodal driver-state monitoring grounded in behavioural and physiological evidence.

Across Chapters 3–5, the thesis presented a staged progression from radar-only monitoring to multimodal state assessment in a realistic environment. Chapter 3 defined a data-efficient and computationally aware design space for large-scale HAM using FMCW radar. Chapter 4 extended radar monitoring to fine-grained movements and motion-robust physiological estimation, using TF-consistent augmentation and VMD-based artefact mitigation. Chapter 5 then demonstrated how behavioural and physiological indicators can be fused into an interpretable human state metric, aligning radar sensing capabilities with the multimodal requirement of **C3**.

Achievement of Research Aims The research aims stated in Chapter 1 were achieved through the combined findings of Chapters 3–5.

Aim 1 sought to investigate whether a data- and computationally-efficient AI-based radar sensing framework could interpret the three scales of human motion, from large-scale activities to fine-grained movements, while also supporting reliable physiological monitoring under motion-induced corruption. This aim was achieved through the results of Chapters 3 and 4. Chapter 3 demonstrated that large-scale activities can be recognised accurately and efficiently using FMCW radar with appropriate representation–model pairings, while Chapter 4 extended this capability to fine-grained upper-body motion and motion-robust physiological monitoring through data augmentation and VMD-based signal decomposition.

Aim 2 aimed to evaluate data-efficient AI methodologies for radar-based human monitoring, particularly under limited and specialised datasets. This aim was achieved by demonstrating that transfer learning and targeted augmentation provide effective mechanisms for improving generalisation in the presence of radar data scarcity. In Chapter 3, transfer learning enabled robust classification across multiple radar representations and CNN backbones, while in Chapter 4, radar-specific SpecAugment-inspired augmentation substantially improved subject-wise recognition performance for subtle fine-grained activities.

Aim 3 aimed to investigate multimodal approaches for comprehensive human state assessment by fusing behavioural and physiological signals into an interpretable decision framework. This aim was achieved in Chapter 5, where HR and gaze features were analysed in real driving conditions and integrated into a weighted fusion framework that produced a continuous alertness score for real-time driver monitoring.

Taken together, these findings show that the thesis met its three research aims while also resolving the broader challenge structure defined by **C1**, **C2**, and **C3**.

6.2 Limitations and Future Work

Although this thesis provided coherent contributions across large-scale HAM, fine-grained monitoring, vital sign estimation, and multimodal human monitoring, several limitations remained. These motivate concrete future research directions aligned with residual aspects of **C1–C3**.

6.2.1 Limitations

Scope of radar hardware and deployment A deliberate choice in the radar-based studies was to use a single transmit and single receive configuration in each experiment. This simplifies processing and reduces data dimensionality, but limits spatial discrimination. Without multiple channels, the system cannot exploit angle-of-arrival information to better localise targets, separate overlapping movers, or disambiguate posture and chest motion in cluttered scenes. In addition, radar experiments were conducted in single-subject scenarios, which limits direct generalisation to multi-person settings.

Large-scale monitoring Chapter 3 evaluated a single public dataset collected under controlled indoor conditions using a 5.8 GHz SISO FMCW radar and a six-class activity taxonomy. While subject-wise evaluation was applied, cross-environment and cross-device shifts (placement, room layout, hardware variation) were not studied. Timing metrics support computational conclusions, but energy consumption was not measured on embedded hardware. The study also focused on TR and TF representations derived from a single range-gating strategy and did not investigate learning directly from the raw radar cube, model quantisation, pruning, or on-device benchmarking.

Fine-grained monitoring The fine-grained study in Chapter 4 demonstrated strong subject-wise performance for six seated upper-body and head movements. Still, data were collected in an indoor setup that emulated a driving posture rather than in a moving vehicle. Consequently, the model was not exposed to in-cabin vibration, road-induced motion, or vehicle-specific clutter. The dataset size and the number of activity classes remain limited, and radar representation maps were derived from a single effective channel, discarding angle cues that could improve separability among confusable classes, such as leaning left or right. The approach has not yet been evaluated in multi-target or multi-occupant scenarios, and no real-time embedded implementation has been performed.

Motion-robust vital sign estimation The VMD-based vital sign pipeline was validated on a paediatric dataset and applied after the radar conventional vital sign estimation pipeline rather than directly to raw IF data. This limits conclusions about information preservation across the full processing chain. The dataset was collected from children in an indoor setting, limiting generalisation to adult or dynamic scenarios. While the proposed approach clearly improves robustness to motion, it remains an offline approach, and its real-time feasibility on low-power hardware remains open.

Multimodal HR–gaze monitoring Chapter 5 used a considerable amount of real-world driving data, but the cohort size was small and demographically narrow. All participants in the study were male, which limits the generalisability of the findings. In addition, the fusion algorithm used manually selected weights and thresholds, and driving segments were labelled using heuristic criteria rather than external ground-truth annotations for distraction or fatigue. As a result, the evaluation primarily demonstrates internal consistency of the proposed framework rather than classification accuracy. A more rigorous assessment would require externally validated labels, such as fatigue events, PERCLOS-based eye-closure measures, or expert annotation of driver state.

Furthermore, radar was not utilised as a non-intrusive modality in this multimodal study, indicating that the framework does not yet constitute a fully non-contact, user-compliant monitoring solution. Future work should therefore integrate radar as a physiological sensing modality within the behaviour–physiology fusion framework.

More broadly, the thesis established effective solutions at large-scale, fine-grained, and physiological scales, but did not yet deliver a single multi-scale radar front end capable of operating across all three simultaneously. Dataset scarcity remains a field-wide bottleneck, particularly for multimodal radar datasets that include behavioural and physiological signals in realistic environments. Finally, multimodal fusion was validated only for HR–gaze, and both personalisation and radar integration remain limited.

6.2.2 Future Work

Future research can build on these contributions along several connected directions that prioritise deployment realism, robustness to domain shift, and multimodal integration.

Energy-aware and edge-optimised system Chapter 3 quantified preprocessing and inference latency but did not measure energy consumption. Future work should deploy the most promising model–representation pairs (for example, MobileNetV2 with STFT-based TF maps) on embedded platforms and report power–latency–accuracy trade-offs using direct measurements. Practical extensions include post-training quantisation, pruning, and hardware-aware optimisation, with evaluation based on energy per inference and energy per correct decision under subject-wise and cross-environment protocols.

Adaptive pre-processing and multi-scale representations Future work should explore learnable front ends that operate closer to the radar cube and jointly optimise range selection, temporal filtering, and TF feature extraction. For motion-robust vital signs, applying VMD and related decomposition directly to raw complex radar channels may improve separation between motion artefacts and physiological components. A unified objective is to produce a single radar front end that yields informative features for **C1.1–C1.3** without separate, task-specific pipelines.

From SISO to MIMO for spatial disambiguation A transition to multi-channel FMCW radars can enable spatial discrimination that supports posture, limb dynamics, head orientation, and chest motion within the same sensing volume. Future work should quantify the extent to which angular information improves separability for confusable fine-grained classes and how MIMO processing affects computational budgets. Efficient strategies include dimensionality reduction and hierarchical pipelines that trigger higher-cost spatial processing only when needed.

Richer datasets and realistic evaluation protocols Future work should develop and release larger datasets that utilise only radar to capture physiological and behavioural metrics across more diverse scenarios, including real in-vehicle collection with broader driver demographics, vehicle types, and secondary-task conditions. Benchmarking should include subject-wise, session-wise, and cross-environment splits, enabling credible assessment of generalisation.

Privacy-preserving deployment Although the experimental studies in this thesis were conducted with consented, opt-in participants, the proposed methodology should be understood within a broader privacy-, safety-, and security-aware deployment framework. Radar sensing is comparatively privacy-preserving relative to camera-based monitoring because it captures motion and physiological signatures rather than directly identifiable imagery. However, radar data may still reveal sensitive behavioural or health-related information, and real-world deployment should therefore include explicit informed consent, a clear right to opt out, on-device or edge processing where possible, and retention policies that prioritise derived state scores over raw radar data. These measures align the proposed approach with data-minimisation principles and reduce unnecessary exposure of participant information.

Security and safe operation deployment Beyond data governance, deployment in ambient-assisted living, clinical, and in-vehicle environments also requires appropriate regulatory and electromagnetic compatibility assessments. Radar devices should be operated within the relevant spectrum-use conditions and assessed for coexistence with nearby communication, electronic, and medical devices. In addition, security must be considered because radar-based monitoring can be vulnerable to intentional interference. For example, jamming could degrade the received signal and reduce monitoring reliability, while spoofing could introduce fabricated reflections that lead to false activity or vital-sign estimates. Robustness against such attacks is therefore identified as an important direction for future work. Together, privacy governance, regulatory compliance, EMC assessment, and security resilience are necessary prerequisites for deploying the proposed radar-based monitoring methods in the real world.

Personalisation and uncertainty-aware inference Inter-individual variability strongly affects radar signatures of physiological and movement baselines. Future work should investigate mechanisms for personalising thresholds and model parameters over time, using limited supervision or self-supervised adaptation. Uncertainty-aware decision making can further improve safety by enabling calibrated confidence and conservative behaviour when signals are degraded or contradictory.

Towards a unified, multi-scale, multimodal AI–radar framework Ultimately, the strands of work in this thesis can be unified into a single architecture in which a radar front-end produces a multi-scale representation of human activity and state. Large-scale activities, fine-grained gestures, and vital signs would be interpreted using a single radar, and the resulting inference would feed into a decision layer that estimates risk, abnormality, or clinical relevance. Realising this vision will require advances in radar hardware utilisation (including MIMO), multi-scale and multimodal feature learning, energy-efficient deployment, and rich datasets. Nevertheless, the results presented in Chapters 3 to 5 demonstrate that each of the fundamental components is feasible. The future directions outlined above chart a path from current prototype systems to robust, widely deployable human-centric monitoring solutions that fully exploit the potential of AI and radar for non-invasive, fine-grained human state monitoring.

Acknowledgement for the Use of AI Tools

I acknowledge the use of two language assistance tools, **ChatGPT** by OpenAI and **Grammarly**, in the proofreading of this thesis. These tools were employed solely to enhance the clarity, coherence, and academic tone of the written text. Their contributions were limited to improving language and flow, while the intellectual content and findings of this work remain entirely my own.

Bibliography

[1] Y. Yin, L. Xie, Z. Jiang, F. Xiao, J. Cao, and S. Lu, “A Systematic Review of Human Activity Recognition Based on Mobile Devices: Overview, Progress and Trends,” *IEEE Communications Surveys & Tutorials*, vol. 26, no. 2, pp. 890–929, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10413312

[2] World Health Organization, *World Report on Ageing and Health*. Geneva: World Health Organization, 2015.

[3] C. Ranieri, S. MacLeod, M. Dragone, P. Vargas, and R. Romero, “Activity recognition for ambient assisted living with videos, inertial units and ambient sensors,” *Sensors*, vol. 21, no. 3, p. 768, 2021.

[4] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, “Monitoring activities of daily living in smart homes: Understanding human behavior,” *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 81–94, March 2016.

[5] “Total world vehicles in use,” [Online] Available: https://www.oica.net/category/vehicles-in-use, accessed: 22 July 2024.

[6] National Highway Traffic Safety Administration, “Drowsy driving: Crashes, injuries, and fatalities,” U.S. Department of Transportation, Tech. Rep. NCSA Research Note, DOT HS 812 747, April 2019, data for 2017.

[7] ——, “Traffic safety facts research note: Distracted driving 2022,” U.S. Department of Transportation, Tech. Rep. DOT HS 813 559, April 2024.

[8] “Passive infrared (pir) detector data sheet,” https://epiuat.tunstall.com/resources/productdatasheets/pir/, accessed: November 2, 2025.

[9] M. Falter, W. Budts, K. Goetschalckx, V. Cornelissen, R. Buys *et al.*, “Accuracy of apple watch measurements for heart rate and energy expenditure in patients with cardiovascular disease: cross-sectional study,” *JMIR mHealth and uHealth*, vol. 7, no. 3, p. e11889, 2019.

[10] S. Seneviratne, Y. Hu, T. Nguyen, G. Lan, S. Khalifa, K. Thilakarathna, M. Hassan, and A. Seneviratne, "A Survey of Wearable Devices and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2573–2620, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7993011

[11] F. Luo, A. Li, S. Khan, K. Wu, and L. Wang, "Bi-DeepViT: Binarized Transformer for Efficient Sensor-Based Human Activity Recognition," *IEEE Transactions on Mobile Computing*, pp. 1–15, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10829799

[12] S. Otmani, T. Pebayle, J. Roge, and A. Muzet, "Effect of driving duration and partial sleep deprivation on subsequent alertness and performance of car drivers," *Physiology & behavior*, vol. 84, no. 5, pp. 715–724, 2005.

[13] Seeing Machines, "Automotive driver monitoring systems," https://seeingmachines.com/products/automotive/, 2025, accessed: November 2, 2025.

[14] Mobileye, "Presenting the mobileye driver monitoring system: Fusing road safety inside the cabin," https://www.mobileye.com/blog/presenting-the-mobileye-driver-monitoring-system-fusing-road-safety-inside-the-cabin/, 2024, accessed: November 2, 2025.

[15] M. Akin, M. B. Kurt, N. Sezgin, and M. Bayram, "Estimating vigilance level by using EEG and EMG signals," *Neural Computing and Applications*, vol. 17, no. 3, pp. 227–236, Jun. 2008. [Online]. Available: https://doi.org/10.1007/s00521-007-0117-7

[16] D. Bouchabou, S. M. Nguyen, C. Lohr, B. LeDuc, and I. Kanellos, "A survey of human activity recognition in smart homes based on iot sensors algorithms: Taxonomies, challenges, and opportunities with deep learning," *Sensors*, vol. 21, no. 18, p. 6037, 2021.

[17] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE journal of biomedical and health informatics*, vol. 17, no. 3, pp. 579–590, 2012.

[18] J. C. Lin, "Microwave sensing of physiological movement and volume change: A review," *Bioelectromagnetics*, vol. 13, no. 6, pp. 557–565, 1992. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/bem.2250130610

[19] S. Ahmed and S. H. Cho, "Machine Learning for Healthcare Radars: Recent Progresses in Human Vital Sign Measurement and Activity Recognition," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 1, pp. 461–495, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10322785

[20] K. Papadopoulos and M. Jelali, “A Comparative Study on Recent Progress of Machine Learning-Based Human Activity Recognition with Radar,” *Applied Sciences*, vol. 13, no. 23, p. 12728, Jan. 2023, number: 23 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2076-3417/13/23/12728

[21] A. Dahal, S. Biswas, S. Z. Gurbuz, and A. C. Gurbuz, “Comparison Between Wi-Fi-CSI and Radar-Based HAR,” in *2024 IEEE Radar Conference (RadarConf24)*, May 2024, pp. 1–6, iSSN: 2375-5318. [Online]. Available: https://ieeexplore.ieee.org/document/10548515

[22] S. Jardak, M.-S. Alouini, T. Kiuru, M. Metso, and S. Ahmed, “Compact mmwave fmcw radar: Implementation and performance analysis,” *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 2, pp. 36–44, 2019.

[23] L. Bergasa, J. Nuevo, M. Sotelo, R. Barea, and M. Lopez, “Real-time system for monitoring driver vigilance,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 1, pp. 63–77, Mar. 2006. [Online]. Available: https://ieeexplore.ieee.org/document/1603553

[24] S. Z. Gurbuz and M. G. Amin, “Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring,” *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 16–28, 2019.

[25] M. G. Amin, Y. D. Zhang, F. Ahmad, and K. D. Ho, “Radar Signal Processing for Elderly Fall Detection: The future for in-home monitoring,” *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 71–80, Mar. 2016.

[26] W. Hu, Z. Zhao, Y. Wang, H. Zhang, and F. Lin, “Noncontact accurate measurement of cardiopulmonary activity using a compact quadrature doppler radar sensor,” *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 725–735, 2014.

[27] L. Anitori, A. de Jong, and F. Nennie, “Fmcw radar for life-sign detection,” in *2009 IEEE Radar Conference*, 2009, pp. 1–6.

[28] A. E. Omer, S. Safavi-Naeini, R. Hughson, and G. Shaker, “Blood glucose level monitoring using an fmcw millimeter-wave radar sensor,” *Remote Sensing*, vol. 12, no. 3, p. 385, 2020.

[29] S. Ishizaka, K. Yamamoto, and T. Ohtsuki, “Non-contact blood pressure measurement using doppler radar based on waveform analysis by lstm,” in *ICC 2021 - IEEE International Conference on Communications*, 2021, pp. 1–6.

[30] M. Nasr, M. M. Islam, S. Shehata, F. Karray, and Y. Quintana, "Smart healthcare in the age of ai: Recent advances, challenges, and future prospects," *IEEE Access*, vol. 9, pp. 145 248–145 270, 2021.

[31] F. Ayaz, B. Alhumaily, S. Hussain, M. A. Imran, K. Arshad, K. Assaleh, and A. Zoha, "Radar Signal Processing and Its Impact on Deep Learning-Driven Human Activity Recognition," *Sensors*, vol. 25, no. 3, p. 724, Jan. 2025, number: 3 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/25/3/724

[32] S. Wang, L. Mei, R. Liu, W. Jiang, Z. Yin, X. Deng, and T. He, "Multi-Modal Fusion Sensing: A Comprehensive Review of Millimeter-Wave Radar and Its Integration With Other Modalities," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10525189

[33] F. Wang, X. Zeng, C. Wu, B. Wang, and K. J. R. Liu, "Driver Vital Signs Monitoring Using Millimeter Wave Radio," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 11 283–11 298, Jul. 2022.

[34] T. Zheng, Z. Chen, C. Cai, J. Luo, and X. Zhang, "V2ifi: In-vehicle vital sign monitoring via compact rf sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–27, 2020.

[35] J. Ludikhuize, S. M. Smorenburg, S. E. de Rooij, and E. de Jonge, "Identification of deteriorating patients on general wards; measurement of vital parameters and potential effectiveness of the modified early warning score," *Journal of critical care*, vol. 27, no. 4, pp. 424–e7, 2012.

[36] J. Chung, H. Lee, H. Moon, and E. Lee, "The Static and Dynamic Analyses of Drivers' Gaze Movement Using VR Driving Simulator," *Applied Sciences*, vol. 12, no. 5, p. 2362, Jan. 2022, number: 5 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2076-3417/12/5/2362

[37] D. Krauss, L. Engel, T. Ott, J. Bräunig, R. Richer, M. Gambietz, N. Albrecht, E. M. Hille, I. Ullmann, M. Braun, P. Dabrock, A. Kölpin, A. D. Koelewijn, B. M. Eskofier, and M. Vossiek, "A Review and Tutorial on Machine Learning-Enabled Radar-Based Biomedical Monitoring," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 5, pp. 680–699, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10520876

[38] X. Tao, D. Gao, W. Zhang, T. Liu, B. Du, S. Zhang, and Y. Qin, "A multimodal physiological dataset for driving behaviour analysis," *Scientific data*, vol. 11, no. 1, p. 378, 2024.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, 2012.

[40] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[41] "Polar h10 chest strap," https://support.polar.com/e_manuals/h10-heart-rate-sensor/polar-h10-user-manual-english/manual.pdf.

[42] "Pupil invisible eye tracking glasses," [Online] Available: https://docs.pupil-labs.com/invisible/, accessed: 20 November 2024.

[43] H. J. Keller, "30 years of passive infrared motion detectors-a technology review," in *Int. Conf. OPTO/IRS2/MTT*, 2000, pp. 1–6.

[44] J. DuCarme, "Developing effective proximity detection systems for underground coal mines," in *Advances in productive, safe, and responsible coal mining*. Elsevier, 2019, pp. 101–119.

[45] J. Yun and S.-S. Lee, "Human Movement Detection and Identification Using Pyroelectric Infrared Sensors," *Sensors*, vol. 14, no. 5, pp. 8057–8081, May 2014, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/14/5/8057

[46] A. Shokrollahi, J. A. Persson, R. Malekian, A. Sarkheyli-Hägele, and F. Karlsson, "Passive infrared sensor-based occupancy monitoring in smart buildings: a review of methodologies and machine learning approaches," *Sensors*, vol. 24, no. 5, p. 1533, 2024.

[47] J. Shin, N. Hassan, A. S. M. Miah, and S. Nishimura, "A Comprehensive Methodological Survey of Human Activity Recognition Across Diverse Data Modalities," *Sensors*, vol. 25, no. 13, p. 4028, Jan. 2025, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/25/13/4028

[48] B. Fu, N. Damer, F. Kirchbuchner, and A. Kuijper, "Sensing Technology for Human Activity Recognition: A Comprehensive Survey," *IEEE Access*, vol. 8, pp. 83 791–83 820, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9083980

[49] J. Yang, Y. Xu, H. Cao, H. Zou, and L. Xie, "Deep learning and transfer learning for device-free human activity recognition: A survey," *Journal of Automation and Intelligence*, vol. 1, no. 1, p. 100007, Dec. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949855422000077

[50] P. P. Ariza-Colpas, E. Vicario, A. I. Oviedo-Carrascal, S. Butt Aziz, M. A. Piñeres-Melo, A. Quintero-Linero, and F. Patara, "Human Activity Recognition Data Analysis: History, Evolutions, and New Trends," *Sensors*, vol. 22, no. 9, p. 3401, Jan. 2022, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/22/9/3401

[51] N. Dua, S. N. Singh, S. K. Challa, V. B. Semwal, and M. L. S. Sai Kumar, "A Survey on Human Activity Recognition Using Deep Learning Techniques and Wearable Sensor Data," in *Machine Learning, Image Processing, Network Security and Data Sciences*, N. Khare, D. S. Tomar, M. K. Ahirwal, V. B. Semwal, and V. Soni, Eds. Cham: Springer Nature Switzerland, 2022, pp. 52–71.

[52] L. Minh Dang, K. Min, H. Wang, M. Jalil Piran, C. Hee Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, p. 107561, Dec. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320320303642

[53] O. D. Lara and M. A. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6365160/

[54] M. A. Hossen and P. E. Abas, "Machine Learning for Human Activity Recognition: State-of-the-Art Techniques and Emerging Trends," *Journal of Imaging*, vol. 11, no. 3, p. 91, Mar. 2025, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2313-433X/11/3/91

[55] S. Bian, M. Liu, B. Zhou, and P. Lukowicz, "The State-of-the-Art Sensing Techniques in Human Activity Recognition: A Survey," *Sensors*, vol. 22, no. 12, p. 4596, Jan. 2022, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/22/12/4596

[56] M. Alarfaj, A. Al Madini, A. Alsafran, M. Farag, S. Chtourou, A. Afifi, A. Ahmad, O. Al Rubayyi, A. Al Harbi, and M. Al Thunaian, "Wearable sensors based on artificial intelligence models for human activity recognition," *Frontiers in artificial intelligence*, vol. 7, p. 1424190, 2024.

[57] S. Zhang, Y. Li, S. Zhang, F. Shahabi, S. Xia, Y. Deng, and N. Alshurafa, "Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances," *Sensors*, vol. 22, no. 4, p. 1476, Jan. 2022, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/22/4/1476

[58] Z. Li, J. Le Kernec, Q. Abbasi, F. Fioranelli, S. Yang, and O. Romain, "Radar-based human activity recognition with adaptive thresholding towards resource

constrained platforms," *Scientific Reports*, vol. 13, no. 1, p. 3473, Mar. 2023, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-023-30631-x

[59] A. Lazar, C. Koehler, T. J. Tanenbaum, and D. H. Nguyen, "Why we use and abandon smart devices," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '15. New York, NY, USA: Association for Computing Machinery, Sep. 2015, pp. 635–646. [Online]. Available: https://doi.org/10.1145/2750858.2804288

[60] M. A. Alanazi, A. K. Alhazmi, O. Alsattam, K. Gnau, M. Brown, S. Thiel, K. Jackson, and V. P. Chodavarapu, "Towards a Low-Cost Solution for Gait Analysis Using Millimeter Wave Sensor and Machine Learning," *Sensors*, vol. 22, no. 15, p. 5470, Jan. 2022, number: 15 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/22/15/5470

[61] N. Twomey, T. Diethe, X. Fafoutis, A. Elsts, R. McConville, P. Flach, and I. Craddock, "A comprehensive study of activity recognition using accelerometers," in *Informatics*, vol. 5, no. 2. MDPI, 2018, p. 27.

[62] D. Beckett, R. Curtis, K. Szeto, and C. Maher, "Changing User Experience of Wearable Activity Monitors Over 7 Years: Repeat Cross-Sectional Survey Study," *Journal of Medical Internet Research*, vol. 27, no. 1, p. e56251, Feb. 2025, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://www.jmir.org/2025/1/e56251

[63] J. Clawson, J. A. Pater, A. D. Miller, E. D. Mynatt, and L. Mamykina, "No longer wearing: investigating the abandonment of personal health-tracking technologies on craigslist," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '15. New York, NY, USA: Association for Computing Machinery, Sep. 2015, pp. 647–658. [Online]. Available: https://doi.org/10.1145/2750858.2807554

[64] H. Zou, J. Gao, J. Yi, C. Tian, and H. Huang, "A Review of Human Activity Recognition Based on Millimeter-Wave Radar," in *2023 China Automation Congress (CAC)*, Nov. 2023, pp. 9306–9313, iSSN: 2688-0938. [Online]. Available: https://ieeexplore.ieee.org/document/10450249

[65] A. Kristoffersson and M. Lindén, "A systematic review of wearable sensors for monitoring physical activity," *Sensors*, vol. 22, no. 2, p. 573, 2022.

[66] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 30 509–30 555, 2020.

[67] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[68] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A Review of Human Activity Recognition Methods," *Frontiers in Robotics and AI*, vol. 2, Nov. 2015, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2015.00028/full

[69] I. Jegham, A. Ben Khalifa, I. Alouani, and M. A. Mahjoub, "Vision-based human action recognition: An overview and real world challenges," *Forensic Science International: Digital Investigation*, vol. 32, p. 200901, Mar. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S174228761930283X

[70] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *Acm Computing Surveys (Csur)*, vol. 43, no. 3, pp. 1–43, 2011.

[71] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[72] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *CVPR 2011*, 2011, pp. 3169–3176.

[73] C. Wang and J. Yan, "A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition," *IEEE Access*, vol. 11, pp. 53 880–53 898, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10143178

[74] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[75] X. Li, Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote Sensing*, vol. 11, no. 9, p. 1068, 2019.

[76] F. Savvidou, S. A. Tegos, P. D. Diamantoulakis, and G. K. Karagiannidis, "Passive Radar Sensing for Human Activity Recognition: A Survey," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 5, pp. 700–706, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10577229

[77] A. M. F and S. Singh, "Computer Vision-based Survey on Human Activity Recognition System, Challenges and Applications," in *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, May 2021, pp. 110–114. [Online]. Available: https://ieeexplore.ieee.org/document/9451736

[78] C. Hinojosa, M. Marquez, H. Arguello, E. Adeli, L. Fei-Fei, and J. C. Niebles, "Privhar: Recognizing human actions from privacy-preserving lens," in *European Conference on Computer Vision*. Springer, 2022, pp. 314–332.

[79] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A Review on Human Activity Recognition Using Vision-Based Method," *Journal of Healthcare Engineering*, vol. 2017, no. 1, p. 3090343, 2017, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2017/3090343. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2017/3090343

[80] S. Mekruksavanich, W. Phaphan, N. Hnoohom, and A. Jitpattanakul, "Attention-Based Hybrid Deep Learning Network for Human Activity Recognition Using WiFi Channel State Information," *Applied Sciences*, vol. 13, no. 15, p. 8884, Jan. 2023, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2076-3417/13/15/8884

[81] M. H. Arshad, M. Bilal, and A. Gani, "Human Activity Recognition: Review, Taxonomy and Open Challenges," *Sensors*, vol. 22, no. 17, p. 6463, Jan. 2022, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/22/17/6463

[82] F. Camarena, M. Gonzalez-Mendoza, L. Chang, and R. Cuevas-Ascencio, "An Overview of the Vision-Based Human Action Recognition Field," *Mathematical and Computational Applications*, vol. 28, no. 2, p. 61, Apr. 2023, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2297-8747/28/2/61

[83] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017.

[84] X. Li, Y. He, and X. Jing, "A Survey of Deep Learning-Based Human Activity Recognition in Radar," *Remote Sensing*, vol. 11, no. 9, p. 1068, Jan. 2019, number: 9 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2072-4292/11/9/1068

[85] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE internet of things journal*, vol. 3, no. 5, pp. 637–646, 2016.

[86] S. Z. Gurbuz, M. M. Rahman, E. Kurtoglu, E. Malaia, A. C. Gurbuz, D. J. Griffin, and C. Crawford, “Multi-Frequency RF Sensor Fusion for Word-Level Fluent ASL Recognition,” *IEEE Sensors Journal*, vol. 22, no. 12, pp. 11 373–11 381, Jun. 2022.

[87] G. Lim, B. Oh, D. Kim, and K.-A. Toh, “Human activity recognition via score level fusion of wi-fi csi signals,” *Sensors*, vol. 23, no. 16, p. 7292, 2023.

[88] P. F. Moshiri, R. Shahbazian, M. Nabati, and S. A. Ghorashi, “A csi-based human activity recognition using deep learning,” *Sensors*, vol. 21, no. 21, p. 7225, 2021.

[89] S. Palipana, D. Rojas, P. Agrawal, and D. Pesch, “Falldefi: Ubiquitous fall detection using commodity wi-fi devices,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018.

[90] W. Wang, A. X. Liu, and M. Shahzad, “Gait recognition using wifi signals,” in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 363–373. [Online]. Available: https://doi.org/10.1145/2971648.2971670

[91] X. Yang, F. Xiong, Y. Shao, and Q. Niu, “Wmfall: Wifi-based multistage fall detection with channel state information,” *International Journal of Distributed Sensor Networks*, vol. 14, no. 10, p. 1550147718805718, 2018. [Online]. Available: https://doi.org/10.1177/1550147718805718

[92] J. Ding and Y. Wang, “Wifi csi-based human activity recognition using deep recurrent neural network,” *IEEE Access*, vol. 7, pp. 174 257–174 269, 2019.

[93] Z. Zhao, “A short overview of multi-modal wi-fi sensing,” *arXiv preprint arXiv:2505.06682*, 2025.

[94] X. Ding, T. Jiang, Y. Zhong, S. Wu, J. Yang, and J. Zeng, “Wi-fi-based location-independent human activity recognition with attention mechanism enhanced method,” *Electronics*, vol. 11, no. 4, p. 642, 2022.

[95] R. El Hail, P. Mehrjouseresht, D. M. M.-P. Schreurs, and P. Karsmakers, “Radar-Based Human Activity Recognition: A Study on Cross-Environment Robustness,” *Electronics*, vol. 14, no. 5, p. 875, Jan. 2025, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2079-9292/14/5/875

[96] A. M. Ashleibta, A. Taha, M. A. Khan, W. Taylor, A. Tahir, A. Zoha, Q. H. Abbasi, and M. A. Imran, “5g-enabled contactless multi-user presence and activity detection for independent assisted living,” *Scientific Reports*, vol. 11, no. 1, p. 17590, 2021.

[97] P. Picazo-Martinez, C. Barroso-Fernández, A. Calvillo-Fernandez, M. Groshev, C. J. Bernardos, A. de la Oliva, and A. Mourad, "802.11 bf multiband passive sensing: Reusing wi-fi signaling for sensing," *arXiv preprint arXiv:2507.22591*, 2025.

[98] X. Xu, J. Yu, C. Ma, Y. Ren, H. Liu, Y. Zhu, Y.-C. Chen, and F. Tang, "mmecg: Monitoring human cardiac cycle in driving environments leveraging millimeter wave," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 2022, pp. 90–99.

[99] W. Taylor, S. A. Shah, K. Dashtipour, A. Zahid, Q. H. Abbasi, and M. A. Imran, "An intelligent non-invasive real-time human activity recognition system for next-generation healthcare," *Sensors*, vol. 20, no. 9, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/9/2653

[100] H. Sadreazami, M. Bolic, and S. Rajan, "Contactless Fall Detection Using Time-Frequency Analysis and Convolutional Neural Networks," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 10, pp. 6842–6851, Oct. 2021.

[101] I. Ullmann, R. G. Guendel, N. C. Kruse, F. Fioranelli, and A. Yarovoy, "A Survey on Radar-Based Continuous Human Activity Recognition," *IEEE Journal of Microwaves*, vol. 3, no. 3, pp. 938–950, Jul. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10107532

[102] N. Nguyen, M. Pham, V.-S. Doan, and V. Le, "Improving human activity classification based on micro-doppler signatures of fmcw radar with the effect of noise," *PloS one*, vol. 19, no. 8, p. e0308045, 2024.

[103] S. A. Shah, A. Tahir, J. Le Kernec, A. Zoha, and F. Fioranelli, "Data portability for activities of daily living and fall detection in different environments using radar micro-doppler," *Neural Computing and Applications*, vol. 34, no. 10, pp. 7933–7953, May 2022. [Online]. Available: https://link.springer.com/10.1007/s00521-022-06886-2

[104] B. Erol and M. G. Amin, "Radar Data Cube Analysis for Fall Detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 2446–2450, iSSN: 2379-190X.

[105] C. Li, Z. Peng, T.-Y. Huang, T. Fan, F.-K. Wang, T.-S. Horng, J.-M. Muñoz-Ferreras, R. Gómez-García, L. Ran, and J. Lin, "A Review on Recent Progress of Portable Short-Range Noncontact Microwave Radar Systems," *IEEE Transactions on Microwave Theory and Techniques*, vol. 65, no. 5, pp. 1692–1706, May 2017.

[106] G. Bhavanasi, L. Werthen-Brabants, T. Dhaene, and I. Couckuyt, "Patient activity recognition using radar sensors and machine learning," *Neural Computing and*

Applications, vol. 34, no. 18, pp. 16 033–16 048, Sep. 2022. [Online]. Available: https://doi.org/10.1007/s00521-022-07229-x

[107] Z. Hussain, Q. Z. Sheng, and W. E. Zhang, "A review and categorization of techniques on device-free human activity recognition," *Journal of Network and Computer Applications*, vol. 167, p. 102738, Oct. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804520302125

[108] G. Li, Y. Ge, Y. Wang, Q. Chen, and G. Wang, "Detection of human breathing in non-line-of-sight region by using mmwave fmcw radar," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[109] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A Survey on Behavior Recognition Using WiFi Channel State Information," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 98–104, Oct. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/8067693/

[110] C. Wang, S. Chen, Y. Yang, F. Hu, F. Liu, and J. Wu, "Literature review on wireless sensing-Wi-Fi signal-based recognition of human activities," *Tsinghua Science and Technology*, vol. 23, no. 2, pp. 203–222, Apr. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8329114/

[111] X. Ma, Y. Zhao, L. Zhang, Q. Gao, M. Pan, and J. Wang, "Practical Device-Free Gesture Recognition Using WiFi Signals Based on Metalearning," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 228–237, Jan. 2020.

[112] F. Wang, W. Gong, and J. Liu, "On Spatial Diversity in WiFi-Based Human Activity Recognition: A Deep Learning-Based Approach," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2035–2047, Apr. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8469082/

[113] D. Wang, S. Yoo, and S. H. Cho, "Experimental comparison of ir-uwb radar and fmcw radar for vital signs," *Sensors*, vol. 20, no. 22, p. 6695, 2020.

[114] F. Khan, S. Azou, R. Youssef, P. Morel, and E. Radoi, "IR-UWB Radar-Based Robust Heart Rate Detection Using a Deep Learning Technique Intended for Vehicular Applications," *Electronics*, vol. 11, no. 16, p. 2505, Jan. 2022, number: 16 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2079-9292/11/16/2505

[115] J. Hu, H. Jiang, D. Liu, Z. Xiao, S. Dustdar, J. Liu, and G. Min, "BlinkRadar: Non-Intrusive Driver Eye-Blink Detection with UWB Radar," in *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*, Jul. 2022, pp.

1040–1050, iSSN: 2575-8411. [Online]. Available: https://ieeexplore.ieee.org/document/9912297

[116] F. Jin, A. Sengupta, and S. Cao, "mmFall: Fall Detection Using 4-D mmWave Radar and a Hybrid Variational RNN AutoEncoder," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 2, pp. 1245–1257, Apr. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9305931

[117] Texas Instruments, "AWR1843 single-chip 77- to 79-GHz FMCW radar sensor, datasheet," Texas Instruments Incorporated, Tech. Rep. SWRS216, 2019. [Online]. Available: https://www.ti.com/lit/ds/symlink/awr1843.pdf

[118] F. J. Abdu, Y. Zhang, and Z. Deng, "Activity Classification Based on Feature Fusion of FMCW Radar Human Motion Micro-Doppler Signatures," *IEEE Sensors Journal*, vol. 22, no. 9, pp. 8648–8662, May 2022.

[119] Z. Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic Continuous Hand Gesture Recognition Using FMCW Radar Sensor," *IEEE Sensors Journal*, vol. 18, no. 8, pp. 3278–3289, Apr. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8300506

[120] M. Chmurski, G. Mauro, A. Santra, M. Zubert, and G. Dagasan, "Highly-Optimized Radar-Based Gesture Recognition System with Depthwise Expansion Module," *Sensors*, vol. 21, no. 21, p. 7298, Jan. 2021, number: 21 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/21/21/7298

[121] Z. Hao, Z. Sun, F. Li, R. Wang, and J. Peng, "Millimeter wave gesture recognition using multi-feature fusion models in complex scenes," *Scientific Reports*, vol. 14, no. 1, p. 13758, Jun. 2024, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-024-64576-6

[122] L. Cao, S. Liang, Z. Zhao, D. Wang, C. Fu, and K. Du, "Human Activity Recognition Method Based on FMCW Radar Sensor with Multi-Domain Feature Attention Fusion Network," *Sensors*, vol. 23, no. 11, p. 5100, Jan. 2023, number: 11 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/23/11/5100

[123] Y. Sun, T. Fei, X. Li, A. Warnecke, E. Warsitz, and N. Pohl, "Real-Time Radar-Based Gesture Detection and Recognition Built in an Edge-Computing Platform," *IEEE Sensors Journal*, vol. 20, no. 18, pp. 10 706–10 716, Sep. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9092991

[124] Z. Gao, L. Ali, C. Wang, R. Liu, C. Wang, C. Qian, H. Sung, and F. Meng, "Real-Time Non-Contact Millimeter Wave Radar-Based Vital Sign Detection," *Sensors*, vol. 22, no. 19, p. 7560, Jan. 2022, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/22/19/7560

[125] M. Xiang, W. Ren, W. Li, Z. Xue, and X. Jiang, "High-Precision Vital Signs Monitoring Method Using a FMCW Millimeter-Wave Sensor," *Sensors*, vol. 22, no. 19, p. 7543, Jan. 2022, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/22/19/7543

[126] W.-Y. Kim and D.-H. Seo, "Radar-Based Human Activity Recognition Combining Range–Time–Doppler Maps and Range-Distributed-Convolutional Neural Networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[127] S.-P. Tseng, W.-Y. Wu, J.-F. Wang, and D. Tao, "Real-Time Driver State Detection Using mmWave Radar: A Spatiotemporal Fusion Network for Behavior Monitoring on Edge Platforms," *Electronics*, vol. 14, no. 17, p. 3556, Jan. 2025, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2079-9292/14/17/3556

[128] J. Zhou, C. Sun, K. Jang, S. Yang, and Y. Kim, "Human activity recognition based on continuous-wave radar and bidirectional gate recurrent unit," *Electronics*, vol. 12, no. 19, p. 4060, 2023.

[129] H.-S. Lim, J. Jung, J.-E. Lee, H.-M. Park, and S. Lee, "DNN-Based Human Face Classification Using 61 GHz FMCW Radar Sensor," *IEEE Sensors Journal*, vol. 20, no. 20, pp. 12 217–12 224, Oct. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9107088

[130] Y. Wang, W. Wang, M. Zhou, A. Ren, and Z. Tian, "Remote Monitoring of Human Vital Signs Based on 77-GHz mm-Wave FMCW Radar," *Sensors*, vol. 20, no. 10, p. 2999, Jan. 2020, number: 10 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/20/10/2999

[131] C. I. Radar, A. Manager, and S. Rao, "The fundamentals of millimeter wave radar sensors the fundamentals of millimeter wave radar sensors 2," 2020.

[132] G. Brooker, "Understanding millimetre wave FMCW radars," *1st International Conference on Sensing Technology*, Jan. 2005.

[133] J. Zhang, R. Xi, Y. He, Y. Sun, X. Guo, W. Wang, X. Na, Y. Liu, Z. Shi, and T. Gu, "A Survey of mmWave-Based Human Sensing: Technology, Platforms and Applications," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2023. [Online]. Available: https://

ieeexplore.ieee.org/abstract/document/10193776?casa_token=vwu-acCBLTMAAAAA:W8Soy-aaDdpzGrE-2WU899UjSM3fL3aiSploGTwoVFy7tlNuLttqv0tcz2D42yL4__a7xsHNpQ

[134] D. Gusland, J. M. Christiansen, B. Torvik, F. Fioranelli, S. Z. Gurbuz, and M. Ritchie, "Open Radar Initiative: Large Scale Dataset for Benchmarking of micro-Doppler Recognition Algorithms," in *2021 IEEE Radar Conference (RadarConf21)*, May 2021, pp. 1–6, iSSN: 2375-5318.

[135] Y. Zhou, X. Yu, M. López-Benítez, L. Yu, and Y. Yue, "Corruption Robustness Analysis of Radar Micro-Doppler Classification for Human Activity Recognition," *IEEE Transactions on Radar Systems*, vol. 2, pp. 504–516, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10522821

[136] L. Qu, Y. Wang, T. Yang, and Y. Sun, "Human Activity Recognition Based on WRGAN-GP-Synthesized Micro-Doppler Spectrograms," *IEEE Sensors Journal*, vol. 22, no. 9, pp. 8960–8973, May 2022.

[137] F. Xiang, X. Nie, C. Cui, W. Nie, and X. Dong, "Radar-based human activity recognition using two-dimensional feature extraction," in *2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Jan. 2023, pp. 267–271.

[138] S. A. Shah and F. Fioranelli, "RF Sensing Technologies for Assisted Daily Living in Healthcare: A Comprehensive Review," *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 11, pp. 26–44, Nov. 2019.

[139] W. Ding, X. Guo, and G. Wang, "Radar-based human activity recognition using hybrid neural network model with multidomain fusion," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 5, pp. 2889–2898, 2021.

[140] A. Shrestha, H. Li, J. Le Kernec, and F. Fioranelli, "Continuous human activity classification from fmcw radar with bi-lstm networks," *IEEE Sensors Journal*, vol. 20, no. 22, pp. 13 607–13 619, 2020.

[141] Y. Shao, S. Guo, L. Sun, and W. Chen, "Human Motion Classification Based on Range Information with Deep Convolutional Neural Network," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, Jul. 2017, pp. 1519–1523. [Online]. Available: https://ieeexplore.ieee.org/document/8110544

[142] J. Jung, J. Kim, S.-C. Kim, and S. Lim, "Eye-Gaze Tracking Based on Head Orientation Estimation Using FMCW Radar Sensor," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–10, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10723245

[143] L. Tang, Y. Jia, Y. Qian, S. Yi, and P. Yuan, "Human Activity Recognition Based on Mixed CNN With Radar Multi-Spectrogram," *IEEE Sensors Journal*, vol. 21, no. 22, pp. 25 950–25 962, Nov. 2021.

[144] Y. Kim and B. Toomajian, "Hand Gesture Recognition Using Micro-Doppler Signatures With Convolutional Neural Network," *IEEE Access*, vol. 4, pp. 7125–7130, 2016. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7590035/similar

[145] J. Kobak, B. J. Richman, L. Washington Jr., and S. A. Hamza, "Continuous Human Activity Recognition using a MIMO Radar for Transitional Motion Analysis," Apr. 2023, arXiv:2304.06173 [eess]. [Online]. Available: http://arxiv.org/abs/2304.06173

[146] X. Gao, H. Liu, S. Roy, G. Xing, A. Alansari, and Y. Luo, "Learning to Detect Open Carry and Concealed Object with 77GHz Radar," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 4, pp. 791–803, Jun. 2022, arXiv:2111.00551 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2111.00551

[147] M. G. Amin and R. G. Guendel, "Radar classifications of consecutive and contiguous human gross-motor activities," *IET Radar, Sonar & Navigation*, vol. 14, no. 9, pp. 1417–1429, 2020, _eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-rsn.2019.0585. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1049/iet-rsn.2019.0585

[148] P. Vaishnav and A. Santra, "Continuous human activity classification with unscented kalman filter tracking using fmcw radar," *IEEE Sensors Letters*, vol. 4, no. 5, pp. 1–4, 2020.

[149] H. D. Mafukidze, A. K. Mishra, J. Pidanic, and S. W. P. Francois, "Scattering Centers to Point Clouds: A Review of mmWave Radars for Non-Radar-Engineers," *IEEE Access*, vol. 10, pp. 110 992–111 021, 2022.

[150] J. Xie, H. Kong, J. Yu, Y. Chen, L. Kong, Y. Zhu, and F. Tang, "mm3DFace: Nonintrusive 3D Facial Reconstruction Leveraging mmWave Signals," in *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, ser. MobiSys '23. New York, NY, USA: Association for Computing Machinery, Jun. 2023, pp. 462–474. [Online]. Available: https://dl.acm.org/doi/10.1145/3581791.3596839

[151] C. Xu, Z. Li, H. Zhang, A. S. Rathore, H. Li, C. Song, K. Wang, and W. Xu, "Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 14–26.

[152] B. Jin, X. Ma, B. Hu, Z. Zhang, Z. Lian, and B. Wang, “Gesture-mmwave: Compact and accurate millimeter-wave radar-based dynamic gesture recognition for embedded devices,” *IEEE Transactions on Human-Machine Systems*, vol. 54, no. 3, pp. 337–347, 2024.

[153] M. Raja, Z. Vali, S. Palipana, D. G. Michelson, and S. Sigg, “3D Head Motion Detection Using Millimeter-Wave Doppler Radar,” *IEEE Access*, vol. 8, pp. 32 321–32 331, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8998250

[154] D. G. Bresnahan and Y. Li, “Classification of Driver Head Motions Using a mm-Wave FMCW Radar and Deep Convolutional Neural Network,” *IEEE Access*, vol. 9, pp. 100 472–100 479, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9481111

[155] Y. Ge, H. Hameed, A. Shafique, W. Zhang, S. Li, M. Z. Khan, J. Cooper, M. A. Imran, and Q. H. Abbasi, “A Driver Fatigue Detection Scheme using 3D mmWave Imaging Radar,” *IEEE Sensors Journal*, pp. 1–1, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/11134704

[156] Y. Chen, J. Yuan, and J. Tang, “A high precision vital signs detection method based on millimeter wave radar,” *Scientific Reports*, vol. 14, no. 1, p. 25535, Oct. 2024, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-024-77683-1

[157] T. K. V. Dai, K. Oleksak, T. Kvelashvili, F. Foroughian, C. Bauder, P. Theilmann, A. E. Fathy, and O. Kilic, “Enhancement of Remote Vital Sign Monitoring Detection Accuracy Using Multiple-Input Multiple-Output 77 GHz FMCW Radar,” *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology*, vol. 6, no. 1, pp. 111–122, Mar. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9439202

[158] S. Lim, G. S. Jang, W. Song, B.-h. Kim, and D. H. Kim, “Non-contact vital signs monitoring of a patient lying on surgical bed using beamforming fmcw radar,” *Sensors*, vol. 22, no. 21, p. 8167, 2022.

[159] Z. Duan and J. Liang, “Non-Contact Detection of Vital Signs Using a UWB Radar Sensor,” *IEEE Access*, vol. 7, pp. 36 888–36 895, 2019.

[160] L. Ren, H. Wang, K. Naishadham, O. Kilic, and A. E. Fathy, “Phase-Based Methods for Heart Rate Detection Using UWB Impulse Doppler Radar,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 10, pp. 3319–3331, Oct. 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7547380

[161] H. Shen, C. Xu, Y. Yang, L. Sun, Z. Cai, L. Bai, E. Clancy, and X. Huang, "Respiration and Heartbeat Rates Measurement Based on Autocorrelation Using IR-UWB Radar," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 10, pp. 1470–1474, Oct. 2018.

[162] X. Yang, X. Zhang, Y. Ding, and L. Zhang, "Indoor Activity and Vital Sign Monitoring for Moving People with Multiple Radar Data Fusion," *Remote Sensing*, vol. 13, no. 18, p. 3791, Jan. 2021, number: 18 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2072-4292/13/18/3791

[163] D. Singh, T. Eleonye, L. Surazynski, H. Ferdinando, A. Kumar, H. D. Joshi, M. Särestöniemi, and T. Myllylä, "Preliminary Studies on mm-Wave Radar for Vital Sign Monitoring of Driver in Vehicular Environment," in *Digital Health and Wireless Solutions*, M. Särestöniemi, P. Keikhosrokiani, D. Singh, E. Harjula, A. Tiulpin, M. Jansson, M. Isomursu, M. van Gils, S. Saarakkala, and J. Reponen, Eds. Cham: Springer Nature Switzerland, 2024, pp. 480–493.

[164] Z. Xie, Y. Hua, and F. Ye, "A Measurement Study of FMCW Radar Configurations for Non-contact Vital Signs Monitoring," in *2022 IEEE Radar Conference (RadarConf22)*, Mar. 2022, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9764200

[165] D. Li, Y. Huang, H. Li, J. Cheng, W. Zhu, and H. Feng, "A Novel Approach to Accurate Respiratory Rate and Heart Rate Estimation via FMCW Radar," *IEEE Sensors Journal*, pp. 1–1, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10912823

[166] Z. Chen, Y. Liu, C. Sui, M. Zhou, and Y. Song, "A Novel Scheme for Suppression of Human Motion Effects in Non-Contact Heart Rate Detection," *IEEE Access*, vol. 11, pp. 84 241–84 257, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10210373

[167] Y. Wang, Z. Wang, J. A. Zhang, H. Zhang, and M. Xu, "Vital Sign Monitoring in Dynamic Environment via mmWave Radar and Camera Fusion," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 4163–4180, May 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10160149

[168] T. Pan, Y. Guo, W. Guo, and C. Kang, "Detection of Vital Sign Based on UWB Radar by a Time Domain Coherent Accumulation Method," *IEEE Sensors Journal*, pp. 1–1, 2023.

[169] U. Ha, S. Assana, and F. Adib, "Contactless seismocardiography via deep learning radars," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. London United Kingdom: ACM, Sep. 2020, pp. 1–14. [Online]. Available: https://dl.acm.org/doi/10.1145/3372224.3419982

[170] Z. Hao, Y. Gao, Y. Tang, Y. Wang, K. Fan, and F. Li, "FMCW-based contactless heart rate monitoring," *Scientific Reports*, vol. 15, no. 1, p. 2616, Jan. 2025, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-025-86438-5

[171] Z. Xie, H. Wang, S. Han, E. Schoenfeld, and F. Ye, "DeepVS: a deep learning approach for RF-based vital signs sensing," in *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, ser. BCB '22. New York, NY, USA: Association for Computing Machinery, Aug. 2022, pp. 1–5. [Online]. Available: https://dl.acm.org/doi/10.1145/3535508.3545554

[172] M. Kebe, R. Gadhafi, B. Mohammad, M. Sanduleanu, H. Saleh, and M. Al-Qutayri, "Human vital signs detection methods and potential using radars: A review," *Sensors*, vol. 20, no. 5, p. 1454, 2020.

[173] S. Yoo, S. Ahmed, S. Kang, D. Hwang, J. Lee, J. Son, and S. H. Cho, "Radar recorded child vital sign public dataset and deep learning-based age group classification framework for vehicular application," *Sensors*, vol. 21, no. 7, p. 2412, Mar 2021. [Online]. Available: http://dx.doi.org/10.3390/s21072412

[174] A. Gorji, H.-U.-R. Khalid, A. Bourdoux, and H. Sahli, "On the Generalization and Reliability of Single Radar-Based Human Activity Recognition," *IEEE Access*, vol. 9, pp. 85 334–85 349, 2021.

[175] Y. Sun, R. Hang, Z. Li, M. Jin, and K. Xu, "Privacy-Preserving Fall Detection with Deep Learning on mmWave Radar Signal," in *2019 IEEE Visual Communications and Image Processing (VCIP)*, Dec. 2019, pp. 1–4, iSSN: 2642-9357. [Online]. Available: https://ieeexplore.ieee.org/document/8965661

[176] J. Jung, S. Lim, B.-K. Kim, and S. Lee, "CNN-Based Driver Monitoring Using Millimeter-Wave Radar Sensor," *IEEE Sensors Letters*, vol. 5, no. 3, pp. 1–4, Mar. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9366804

[177] H. N. Nguyen, S. Lee, T. T. Nguyen, and Y. H. Kim, "One shot learning based drivers head movement identification using a millimetre wave radar sensor," May 2023, arXiv:2306.05291 [eess]. [Online]. Available: http://arxiv.org/abs/2306.05291

[178] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-Pose: Real-Time Human Skeletal Posture Estimation using mmWave Radars and CNNs," *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10 032–10 044, Sep. 2020, arXiv:1911.09592 [cs, eess, stat]. [Online]. Available: http://arxiv.org/abs/1911.09592

[179] L. Zhang, X. Chen, Z. Chen, J. Zheng, and Y. Diao, “Driver Head–Hand Cooperative Action Recognition Based on FMCW Millimeter-Wave Radar and Deep Learning,” *Sensors*, vol. 25, no. 8, p. 2399, Jan. 2025, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/25/8/2399

[180] Y. Zhou, X. Yu, M. López-Benítez, L. Yu, and Y. Yue, “Radar-Based Swimming Activity Recognition With Temporal Dynamic Convolution and Spectral Data Augmentation,” *IEEE Sensors Journal*, vol. 25, no. 13, pp. 23 241–23 252, Jul. 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10684053

[181] R. G. Guendel, F. Fioranelli, and A. Yarovoy, “Distributed radar fusion and recurrent networks for classification of continuous human activities,” *IET Radar, Sonar & Navigation*, vol. 16, no. 7, pp. 1144–1161, 2022, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1049/rsn2.12249. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1049/rsn2.12249

[182] S. Z. Gurbuz, M. M. Rahman, E. Kurtoglu, T. Macks, and F. Fioranelli, “Cross-frequency training with adversarial learning for radar micro-Doppler signature classification (Rising Researcher),” in *Radar Sensor Technology XXIV*, vol. 11408. SPIE, May 2020, pp. 58–68. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11408/114080A/Cross-frequency-training-with-adversarial-learning-for-radar-micro-Doppler/10.1117/12.2559155.full

[183] M. M. Rahman, S. Z. Gurbuz, and M. G. Amin, “Physics-Aware Generative Adversarial Networks for Radar-Based Human Activity Recognition,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 3, pp. 2994–3008, Jun. 2023.

[184] M. Mahbubur Rahman and S. Z. Gurbuz, “Multi-Frequency RF Sensor Data Adaptation for Motion Recognition with Multi-Modal Deep Learning,” in *2021 IEEE Radar Conference (RadarConf21)*, May 2021, pp. 1–6, iSSN: 2375-5318.

[185] M. M. Rahman, E. A. Malaia, A. C. Gurbuz, D. J. Griffin, C. Crawford, and S. Z. Gurbuz, “Effect of Kinematics and Fluency in Adversarial Synthetic Data Generation for ASL Recognition With RF Sensors,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 58, no. 4, pp. 2732–2745, Aug. 2022.

[186] O. Pavliuk, M. Mishchuk, and C. Strauss, “Transfer learning approach for human activity recognition based on continuous wavelet transform,” *Algorithms*, vol. 16, no. 2, p. 77, 2023.

[187] A. Shrestha, C. Murphy, I. Johnson, A. Anbulselvam, F. Fioranelli, J. Le Kernec, and S. Z. Gurbuz, “Cross-Frequency Classification of Indoor Activities with DNN Transfer

Learning," in *2019 IEEE Radar Conference (RadarConf)*, Apr. 2019, pp. 1–6, iSSN: 2375-5318.

[188] X. Li, Y. He, J. A. Zhang, and X. Jing, "Supervised Domain Adaptation for Few-Shot Radar-Based Human Activity Recognition," *IEEE Sensors Journal*, vol. 21, no. 22, pp. 25 880–25 890, Nov. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9558812

[189] H.-U.-R. Khalid, A. Gorji, A. Bourdoux, S. Pollin, and H. Sahli, "Multi-View CNN-LSTM Architecture for Radar-Based Human Activity Recognition," *IEEE Access*, vol. 10, pp. 24 509–24 519, 2022.

[190] N. Kumchaiseemak, F. Fioranelli, and T. Wilaiprasitporn, "Capturing Head Poses Using FMCW Radar and Deep Neural Networks," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–12, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10839548

[191] C. Yu, Z. Xu, K. Yan, Y.-R. Chien, S.-H. Fang, and H.-C. Wu, "Noninvasive Human Activity Recognition Using Millimeter-Wave Radar," *IEEE Systems Journal*, vol. 16, no. 2, pp. 3036–3047, Jun. 2022.

[192] H. Hameed, A. Tahir, M. Usman, J. Zhu, Lubna, H. Abbas, N. Ramzan, T. J. Cui, M. A. Imran, and Q. H. Abbasi, "Wi-Fi and Radar Fusion for Head Movement Sensing Through Walls Leveraging Deep Learning," *IEEE Sensors Journal*, vol. 24, no. 9, pp. 14 952–14 961, May 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10345499

[193] A. Sen, A. Das, P. Karmakar, and S. Chakraborty, "mmAssist : Passive Monitoring of Driver's Attentiveness Using mmWave Sensors," in *2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS)*, Jan. 2023, pp. 545–553, iSSN: 2155-2509. [Online]. Available: https://ieeexplore.ieee.org/document/10041297

[194] R. W. Picard, *Affective computing*. MIT press, 2000.

[195] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, no. 1, pp. 116–134, Oct. 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314206002335

[196] R. Gravina, P. Alinia, H. Ghasemzadeh, and G. Fortino, "Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges," *Information Fusion*, vol. 35, pp. 68–80, May 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156625351630077X

[197] D. Lahat, T. Adali, and C. Jutten, "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015. [Online]. Available: https://ieeexplore.ieee.org/document/7214350

[198] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, Feb. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8269806

[199] M. Lohani, B. R. Payne, and D. L. Strayer, "A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving," *Frontiers in Human Neuroscience*, vol. 13, Mar. 2019, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2019.00057/full

[200] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, "Smart Homes that Monitor Breathing and Heart Rate," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul Republic of Korea: ACM, Apr. 2015, pp. 837–846. [Online]. Available: https://dl.acm.org/doi/10.1145/2702123.2702200

[201] G. Sikander and S. Anwar, "Driver Fatigue Detection Systems: A Review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2339–2352, Jun. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8482470/

[202] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 58–75, Jul. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0149763412001704

[203] J. Veltman and A. Gaillard, "Physiological workload reactions to increasing levels of task difficulty," *Ergonomics*, vol. 41, no. 5, pp. 656–669, 1998.

[204] J. Vicente, P. Laguna, A. Bartra, and R. Bailón, "Drowsiness detection using heart rate variability," *Medical & Biological Engineering & Computing*, vol. 54, no. 6, pp. 927–937, Jun. 2016. [Online]. Available: https://doi.org/10.1007/s11517-015-1448-7

[205] D. C. Evans and M. Fendley, "A multi-measure approach for connecting cognitive workload and automation," *International Journal of Human-Computer Studies*, vol. 97, pp. 182–189, 2017.

[206] F. Dehais, W. Karwowski, and H. Ayaz, "Brain at work and in everyday life as the next frontier: grand field challenges for neuroergonomics," p. 583733, 2020.

[207] S. Cao, P. Feng, W. Kang, Z. Chen, and B. Wang, “Optimized driver fatigue detection method using multimodal neural networks,” *Scientific Reports*, vol. 15, no. 1, p. 12240, 2025.

[208] K. A. Brookhuis and D. De Waard, “The use of psychophysiology to assess driver status,” *Ergonomics*, vol. 36, no. 9, pp. 1099–1110, 1993.

[209] H. Gorin, J. Patel, Q. Qiu, A. Merians, S. Adamovich, and G. Fluet, “A review of the use of gaze and pupil metrics to assess mental workload in gamified and simulated sensorimotor tasks,” *Sensors*, vol. 24, no. 6, p. 1759, 2024.

[210] K. Dargahi Nobari and T. Bertram, “A multimodal driver monitoring benchmark dataset for driver modeling in assisted driving automation,” *Scientific Data*, vol. 11, no. 1, p. 327, Mar. 2024, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41597-024-03137-y

[211] Q. Massoz, T. Langohr, C. François, and J. G. Verly, “The ulg multimodality drowsiness database (called drozy) and examples of use,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–7.

[212] P. Susarla, A. Mukherjee, M. L. Cañellas, C. Á. Casado, X. Wu, O. Silvén, D. B. Jayagopi, M. B. López *et al.*, “Non-contact multimodal indoor human monitoring systems: A survey,” *Information Fusion*, vol. 110, p. 102457, 2024.

[213] H. Kong, C. Huang, J. Yu, and X. Shen, “A Survey of mmWave Radar-Based Sensing in Autonomous Vehicles, Smart Homes and Industry,” *IEEE Communications Surveys & Tutorials*, vol. 27, no. 1, pp. 463–508, Feb. 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10554983

[214] S.-H. Jo, J.-M. Kim, and D. K. Kim, “Heart rate change while drowsy driving,” *Journal of Korean medical science*, vol. 34, no. 8, 2019.

[215] E. T. Solovey, M. Zec, E. A. Garcia Perez, B. Reimer, and B. Mehler, “Classifying driver workload using physiological and driving performance data: two field studies,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2014, pp. 4057–4066.

[216] U. Trutschel, C. Heinze, B. Sirois, M. Golz, D. Sommer, and D. Edwards, “Heart rate measures reflect the interaction of low mental workload and fatigue during driving simulation,” in *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2012, pp. 261–264.

[217] H. S. Yoon, H. G. Hong, D. E. Lee, and K. R. Park, "Driver's eye-based gaze tracking system by one-point calibration," *Multimedia Tools and Applications*, vol. 78, pp. 7155–7179, 2019.

[218] K. Das, M. Papakostas, K. Riani, A. Gasiorowski, M. Abouelenien, M. Burzo, and R. Mihalcea, "Detection and recognition of driver distraction using multimodal signals," *ACM Transactions on Interactive Intelligent Systems*, vol. 12, no. 4, pp. 1–28, 2022.

[219] W. Jiao, R. Li, J. Wang, D. Wang, and K. Zhang, "Activity recognition in rehabilitation training based on ensemble stochastic configuration networks," *Neural Computing and Applications*, vol. 35, no. 28, pp. 21 229–21 245, 2023.

[220] H. Storf, T. Kleinberger, M. Becker, M. Schmitt, F. Bomarius, and S. Prueckner, "An event-driven approach to activity recognition in ambient assisted living," in *Ambient Intelligence: European Conference, AmI 2009, Salzburg, Austria, November 18–21, 2009. Proceedings*. Springer, 2009, pp. 123–132.

[221] A. Zam, A. Bohlooli, and K. Jamshidi, "Unsupervised deep domain adaptation algorithm for video-based human activity recognition via recurrent neural networks," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108922, 2024.

[222] Y. Lu, L. Zhou, A. Zhang, S. Zha, X. Zhuo, and S. Ge, "Application of deep learning and intelligent sensing analysis in smart home," *Sensors*, vol. 24, no. 3, p. 953, 2024.

[223] N. C. Krishnan and D. J. Cook, "Activity recognition on streaming sensor data," *Pervasive and Mobile Computing*, vol. 10, pp. 138–154, 2014.

[224] S. Huan, L. Wu, M. Zhang, Z. Wang, and C. Yang, "Radar human activity recognition with an attention-based deep learning network," *Sensors*, vol. 23, no. 6, p. 3185, 2023.

[225] F. Ayaz, B. Alhumaily, S. Hussain, L. Mohjazi, M. A. Imran, and A. Zoha, "Integrating Millimeter-Wave FMCW Radar for Investigating Multi-Height Vital Sign Monitoring," in *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2024, pp. 1–6, iSSN: 1558-2612. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10570899

[226] S. Iyer, L. Zhao, M. P. Mohan, J. Jimeno, M. Y. Siyal, A. Alphones, and M. F. Karim, "mm-Wave Radar-Based Vital Signs Monitoring and Arrhythmia Detection Using Machine Learning," *Sensors*, vol. 22, no. 9, p. 3106, Apr. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/9/3106

[227] X. Li, Y. He, F. Fioranelli, and X. Jing, "Semisupervised human activity recognition with radar micro-doppler signatures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[228] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[229] A. Alkasimi, A.-V. Pham, C. Gardner, and B. Funsten, "Human activity recognition based on 4-domain radar deep transfer learning," in *Proc. 2023 IEEE Radar Conference (RadarConf23)*. IEEE, 2023, pp. 1–6.

[230] A. Dixit, V. Kulkarni, and V. Reddy, "Cross frequency adaptation for radar-based human activity recognition using few-shot learning," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[231] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[232] A. Faghihi, M. Fathollahi, and R. Rajabi, "Diagnosis of skin cancer using vgg16 and vgg19 based transfer learning models," *Multimedia Tools and Applications*, vol. 83, no. 19, pp. 57 495–57 510, 2024.

[233] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *Pattern Recognition*, vol. 90, pp. 119–133, 2019.

[234] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 4510–4520.

[235] F. Fioranelli *et al.*, "Radar signatures of human activities," University of Glasgow Research Data, 2019. [Online]. Available: https://researchdata.gla.ac.uk/848/

[236] S. Yang, J. L. Kernec, O. Romain, F. Fioranelli, P. Cadart, J. Fix, C. Ren, G. Manfredi, T. Letertre, I. D. H. Sáenz, J. Zhang, H. Liang, X. Wang, G. Li, Z. Chen, K. Liu, X. Chen, J. Li, X. Wu, Y. Chen, and T. Jin, "The Human Activity Radar Challenge: Benchmarking Based on the 'Radar Signatures of Human Activities' Dataset From Glasgow University," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 4, pp. 1813–1824, Apr. 2023.

[237] M. Ghosian Moghaddam, A. Asghar Nazari Shirehjini, and S. Shirmohammadi, "Device-Free Human Activity Recognition: A Systematic Literature Review," *IEEE Open Journal of Instrumentation and Measurement*, vol. 4, pp. 1–34, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10766409

[238] K. Edanami, M. Kurosawa, H. T. Yen, T. Kanazawa, Y. Abe, T. Kirimoto, Y. Yao, T. Matsui, and G. Sun, "Remote sensing of vital signs by medical radar time-series signal using cardiac peak extraction and adaptive peak detection algorithm: Performance validation

on healthy adults and application to neonatal monitoring at an nicu," *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107163, 2022.

[239] G. Beltrão, R. Stutz, F. Hornberger, W. A. Martins, D. Tatarinov, M. Alaee-Kerahroodi, U. Lindner, L. Stock, E. Kaiser, S. Goedicke-Fritz *et al.*, "Contactless radar-based breathing monitoring of premature infants in the neonatal intensive care unit," *Scientific Reports*, vol. 12, no. 1, p. 5150, 2022.

[240] European Telecommunications Standards Institute (ETSI), "ETSI EN 302 264: Short range devices; transport and traffic telematics (ttt); short range radar equipment operating in the 77 GHz to 81 GHz band; harmonised standard covering the essential requirements of article 3.2 of directive 2014/53/eu," ETSI, Tech. Rep. V2.1.1, 2017.

[241] Office of Communications (Ofcom), "UK Interface Requirement 2030 (IR 2030): Licence exempt short range devices," Ofcom, Tech. Rep., 2023. [Online]. Available: https://www.ofcom.org.uk/siteassets/resources/documents/spectrum/interface-requirements/ir-2030.pdf

[242] A. Singh, S. U. Rehman, S. Yongchareon, and P. H. J. Chong, "Multi-resident non-contact vital sign monitoring using radar: A review," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 4061–4084, 2021.

[243] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE transactions on signal processing*, vol. 62, no. 3, pp. 531–544, 2013.

[244] W. H. Organization, *Global status report on road safety 2018*. World Health Organization, 2019.

[245] S. P. McEvoy, M. R. Stevenson, and M. Woodward, "The impact of driver distraction on road safety: results from a representative survey in two australian states," *Injury Prevention*, vol. 12, no. 4, pp. 242–247, 2006. [Online]. Available: https://injuryprevention.bmj.com/content/12/4/242

[246] J. Reason, A. Manstead, S. Stradling, J. Baxter, and K. Campbell, "Errors and violations on the roads: a real distinction?" *Ergonomics*, vol. 33, no. 10-11, pp. 1315–1332, 1990.

[247] J. E. Meseguer, C. T. Calafate, and J. C. Cano, "On the correlation between heart rate and driving style in real driving scenarios," *Mobile Networks and Applications*, vol. 23, pp. 128–135, 2018.

[248] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 symposium on Eye tracking research & applications*, 2000, pp. 71–78.

[249] A. T. Duchowski and A. T. Duchowski, *Eye tracking methodology: Theory and practice*. Springer, 2017.

[250] M. Land and B. Tatler, *Looking and acting: vision and eye movements in natural behaviour*. Oxford University Press, 2009.

[251] J. W. Navalta, D. W. Davis, E. M. Malek, B. Carrier, N. G. Bodell, J. W. Manning, J. Cowley, M. Funk, M. M. Lawrence, and M. DeBeliso, "Heart rate processing algorithms and exercise duration on reliability and validity decisions in biceps-worn polar verity sense and oh1 wearables," *Scientific Reports*, vol. 13, no. 1, p. 11736, 2023.

[252] M. Tonsen, C. K. Baumann, and K. Dierkes, "A high-level description and performance evaluation of pupil invisible," *arXiv preprint arXiv:2009.00508*, 2020.

[253] P. Lab, "Pupil labs blink detector," https://assets.pupil-labs.com/pdf/Pupil_Labs_Blink_Detector.pdf, 2023, accessed: 18-11-2024.

[254] M. Drews and K. Dierkes, "Strategies for enhancing automatic fixation detection in head-mounted eye tracking," *Behavior Research Methods*, pp. 1–23, 2024.

[255] B. L. Welch, "The generalization of 'student's'problem when several different population varlances are involved," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 1947.

[256] S. Chakraborty, P. Kiefer, and M. Raubal, "Estimating perceived mental workload from eye-tracking data based on benign anisocoria," *IEEE Transactions on Human-Machine Systems*, 2024.

[257] P. Zhang, X. Wang, J. Chen, and W. You, "Feature weight driven interactive mutual information modeling for heterogeneous bio-signal fusion to estimate mental workload," *Sensors*, vol. 17, no. 10, p. 2315, 2017.

[258] E. Debie, R. Fernandez Rojas, J. Fidock, M. Barlow, K. Kasmarik, S. Anavatti, M. Garratt, and H. Abbass, "Multimodal Fusion for Objective Assessment of Cognitive Workload: A Review," *IEEE Transactions on Cybernetics*, vol. PP, pp. 1–14, Sep. 2019.

[259] V. Adarsh, G. Gangadharan, U. Fiore, and P. Zanetti, "Multimodal classification of alzheimer's disease and mild cognitive impairment using custom mkscddl kernel over cnn with transparent decision-making for explainable diagnosis," *Scientific Reports*, vol. 14, no. 1, p. 1774, 2024.

[260] P. Visconti, G. Rausa, C. Del-Valle-Soto, R. Velázquez, D. Cafagna, and R. De Fazio, "Innovative driver monitoring systems and on-board-vehicle devices in a smart-road scenario based on the internet of vehicle paradigm: A literature and commercial solutions overview," *Sensors*, vol. 25, no. 2, p. 562, 2025.

[261] H. Wiberg, E. Nilsson, P. Lindén, B. Svanberg, and L. Poom, "Physiological responses related to moderate mental load during car driving in field conditions," *Biological psychology*, vol. 108, pp. 115–125, 2015.

[262] V. Faure, R. Lobjois, and N. Benguigui, "The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior," *Transportation research part F: traffic psychology and behaviour*, vol. 40, pp. 78–90, 2016.

[263] N. Galley and G. Andres, "Saccadic eye movements and blinks during long-term driving on the autobahn with minimal alcohol ingestion," *Vision in vehicles*, vol. 5, pp. 381–388, 1996.

[264] R. Schleicher, N. Galley, S. Briest, and L. Galley, "Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired?" *Ergonomics*, vol. 51, no. 7, pp. 982–1010, 2008.

[265] G. Cardona and N. Quevedo, "Blinking and driving: the influence of saccades and cognitive workload," *Current eye research*, vol. 39, no. 3, pp. 239–244, 2014.