A Thesis for the Degree of Ph.D. in Engineering

Contactless Vital Signal Detection and Activity Recognition via Deep Learning with Radar and Doppler Ultrasound

September 2025

Graduate School of Science and Technology
Keio University

Shi Xintong

Contents

List of Tables . . . . . **v**

List of Figures . . . . . **vii**

Abstract . . . . . **ix**

Acknowledgments . . . . . **xi**

1 Introduction **1**

1.1 Background and Motivation 2

1.2 Challenges in Non-Contact Biomedical Sensing 2

1.3 Overview of thesis 5

1.4 Summary of Contributions 12

1.5 Abbreviations 13

2 A Robust Approach Assisted by Signal Quality Assessment for Fetal Heart Rate Estimation from Doppler Ultrasound Signal **17**

2.1 Introduction 18

2.2 Related Work 21

2.3 Preliminaries 24

2.3.1 Doppler Ultrasound (DUS) Signal 24

2.3.2 Autocorrelation Function (ACF) 25

2.3.3 Variational Autoencoder (VAE) 25

2.3.4 Self-Organizing Map (SOM) 26

2.4 Proposed Method 26
2.4.1 Pre-Processing 27
2.4.2 FRRI Estimation 29
2.4.3 SQA 31
2.4.4 FRRI Estimation Refinement 34
2.5 Results and Discussion 35
2.5.1 Experimental Setup 35
2.5.2 Experimental Results 37
2.5.3 Limitation and Future Works 42
2.6 Conclusions 43

3 Rough-to-Fine Model-based Non-contact Heart Rate Estimation using MIMO FMCW Radar 45
3.1 Introduction 46
3.2 Proposed Method 50
3.2.1 Person Position Estimation & Rough HR estimation 51
3.2.2 CNN-based Fine HR Estimation 53
3.3 Experiment and Results 56
3.3.1 Experimental Setup 56
3.3.2 Results and Discussion 59
3.3.3 Limitations and Future Plan 61
3.4 Conclusion 62

4 A Robust Multi-frame mmWave Radar Point Cloud-based Human Skeleton Estimation Approach with Point Cloud Reliability Assessment 63
4.1 Introduction 64
4.2 Proposed Method 65
4.2.1 Point Cloud Processing 65
4.2.2 CNN and BiLSTM-based Skeleton Estimation 66
4.2.3 LSTM-based Point Cloud Reliability Assessment 67

4.3 Experiment and Results 68
4.3.1 Experimental Setup 68
4.3.2 Results and Discussion 70
4.4 Conclusion 71

5 mmPoint-Attention: A Unified Attention Framework for Human Pose and Activity Recognition from mmWave Radar Point Clouds 73
5.1 Introduction 74
5.2 Related Work 75
5.2.1 Pose Estimation 75
5.2.2 Activity Recognition 77
5.3 Preliminaries 78
5.3.1 mmWave Radar 78
5.3.2 Attention Mechanism 79
5.4 Proposed Method 81
5.4.1 Point Cloud Data Pre-processing 82
5.4.2 Human-related Feature Extraction 84
5.4.3 Context Information Extraction 85
5.4.4 Pose Estimation & Activity Recognition 85
5.4.5 Domain-specific Fine-tuning 86
5.5 Experiments and Discussion 88
5.5.1 Experimental Settings 88
5.5.2 Experimental Results 93
5.5.3 Discussion 99
5.6 Conclusion 102

6 Conclusions and Future Work 103
6.1 Summary of Contributions 104
6.2 Future Work 106
6.2.1 FHR Estimation using Doppler Ultrasound 106

6.2.2 Heart Rate Estimation using FMCW Radar 106
6.2.3 Pose Estimation and Activity Recognition 107
6.3 Long-Term Vision and Integration 107
6.4 Closing Remarks 108

Appendix A List of Author's Publications and Awards 109
A.1 Articles on periodicals 109
A.2 Articles on international conference proceedings (reviewed full-length articles) 110
A.3 Presentations at domestic conferences 110
A.4 Presentations at domestic technical meetings 111
A.5 Others (Co-authored papers) 111

References 113

List of Tables

1.1 Limitations of conventional approaches and the contributions of Chapter 2. 7

1.2 Limitations of conventional approaches and the contributions of Chapter 3. 9

1.3 Limitations of conventional approaches and the contributions of Chapter 4. 10

1.4 Limitations of conventional approaches and the contributions of Chapter 5. 11

2.1 Comparison of the DUS-based FHR estimation performances of five scenarios for each subject. 42

3.1 The parameters of the CNN-based neural network for fine HR estimation. 53

3.2 The specifications of MIMO FMCW mmWave radar (S-Takaya T14_01120112_2D) utilized in the proposed experiment for HR estimation. 56

3.3 The descriptions of 4 cases using for the MIMO FMCW mmWave radar-based HR estimation experiments. 58

3.4 Comparision of the errors (MAE (bpm)) of HR estimation using MIMO FMCW mmWave radar while using different $Weight_I$. 59

3.5 The HR estimation errors (MAE (bpm)) using conventional methods and the proposed rough-to-fine method. 60

3.6 The HR estimation errors (MAE (bpm)) using the proposed rough-to-fine method with and without sub-window. 61

4.1 The estimation errors (MAE (cm) and RMSE (cm)) of estimated 19 joints location from mmWave radar point cloud using different methods 68

4.2 The estimation errors (MAE (cm) and RMSE (cm)) of estimated 19 joints location from mmWave radar point cloud in two scenarios: keeping or removing unreliable data detected by point cloud assessment 68

5.1 The parameters of CNN and max-pooling layers in human-related feature extraction block. 84

5.2 The activity classes in *MM-Fi* [1] dataset 89

5.3 The activity labels in *mRI* [2] dataset 89

5.4 Multi-level noise parameter settings for noise-level-based data augmentation. 90

5.5 Train-test split settings and activity selection protocols for performance evaluation of mmWave Radar point cloud-based pose estimation and activity recognition. 92

5.6 The performance comparison of mmWave Radar point cloud-based pose estimation using *MM-Fi* dataset under conventional methods and the proposed mmPoint-Attention. 94

5.7 The performance comparison of mmWave Radar point cloud-based pose estimation using *mRI* dataset under conventional methods and the proposed mmPoint-Attention. 95

5.8 Number of trainable parameters for mmWave Radar point cloud-based pose estimation using conventional methods and the proposed mmPoint-Attention. 96

5.9 Performance evaluation (*AP*) of mmWave Radar point cloud-based activity recognition using mRI dataset [2]. 98

List of Figures

1.1 An overview of this thesis. 5

2.1 DUS signal measurement for prenatal monitoring. 19
2.2 The overview of the proposed method for DUS-based FHR estimation assisted by SQA. 27
2.3 An example of filtered DUS data, spectrograms, and corresponding integrated spectrum data. 28
2.4 The workflow for FRRI estimation from DUS signals. 29
2.5 The proposed VAE and SOM-based SQA method for DUS signals. 32
2.6 The training loss of the FCN-based VAE training process for DUS SQA. . 37
2.7 The combined SQI values of the low- and high-quality integrated spectrum segments processed from raw DUS signals. 38
2.8 An example: The filtered DUS signal, the integrated spectrum, and the combined SQI values (from top to bottom) of a 5 s signal segment (20 s–25 s of subject 9). 39
2.9 Comparison of the DUS-based FHR estimation performances of five scenarios for each subject. 41

3.1 The MIMO FMCW Radar used in the proposed experiments and the principle of MIMO FMCW Radar. 47
3.2 The overall structure of the proposed rough-to-fine HR estimation method using MIMO FMCW mmWave radar. 50
3.3 Range-angle map generated from raw radar ADC data by the CL method. 51

3.4 An example of how to calculate HR while using sliding windows and sub-windows. 55
3.5 Comparison of estimation errors (MAE (bpm)) of HR estimation from MIMO FMCW mmWave radar using different weights of loss function I. . 59
3.6 An example: the ground truth and estimation of HR of subset 1 using conventional methods and the proposed rough-to-fine method. 61

4.1 CNN and BiLSTM-based neural network for mmWave radar point cloud-based human skeleton estimation. 65
4.2 LSTM-based neural network for mmWave radar point cloud reliability assessment. 67
4.3 The estimation errors (MAE (cm) and RMSE (cm)) of estimated 19 joints location from mmWave radar point cloud when selecting different values of N. 68

5.1 Point cloud generation from mmWave Radar. 79
5.2 Attention and Transformer-based deep neural network for human pose estimation and activity recognition from mmWave Radar point clouds. . . 81
5.3 Data pre-processing method for mmWave Radar point clouds. 83
5.4 The post-processing workflow for activity labels estimated from attention and Transformer-based deep neural network. 87
5.5 The estimation errors (MPJPE (m)) of LSTM, BiLSTM, Transformer-based (include w/o and w/ attention mode)-based model for pos estimation using mmWave Radar point cloud under different noise levels. 96
5.6 mmPoint-Attention-based activity recognition classification confusion matrix using *MM-Fi* dataset. 97

Contactless Vital Signal Detection and Activity Recognition via Deep Learning with Radar and Doppler Ultrasound

Xintong Shi
shixintong@keio.jp.
Keio University, 2025

Supervisor: Prof. Tomoaki Ohtsuki, Ph.D.
ohtsuki@ics.keio.ac.jp

Abstract

This thesis investigates robust, non-contact sensing methodologies for physiological signal estimation and human behavior understanding using Doppler ultrasound (DUS) and millimeter-wave (mmWave) radar technologies. In response to growing demand for non-contact, privacy-preserving monitoring in healthcare, smart home, and rehabilitation applications, this work aims to develop deep learning-based systems that can reliably interpret sparse, noisy, and environment-sensitive signals in contactless sensing systems.

Chapter 2 of this thesis addresses fetal heart rate (FHR) estimation from DUS signals. We propose a robust framework that integrates time-frequency domain information of DUS signals via short-time Fourier transform (STFT), dual autocorrelation-based peak tracking, and an unsupervised signal quality assessment (SQA) module based on variational autoencoders (VAEs) and self-organizing maps (SOMs). This design enables fetal heart rate estimation while adaptively excluding low-quality signal segments, achieving improved accuracy and robustness without requiring ground truth annotations.

Chapter 3 presents a rough-to-fine heart rate estimation framework using MIMO FMCW radar. This system integrates Curve-Length (CL)-based person localization and a deep Convolutional Neural Network (CNN) model that refines heart rate (HR) estimation using both filtered IQ signals and coarse HR estimates. The approach supports short-window inference, making it suitable for real-time health monitoring.

Chapter 4 focuses on human skeleton estimation using multi-frame mmWave radar point cloud sequences. A CNN and Bi-directional Long Short-Term Memory (BiLSTM) architecture is developed to capture both spatial structure and temporal dynamics, while an Long Short-Term Memory (LSTM)-based point cloud reliability assessment mechanism selectively suppresses unreliable input frames. The proposed model avoids voxelization and maintains real-time performance, significantly enhancing accuracy in low-quality data scenarios.

In Chapter 5, we extend the framework introduced in Chapter 4 and introduce *mmPoint-Attention*, a unified Transformer-based architecture for human pose estimation and activity recognition. This model incorporates attention and Transformer encoder layers for both human pose estimation and activity recognition. Evaluations on MM-Fi and mRI benchmark datasets demonstrate that *mmPoint-Attention* outperforms prior voxelized or two-stage models while maintaining computational efficiency.

Collectively, the contributions of this thesis form a comprehensive foundation for robust and efficient contactless sensing systems. By advancing deep learning-based technologies, this work provides practical methodologies for future smart sensing platforms in healthcare intelligence.

Acknowledgments

Words cannot fully convey my deep gratitude to my supervisor, Prof. Tomoaki Ohtsuki, for his unwavering patience, insightful feedback, and continuous support throughout my Ph.D. journey. Under his expert guidance, I not only learned how to conduct research projects but also how to grow as a researcher and continuously strive for improvement. This thesis and indeed the completion of my doctoral studies would not have been possible without his encouragement and thoughtful mentorship. I am especially thankful for his step-by-step guidance during the writing of this thesis, as well as his careful revisions of my journal and conference papers.

I am also sincerely grateful to the members of my defense committee, Prof. Ikehara, Prof. Gui, and Prof. Isogawa, for kindly accepting to serve as members of the jury, and for their valuable comments and constructive suggestions, which have greatly improved the technical depth and clarity of this thesis.

My heartfelt thanks go to JST ASPIRE (Grant Number JPMJAP2326), the Keio Leading-edge Laboratory (KLL), and the Keio "Design the Future" Award for providing research grants and scholarships that supported my Ph.D. studies.

I would like to express my appreciation to all the members of the Ohtsuki Laboratory, especially Kohei Yamamoto and Mondher Bouazizi, for their constant support both in research and daily life during my time abroad. In particular, I am deeply thankful to Associate Professor Mondher Bouazizi and Visiting Assistant Professor Kohei Yamamoto for their generous help with coding and manuscript editing.

Finally, I would like to extend my heartfelt thanks to my family and friends. Their unwavering belief in me has been a constant source of strength and motivation throughout this journey.

Xintong Shi

Chapter 1

Introduction

1.1 Background and Motivation

The global demand for remote health monitoring [1–8], personalized medicine [9, 10], and intelligent ambient systems [11–13] has increased rapidly. As a result, non-contact sensing technologies [14–18] have become indispensable in modern biomedical engineering, smart home infrastructure, and human-computer interaction (HCI) applications. These technologies facilitate unobtrusive, continuous, and remote monitoring of vital signs [19–24] and human behaviors [25–30], especially in the best interests of vulnerable populations such as elderly, newborns, pregnant women, and patients who require long-term care.

Despite their increasing prevalence, traditional modalities such as contact-based electrocardiogram (ECG) [31–36], RGB cameras [37–43], and wearable inertial sensors [44–47] exhibit inherent limitations such as discomfort, privacy concerns, susceptibility to light or motion artifacts, and maintenance overhead.

To overcome these barriers, the research community has turned to emerging modalities such as Doppler ultrasound [48–51] and millimeter-wave (mmWave) radar [52–57]. These modalities provide distinct advantages: Doppler ultrasound supports non-invasive fetal monitoring, and mmWave radar enables contactless sensing in different environments, offering robust privacy-aware alternatives to camera and wearable systems.

1.2 Challenges in Non-Contact Biomedical Sensing

Despite the growing promise of non-contact sensing technologies in healthcare and ambient intelligence applications, several core technical and practical challenges remain. These challenges span signal reliability, modeling complexity, computational cost, and real-time adaptability, each of which constrains the deployment of robust biomedical systems in real-world environments.

Fetal Heart Rate (FHR) Estimation: Accurate FHR estimation plays a critical role in prenatal care, allowing early detection of fetal distress and guiding obstetric decision making. Doppler ultrasound (DUS) has become a standard tool in clinical settings due to

its non-invasive and cost-effective nature [58]. However, the quality of the acquired DUS signal is highly susceptible to various physiological and environmental factors, such as maternal breathing, fetal motion, and sensor displacement. These factors introduce non-stationary noise and lead to significant degradation in signal-to-noise ratio (SNR), complicating the detection of fetal heartbeats. Traditional peak detection approaches based on autocorrelation functions (ACF) [3, 59] are vulnerable to such artifacts and often fail under poor SNR conditions. Moreover, while signal quality assessment (SQA) can help filter unreliable segments prior to FHR computation, most existing SQA models are supervised [60], requiring large volumes of expert-annotated data. Due to the clinical difficulty and cost of generating high-quality DUS labels, these methods often lack generalization to unseen patient populations and diverse noise patterns. Consequently, there is a strong need for frameworks that do not rely on labeled data. These frameworks should be capable of assessing signal quality and estimating FHR robustly in real-world clinical conditions. Strictly speaking, DUS for fetal heart monitoring is not a completely contactless technology, as it involves contact with the maternal body. However, from a technical perspective, the ultrasound transducer only touches the surface of the mother's abdomen and does not come into direct contact with the fetus. Therefore, in this thesis, DUS for the estimation of FHR is also classified as contactless sensing technologies.

Heart Rate Estimation via FMCW mmWave Radar: Frequency Modulated Continuous Wave (FMCW) mmWave radar systems provide the capability to extract subtle physiological movements, such as chest vibrations due to respiration and heartbeat. This opens up opportunities for contactless heart rate (HR) monitoring in scenarios such as sleep tracking, elderly care, and post-operative surveillance [61, 62]. However, existing systems are often limited in three key aspects. First, most use single-input single-output (SISO) radar hardware, which offers low angular resolution and restricts the ability to isolate signals from a specific body region. Second, to achieve reliable HR estimation, these conventional systems [63, 64] typically rely on long observation windows (10 seconds or more) to suppress noise, which hinders real-time applicability. Third, the signal quality is

highly sensitive to body posture, motion artifacts, and radar-target alignment, causing instability in dynamic environments. While deep learning methods have been introduced to assist spectral estimation [61], they are typically trained on SISO radar data and struggle to generalize across subjects or recording conditions. In summary, robust HR estimation using FMCW mmWave radar remains challenging due to hardware limitations, temporal constraints for signal processing, and sensitivity to environmental and inter-subject variability. Addressing these issues requires both improved spatial resolution and adaptive, context-aware learning algorithms.

Pose Estimation Using mmWave Radar: The mmWave radar offers a privacy-preserving and lighting-agnostic alternative to RGB cameras for human pose estimation, especially in indoor environments. Its ability to generate point cloud representations of human subjects enables fine-grained skeletal tracking [65, 66]. However, several critical challenges remain. First, most existing models rely on single-frame inputs or use voxelization [67], which quantizes point cloud data into regular 3D grids. Although this facilitates the use of convolutional architectures, it introduces spatial information loss and computational overhead. Second, radar point clouds are inherently sparse, noisy, and incomplete due to occlusion, multi-path reflection, and antenna resolution limits. These characteristics make pose estimation more difficult compared to vision-based approaches. Furthermore, prior models [65,66] tend to ignore point cloud reliability that may decrease due to radar sensor movement or environmental interference. This leads to unstable pose predictions over time. To address these issues, more temporally coherent and reliability-aware models are necessary, capable of integrating sequential radar data and assess the point cloud reliability for removing unreliable frames.

Activity Recognition: Radar-based human activity recognition (HAR) has gained traction due to its non-contact nature and resilience to lighting changes. Early HAR systems focused on classifying simple gestures or daily activities using human-defined features and conventional classifiers [68]. However, the recent introduction of complex radar datasets, such as MM-Fi and mRI [1,2], has raised the bar by including fine-grained

actions (e.g., rehabilitation activities) and longer temporal sequences. Many existing approaches rely on a two-stage pipeline. First, a pose estimation model predicts human skeleton joints from radar input. Then, a downstream classifier recognizes actions from the pose sequence. This introduces error propagation: inaccuracies in pose estimation can severely affect the accuracy of activity recognition. Moreover, this approach discards potentially useful spatial and temporal features embedded in the raw point cloud. End-to-end models that simultaneously learn both pose-related and activity-related features from spatiotemporal radar data are still rare. The challenge lies in developing architectures that can jointly model motion dynamics, spatial structure, and temporal dependencies while being robust to radar-specific noise and changing environments.

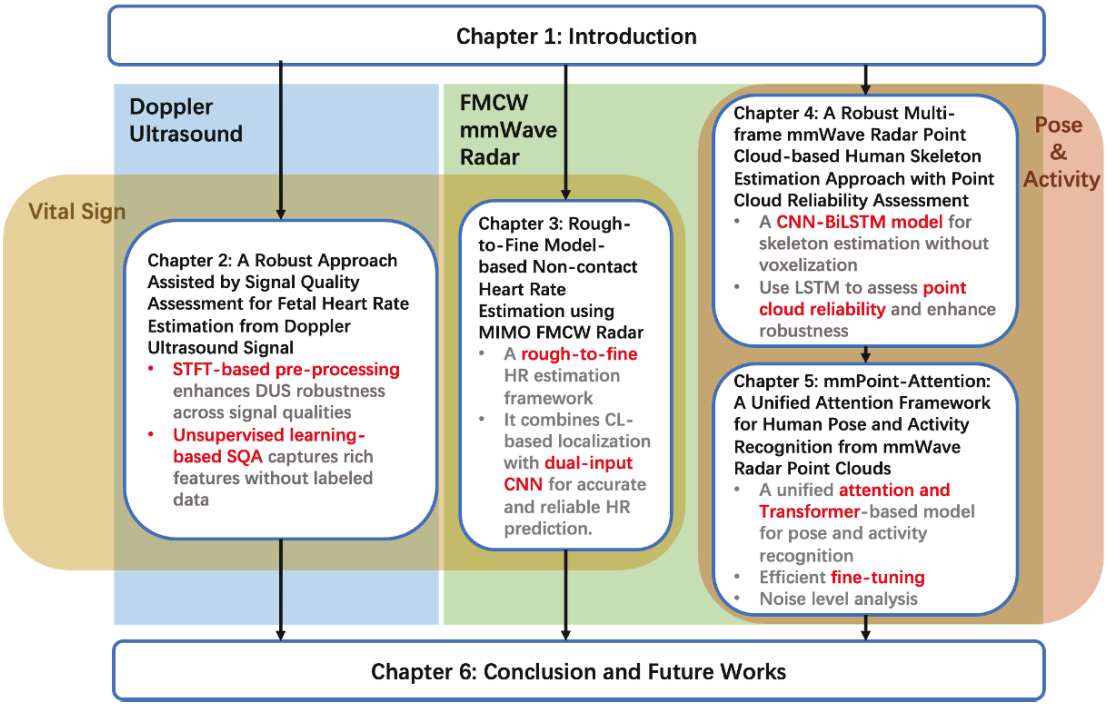


Figure 1.1. An overview of this thesis.

1.3 Overview of thesis

To overcome the challenges outlined above, this thesis presents four integrated research directions that related to advanced contactless sensing technologies with signal-aware machine learning methodologies. Each direction corresponds to a core biomedical task, and collectively they form a cohesive strategy toward building robust, interpretable, and

generalizable non-contact monitoring systems. Following this, Figure 1.1 illustrates an overview of this thesis. This thesis consists of six chapters which are introduced as follows:

- **A Robust Approach Assisted by Signal Quality Assessment for Fetal Heart Rate Estimation from Doppler Ultrasound Signal (Chapter 2):** This chapter introduces a novel unsupervised learning framework that integrates signal quality assessment with Short-Time Fourier Transform (STFT) and Autocorrelation Function (ACF)-based FHR estimation, enabling reliable fetal heart rate detection from noisy Doppler ultrasound (DUS) signals. The method bypasses the need for annotated training data and utilizes the unsupervised-learning method for signal quality assessment. The validation on real data collected from hospital demonstrates its robustness under diverse noise conditions. The contributions of Chapter 2 are summarized in Table 1.1. Here, the full terms represented by the abbreviations in the Table 1.1 are explained. DUS: Doppler ultrasound; SQA: Signal Quality Assessment; STFT: Short-Time Fourier Transform; FHR: Fetal Heart Rate; FRRI: Fetal R-R Interval; VAE: Variational Autoencoder; SOM: Self-Organizing Map; SQI: Signal Quality Index.

- **Rough-to-Fine Model-based Non-contact Heart Rate Estimation using MIMO FMCW Radar (Chapter 3):** A rough-to-fine signal processing and learning framework is developed for heart rate estimation using high-resolution Multiple-Input and Multiple-Output (MIMO) FMCW mmWave radar. The pipeline includes human localization, rough HR estimation through traditional method, and fine estimation using a dual-input CNN model. This approach reduces the length of required time windows while maintaining accuracy, making it suitable for real-time deployment in smart-home applications. The contributions of Chapter 3 are summarized in Table 1.2. Here is the list of abbreviations used in the Table 1.2 along with their full

Table 1.1. Limitations of conventional approaches and the contributions of Chapter 2.

Limitations of conventional methods	**Solutions**	**Contributions**
	FHR–estimation stage	
Only a single domain (time or frequency) is analyzed [3, 59, 69]	Combine time & frequency cues	STFT–based time–frequency preprocessing
Performance is highly dependent on raw DUS quality [70]	Introduce SQA	SQA block that gates unreliable segments
	SQA stage	
Annotation of quality labels is time-consuming [60, 70]	Use unsupervised learning	VAE + SOM unsupervised latent-space learning
Reliance on hand-crafted features limits expressiveness [60, 70]	Learn representations from data	Deep representation learning for SQIs
Discarding bad segments lowers overall coverage [70]	Post-refinement rather than removal	Kalman-filter FRRI correction

forms. DL: Deep Learning; CNN: Convolutional Neural Network; SNR: Signal-to-Noise Ratio; HR: Heart Rate; NOMP: Newtonized Orthogonal Matching Pursuit; SISO: Single-Input Single-Output; MIMO: Multiple-Input Multiple-Output; FMCW: Frequency-Modulated Continuous Wave.

- **A Robust Multi-frame mmWave Radar Point Cloud-based Human Skeleton Estimation Approach with Point Cloud Reliability Assessment (Chapter 4):** This chapter proposes a Convolutional Neural Network (CNN) and Bi-directional Long Short-Term Memory (BiLSTM)-based model that operates directly on sequential radar point clouds for human skeleton estimation (pose estimation). By treating each point as a dynamic entity over time, the model learns both spatial structure and motion patterns without voxelization. A point cloud reliability assessment mechanism is introduced to improve the robustness for pose estimation by recognizing and removing the unreliable frames. The model shows strong robustness to sparse and noisy sequences. The contributions of Chapter 4 are summarized in Table 1.3.

- **mmPoint-Attention: A Unified Attention Framework for Human Pose and Activity Recognition from mmWave Radar Point Clouds (Chapter 5):** Addressing the problems of conventional pipelines, this chapter presents an end-to-end architecture that simultaneously performs pose estimation and activity classification fro radar point cloud data. The model integrates some attention mechanisms and Transformer encoder layers for both human pose estimation and activity recognition. Evaluations on several datasets, including rehabilitation tasks, highlight the model's capability in real-world scenarios. The contributions of Chapter 5 are summarized in Table 1.4.

- **Conclusion and Future Works (Chapter 6):** This research builds the groundwork for a new type of contactless sensing system that is smart, reliable, and easy to use in everyday situations. To make these ideas a reality, future work with hospitals and companies will be important.

Table 1.2. Limitations of conventional approaches and the contributions of Chapter 3.

Limitations of conventional methods	**Solutions**	**Contribution**
	Non-DL-based Methods	
Conventional peak detection methods suffer from interferences [64, 71]	Signal decomposition-based methods isolate heartbeat signal [64]	Propose DL-based method that works with short windows and improves robustness
Signal decomposition needs long time windows [72]	Utilize short-window-based CNN refinement for HR estimation	
	DL-based Methods (Only 2 works are published [61, 72])	
Paper [72]: - Low SNR degrades HR estimation with NOMP - Lack of context info in DL methods	- More robust DL method - Uses nearby sub-window context for DL learning	Employ CNN-based model and smaller sub-windows to refine the rough HR
Paper [61]: - DL method only uses range profile - SISO radar limits generalizability	- Extract heart-related signals from more accurate location of human - Employs MIMO FMCW radar	- Utilize CL-based approach to detect the subject's position using MIMO FMCW Radar - Propose a novel reliability matrix for refining the rough HR.

Table 1.3. Limitations of conventional approaches and the contributions of Chapter 4.

Limitations of conventional methods	**Solutions**	**Contributions**
Based on voxelization [67] - High computational cost - Reduced spatial resolution	Without voxelization Directly use raw point cloud data	Reduced computational overhead while maintaining fine-grained spatial features
Single-frame inputs [2, 65, 66, 73] - Lack of temporal context - Low robustness	Multi-frame inputs Use of model understanding sequential data	Better robustness and contextual understanding through temporal modeling
Point cloud reliability have not been proposed yet	Reliability assessment network Frame-wise confidence evaluation	Introduced a neural reliability estimator to adaptively weigh unreliable inputs

Table 1.4. Limitations of conventional approaches and the contributions of Chapter 5.

Limitations of Conventional Methods	**Solutions**	**Contributions**
Two-step estimation may cause error accumulation [2] -Activity recognition relies on estimation of joint positions	End-to-end framework which can directly predict activity class	Improved accuracy of activity recognition
Pre-processing highly depends on voxelization -increases computation -reduces spatial detail [68]	Directly use of raw point cloud	Pre-processing without voxelization
Single-frame input lacks temporal context [65, 66]	Multi-frame input and temporal modeling	Build a model using transformer encoder and attention mechanisms
Domain shift leads to degraded performance [74] -Changing environment and subject may reduce accuracy	Fine-tuning	Propose finetuning techniques which provide efficient adaptation to new environments with limited data

These four research threads are unified under a common goal: to design deep learning-driven non-contact sensing systems that are adaptable to signal variations, capable of leveraging temporal context, and deployable in practical scenarios involving vulnerable users or privacy-sensitive conditions.

1.4 Summary of Contributions

This thesis contributes both methodological innovations and application-level validations that advance the field of contactless biomedical sensing. The main contributions can be summarized as follows:

- A novel pre-processing strategy is proposed for DUS signals by combining time and frequency information through Short-Time Fourier Transform (STFT). This approach is inherently robust to variations in DUS signal quality, as it is supported by a SQA module. To enable this, we introduce an unsupervised representation learning-based SQA framework that eliminates the need for large annotated datasets of quality labels. Moreover, the use of representation learning allows our model to capture richer and more abstract information than conventional hand-crafted features. Compared to the baseline MAE of 2.51 bpm, the proposed method achieves a substantially lower error of 0.74 bpm in FHR estimation.

- A rough-to-fine HR estimation framework using MIMO FMCW mmWave radar is introduced, combining a curve-length (CL)-based human localization method with a two-stage estimation pipeline. After detecting the target's position, the system extracts heartbeat-related signals from selected radar cells. A rough HR estimate is obtained via peak detection, followed by a refined estimate using a dual-input, dual-output CNN model that captures richer temporal context and is capable to assign weights to the selected cells for improving the HR estimation accuracy. The proposed model reduces the MAE of HR estimation by 3.64 bpm compared to baseline.

- A novel CNN and BiLSTM-based human skeleton estimation model is introduced that utilizes multi-frame point cloud data without the need for voxelization. Additionally, a Long Short-Term Memory (LSTM)-based neural network is introduced to assess the reliability of point cloud data segments, further enhancing the robustness of our method. This approach achieves a 3 cm reduction in joint location MAE and reduces the number of trainable parameters by 75% compared to the baseline.

- A framework *mmPoint-Attention* is proposed, which is a unified deep learning framework that combines attention mechanisms with Transformer encoders to perform human pose estimation and activity recognition directly from multi-frame point cloud data, eliminating the need for voxelization. Our model incorporates attention mechanisms during feature extraction and context modeling, leading to improved accuracy in 3D joint localization and complex activity classification compared to prior methods. A lightweight post-processing step enhances prediction consistency in continuous sequences. Furthermore, we address domain shift challenges through efficient fine-tuning with limited target data, achieving strong generalization while maintaining computational efficiency. Our method improves pose estimation accuracy by 2 cm PA-MPJPE and improves 9% in mean average precision (mAP) for activity recognition, while reducing the number of trainable parameters by 53% compared to the baseline.

The effectiveness of these contributions is demonstrated through extensive experiments on diverse datasets, including real-world clinical DUS recordings, mmWave radar point clouds, and FMCW mmWave radar raw Analog-to-Digital Converter (ADC) data. Together, these contributions push forward the integration of signal processing and deep learning for future-proof, contactless health monitoring in healthcare intelligence.

1.5 Abbreviations

Table 1.5 shows the abbreviations used in this thesis.

Abbreviation	Full Term
AAE	Average Absolute Error
ACF	Autocorrelation Function
ADC	Analog-to-Digital Converter
AE	Autoencoder
BiLSTM	Bidirectional Long Short-Term Memory
CL	Curve Length
CNN	Convolutional Neural Network
CTG	Cardiotocography
DL	Deep Learning
DUS	Doppler Ultrasound
ECG	Electrocardiogram
ELU	Exponential Linear Unit
EMD	Empirical Mode Decomposition
FCN	Fully Convolutional Network
FHR	Fetal Heart Rate
FMCW	Frequency-Modulated Continuous Wave
FRRI	Fetal R-R Interval
HCI	Human-Computer Interaction
HR	Heart Rate
IMF	Intrinsic Mode Function
KF	Kalman Filter
KLD	Kullback–Leibler Divergence
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MIMO	Multiple-Input Multiple-Output
MLP	Multi-Layer Perceptron
mRI	mmWave Radar Intelligence dataset
mmWave	Millimeter Wave
NAE	Normalized Absolute Error
NLP	Natural Language Processing
NOMP	Newtonized Orthogonal Matching Pursuit
PC	Point Cloud
PPG	Photoplethysmogram
QE	Quantization Error
ReLU	Rectified Linear Unit

Abbreviation	Full Term
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RRI	R-R Interval
SISO	Single-Input Single-Output
SNR	Signal-to-Noise Ratio
SOM	Self-Organizing Map
SQA	Signal Quality Assessment
SQI	Signal Quality Index
STFT	Short-Time Fourier Transform
TI	Texas Instruments
VAE	Variational Autoencoder

Chapter 2

A Robust Approach Assisted by Signal Quality Assessment for Fetal Heart Rate Estimation from Doppler Ultrasound Signal

2.1 Introduction

Fetal heart rate (FHR) serves as a vital metric for assessing the well-being of a fetus. It aids in identifying high-risk pregnancies and reducing fetal mortality rates. Cardiotocography (CTG) is a commonly employed method that allows the simultaneous recording of both maternal and fetal heartbeats. Among the techniques used to extract FHR from CTG recordings, the Doppler ultrasound (DUS) technique has gained prominence [58]. This involves attaching an ultrasound (US) transducer to the mother's abdomen to enable continuous monitoring, as depicted in Figure 2.1. Previous studies [75, 76] have established that the most accurate FHR estimation relies on invasive fetal electrocardiogram (ECG), with non-invasive fetal ECG representing the second most precise method. While FHR estimation based on DUS signals does not attain the same level of accuracy as fetal ECG, it offers distinct advantages. Notably, invasive fetal ECG is restricted to specific cases due to the undesirability of long-term attachment of fetal scalp electrodes for FHR monitoring. Conversely, non-invasive fetal ECG signals obtained by placing electrodes on the mother's abdomen necessitate expensive equipment and expert operation [77]. To capture a DUS signal, the ultrasound (US) transducer is positioned on the mother's abdomen. The primary technique for estimating fetal heart rate (FHR) from these DUS signals is based on the autocorrelation function (ACF) [3, 59, 69]. This method leverages the inherent periodicity of the DUS signals to extract the FHR information accurately. Another notable approach for FHR estimation involves empirical mode decomposition (EMD). By employing EMD, it becomes possible to identify signal components associated with the valve motion of the fetal heart [78, 79]. The ACF-based approaches have demonstrated superior consistency and applicability compared to EMD-based methods, primarily due to their simpler computations and fewer parameters. However, current approaches still face several challenges that warrant attention. Firstly, prior studies [3, 59, 69] have predominantly focused on a single domain (time or frequency) during ACF calculations, potentially resulting in the loss of valuable information from the other domain. Moreover, DUS signals are susceptible to various interferences, including maternal, fetal, and instrument movements, leading to a degradation in the signal-to-noise ratio (SNR). Accurately estimating

FHR using noisy DUS data becomes challenging, as the quality of the fetal DUS signals significantly impacts the accuracy of the FHR estimation [70]. To enhance FHR estimation accuracy, a common technique employed is signal quality assessment (SQA), which involves identifying segments of low-quality signals and subsequently eliminating or interpolating erroneous FHR values derived from these segments. Presently, the available literature on signal quality assessment (SQA) methods for DUS signals remains limited, with only a few approaches published [60, 70]. However, these approaches based on supervised machine learning techniques usually require large labeled datasets, and there are few public DUS datasets with professional annotations. Furthermore, the existing SQA methods primarily focus on removing the estimated fetal heart rates (FHRs) from the identified segments of low-quality DUS signals. However, this approach may result in a reduced proportion of reserved FHRs in relation to the overall number of estimated FHRs within a record. Hence, alternative strategies are required to address this issue effectively.

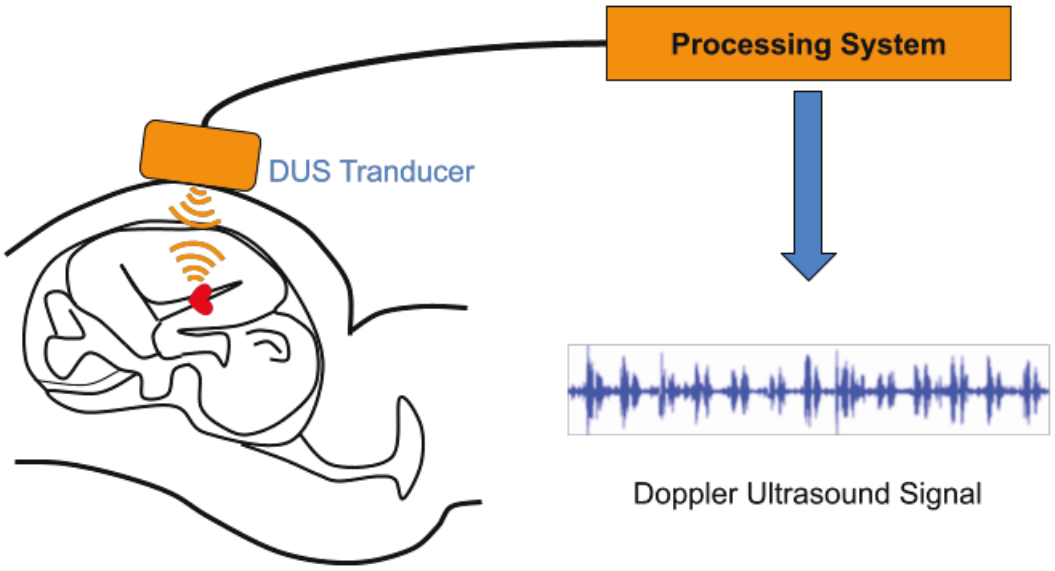


Figure 2.1. DUS signal measurement for prenatal monitoring.

Inspired by the shortcomings of existing DUS-based FHR estimation and DUS SQA methods, this study present a robust DUS-based FHR estimation method supported by DUS SQA based on unsupervised representation learning. By means of a band-pass filter and a short-time Fourier transform (STFT), the DUS signals are pre-processed in the

form of spectrograms. The spectra are integrated over the spectrograms to produce integrated spectral data. The ACF is applied to a 3.75 s integrated spectrum segment and a peak is detected. An approximate fetal RR interval (FRRI) is first determined from the detected ACF peak, which is used to resize the window for the second ACF calculation. The detected peak of the second ACF is defined as the FRRI of that segment. In addition, the length of the window movement is determined based on the FRRI estimated from the previous window. The proposed DUS SQA uses a fully convolutional network (FCN)-based variational autoencoder (VAE) to obtain learning representations of pre-processed DUS signals by integrating the autoencoder (AE) models mentioned in [80, 81]. By feeding these learned representations into a self-organizing map (SOM), the study generate a combined signal quality indicator (SQI) that incorporates the quantization error (QE) of the SOM [82]. The combined SQI is used to modify the measurement noise covariance, and a Kalman filter (KF) is then used to correct the inaccurate FRRIs computed from low-quality signals [83, 84]. The study would like to mention that the proposed DUS SQA method has been previously reported [85], serving as a foundation for the proposed current research. In this thesis, the study utilize the DUS SQA method proposed in the proposed previous work [85]. However, the fetal heart rate estimation section is entirely novel and presented for the first time. Additionally, the study include a comparative analysis of the experimental results between the proposed approach and a conventional method, which remains unpublished. The study conducted a thorough evaluation of the proposed method using DUS recordings obtained from ten subjects. The experimental results clearly demonstrated that the proposed method outperformed the conventional approach [3] in terms of FHR estimation accuracy. Furthermore, the integration of the proposed DUS SQA method led to a substantial reduction in the estimation errors of FHRs. The main contributions of the proposed work can be summarized as follows:

- This study combine time and frequency information using STFT for DUS signal pre-processing so that the ACF-based FHR estimation algorithm does not focus only on a single domain as in previous works.

- This method is robust to DUS recordings of different qualities because it is supported by DUS SQA.
- An unsupervised representation learning-based DUS SQA approach is proposed in this thesis, which eliminates the need for a large dataset of quality labels. Furthermore, representation learning enables the proposed method to exploit deeper information than human-defined features.

The structure of this chapter is as follows. Section 2.2 summarises and discusses related work. Subsequently, the preliminary work is presented in Section 2.3. The proposed method is described in Section 2.4. In Section 2.5, this study evaluate the performance of the proposed proposed method and discuss its strengths and limitations. Finally, this study conclude this chapter in Section 2.6.

2.2 Related Work

Some FHR estimation techniques using DUS signals have been reported [3,59,69,78,79]. ACF-based FHR estimators are the most commonly used, which have been shown to be computationally inexpensive and can achieve accurate FHR estimation. A two-step approach for FHR extraction using DUS signals has been proposed by Peter et al. [69], based mainly on the ACF technique. The cardiac cycle times are roughly estimated based on low-pass filtering of the envelope signals, and then the FHRs are calculated using ACF in the frequency domain. However, they mentioned that this method has a relatively high sensitivity to signal-to-noise ratio (SNR), which means that low SNR can lead to inaccurate results. Jezewski et al. [59] presented an ACF-based technique for FHR estimation from DUS signals, which includes three main steps: dynamic adjustment of the ACF window, adaptive ACF peak detection, and determination of beat-to-beat intervals. The ACF windows are continuously adjusted based on the most recent FHR estimate. In addition, if ACF peaks fall within the previously identified FHR range, they are assigned a higher probability of correlating with the correct cardiac cycle. To assess the quality of DUS signals, they also propose that a lower ACF peak indicates lower signal

quality, which increases the possibility of inaccurate estimation of cardiac cycle duration. Although Jezewski et al. [59] introduced an indicator to reduce the negative effects of low-quality signals, a single indicator is not reliable enough to remove all low-quality signals. Another research paper [3] provides a generalizable, reproducible, open-source ACF-based technique that can accurately estimate FHR from 1D-DUS obtained using a low-cost handheld transducer. A total of 721 DUS signal segments of 3.75 s length were used to train an FHR estimator, and the parameters of the estimator were adjusted by Bayesian optimization. In addition, it has been mentioned in this thesis that only high-quality signal segments are selected for experiments based on the SQA method proposed in [70]. However, by using a rectangular window, the ACF was calculated only in the time domain, so some information in the frequency domain may be lost. Moreover, the fixed window length for ACF calculation limits the accuracy of FHR estimates.

In addition to ACF-based approaches, the empirical mode decomposition (EMD) method has also been used for FHR estimation. Rouvre et al. [78] proposed that the performance of FHR estimation can be improved by applying ACF to the intrinsic mode functions (IMFs) after empirical mode decomposition. However, they also mentioned that this method has not yet been fully formalized theoretically and that the computational complexity is highly dependent on the intrinsic signal properties. Furthermore, an EMD kurtosis method for FHR extraction from DUS signals was presented by Al-Angari et al. [79]. In their work, the kurtosis function is calculated on the FHRs extracted from the DUS signal. It is worth mentioning that the most important indicator of this approach is the appropriate window size used to apply the kurtosis function. It has been shown that the EMD-based methods have better performance than the ACF-based methods when processing signals with relatively high SNR, in other words, the EMS-based methods are more robust. However, it is very possible that some related parameters, including the number of IMFs and the size of the window used to calculate kurtosis, may be overfitted during optimization due to the limited amount of data.

Compared to EMD-based techniques, ACF-based methods show higher reproducibility and practicality due to the relatively low computational complexity and a small number

of parameters. However, there are still several drawbacks to the existing methods that can be improved:

- The ACF calculations in these existing works only focus on one domain, time, or frequency. Thus, the information about the other domain may be lost during the calculation.

- The performance of ACF-based FHR estimation techniques is highly dependent on the quality of the DUS signal. Although some papers [3, 59] use simple SQA approaches to solve this problem, these DUS SQA methods can still be enhanced.

Up to now, a small number of SQA methods for DUS signals have been introduced. A DUS SQA based on the Support Vector Machine was proposed by C. E. Valderrama et al. [70]. Some innovative template-based SQIs derived from correlation coefficients between DUS signals and the fetal heartbeat-based template signals are used for DUS SQA. The Naive Bayes classifier was also used for DUS SQA [60], where twelve SQIs were used. These SQIs are mostly related to the range of valve movement. Although it has been demonstrated that these DUS SQA approaches can discriminate between different quality levels, there are still certain limitations:

- Large labeled datasets are usually required for these research works based on supervised machine learning methods. However, there are few DUS datasets with quality annotations. In addition, annotation of DUS quality levels is laborious and requires expert knowledge.

- The human-defined signal quality features used in these works limit the ability to mine deeper signal quality information in DUS signals.

- These existing methods simply eliminate the estimated FHRs from the detected low-quality DUS signal segments, which may result in a reduction in the proportion of reserved FHRs to all estimated FHRs in each recording.

Although not for DUS signals, some unsupervised learning-based SQA methods have been introduced to improve signal analyzing performance. A paper [86] evaluates photoplethysmography signal quality by computing seven SQIs associated with entropy and

waveform morphology and training a SOM for quality-level classification. In addition, two AE-based ECG SQIs associated with reconstruction errors and reconstruction reliabilities have been proposed by N. Seeuws et al. [80]. Recently, unsupervised representation learning has gained popularity for time series classification tasks. In the paper [87], time series data can be transformed into an instance-feature matrix using a proposed effective unsupervised representation learning methodology, where they also demonstrated that these features can be used to perform precise time series clustering tasks. J. Pereira et al. [81] have also proposed a VAE-based representation learning strategy for anomaly identification. How to appropriately apply the SQA results for achieving better FHR estimation accuracy is another crucial task, in addition to the evaluation of the signal quality. Li et al. [83] introduced a method that utilizes KF to correct heart rate estimates, incorporating four ECG SQIs to adaptively adjust the measurement noise covariance. In addition, for non-invasive fetal ECG signals, another paper [84] also uses KF to enhance the precision of FHR estimation.

2.3 Preliminaries

2.3.1 Doppler Ultrasound (DUS) Signal

For continuous recording, the US transducer is placed on the maternal abdomen to extract DUS signals, as shown in Figure 2.1. When US waves are reflected from an object, the frequency of the US waves changes, which is called the Doppler frequency shift. This frequency shift, f_s, is given by:

$$f_s = \frac{2 f_0 v \cos\theta}{c}, \tag{2.1}$$

where c is the US propagation velocity and f_0 is the transmission frequency. θ is the angle of the object along the direction of the US wave, and $v\cos\theta$ is the velocity of the object along that direction [77]. The US waves propagate through the mother's skin and some tissues. The transmitted US waves are reflected by the fetal chest wall. During reflection, the frequency of the US waves changes due to the movement of the heart wall and valves

of the fetal heart and, in some cases, blood flow. The reflected US waves are received by the transducer on the mother's abdomen. The received US waves contain such fetal physical movements, including heart activity, and thus it is possible to estimate the FHR by analyzing the received US wave, i.e., a DUS signal.

2.3.2 Autocorrelation Function (ACF)

It provides a statistical representation of the similarity between a time series and its delayed version. By utilizing the ACF, the analysts can compare the current value of a dataset with its past values. This function operates by comparing a given time series with a lagged version of itself across one or multiple time periods [88]. For a given time series data $[Y_1, Y_2, ..., Y_N]$, the ACF with a lag of k is defined as follows:

$$acf_k = \frac{\sum_{i=1}^{N-k}(Y_i - \overline{Y})(Y_{i+k} - \overline{Y})}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}. \tag{2.2}$$

2.3.3 Variational Autoencoder (VAE)

Autoencoder (AE), which is a neural network consisting of an encoder and a decoder, can extract representations from the inputs and reconstruct the inputs from these representations [89]. By inputting a vector $\boldsymbol{x}$ into AE, a representation vector $\boldsymbol{z}$ in the latent layer (between encoder and decoder) can be compressed by the encoder neural network and reconstructed to the reconstructed vector $\hat{\boldsymbol{x}}$ by the decoder neural network. The redundancies associated with unnecessary information can be eliminated during representation learning, while the most important information of the input data can be retained. Therefore, the reconstruction of the input vector $\boldsymbol{x}$ from the representation vector $\boldsymbol{z}$ is possible after training for several epochs. The proposal of VAE [90] rapidly promoted the development of AE technology to a great extent. The most significant difference between VAE and AE is that the representation vector $\boldsymbol{z}$ in the latent layer of VAE is a random variable with a Gaussian distribution, $\mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z^2)$. In addition, it is worth noting that while the reconstruction error typically serves as the primary loss function for AE, VAE introduces an additional loss component based on the Kullback–Leibler divergence (KLD) between

the Gaussian distribution $\mathcal{N}(\mu_z, \sigma_z^2)$ and the standard normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. This KLD loss term further enhances the capability to capture and model the latent space distribution.

2.3.4 Self-Organizing Map (SOM)

To address the challenges of SQA, SOM, a clustering method based on unsupervised learning, has been used in some papers [86, 91]. SOM employs a two-layer neural network to map an N-dimensional input layer onto a two-dimensional output layer. The output layer consists of multiple neurons, each serving as a potential match for an input sample, thereby identifying a winning neuron. SOM training typically involves competitive learning. The QE of an input vector can be regarded as the Euclidean distance between that vector and its corresponding winning neuron. T. Kohonen et al. [82] noted that a smaller QE signifies a better alignment between the input sample and the trained SOM model [82]. By leveraging the QE of a set of input data, it becomes possible to identify anomalies within the data. For instance, it can be utilized to detect low-quality signal segments when the majority of signal segments exhibit high quality.

2.4 Proposed Method

As shown in Figure 2.2, four components make up the framework of the robust FHR estimation approach assisted by DUS SQA. These four blocks are described separately as follows.

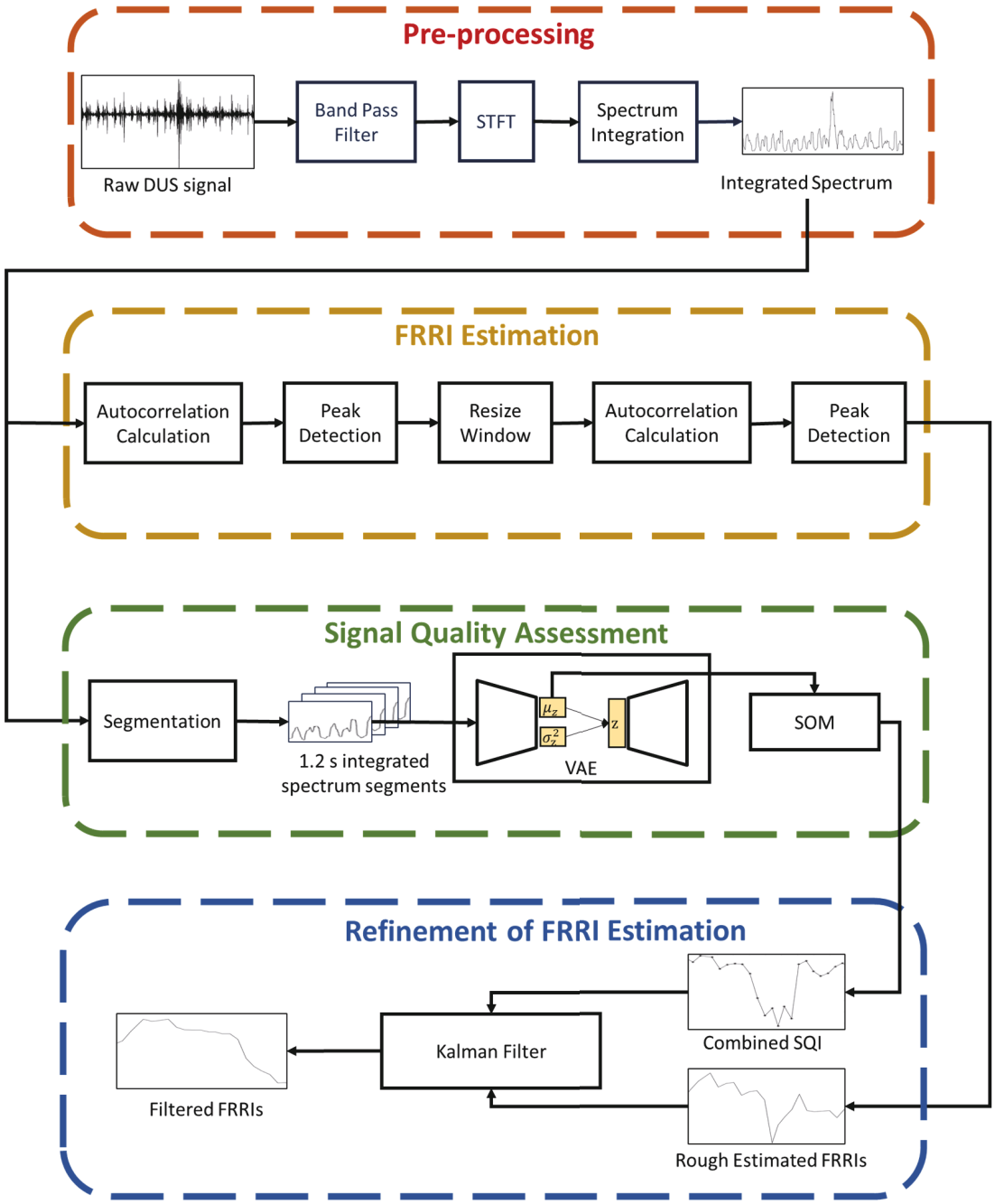
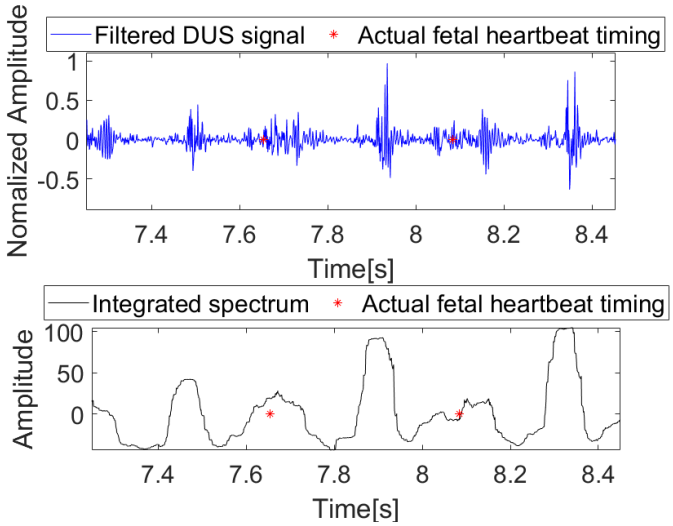


Figure 2.2. The overview of the proposed method for DUS-based FHR estimation assisted by SQA.

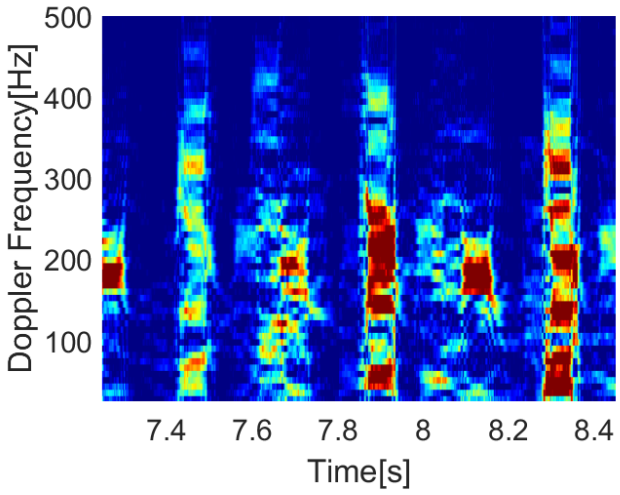
2.4.1 Pre-Processing

In general, fetal heart activity occurs at frequencies below 500 Hz. In addition, frequencies below 25 Hz are primarily correlated with noise generated by fetal movement or

equipment. Thus, this study apply a bandpass filter to raw DUS signals with cutoff frequencies of 25 Hz and 500 Hz to focus on fetal heartbeats [3]. In addition, 2D spectrograms are generated using STFT with a window length of 64 ms and a step size of 1 ms. To generate integrated spectrum data from spectrograms, which are time series data, the spectra are integrated over the spectrograms in the frequency range [25, 500] Hz [92]. An example of filtered DUS data, spectrograms, and integrated spectrum data is shown in Figure 2.3. The actual fetal heartbeat timings (R-peaks) are marked in red.



(a) Filtered DUS signal and integrated spectrum



(b) Spectrogram

Figure 2.3. An example of filtered DUS data, spectrograms, and corresponding integrated spectrum data.

2.4.2 FRRI Estimation

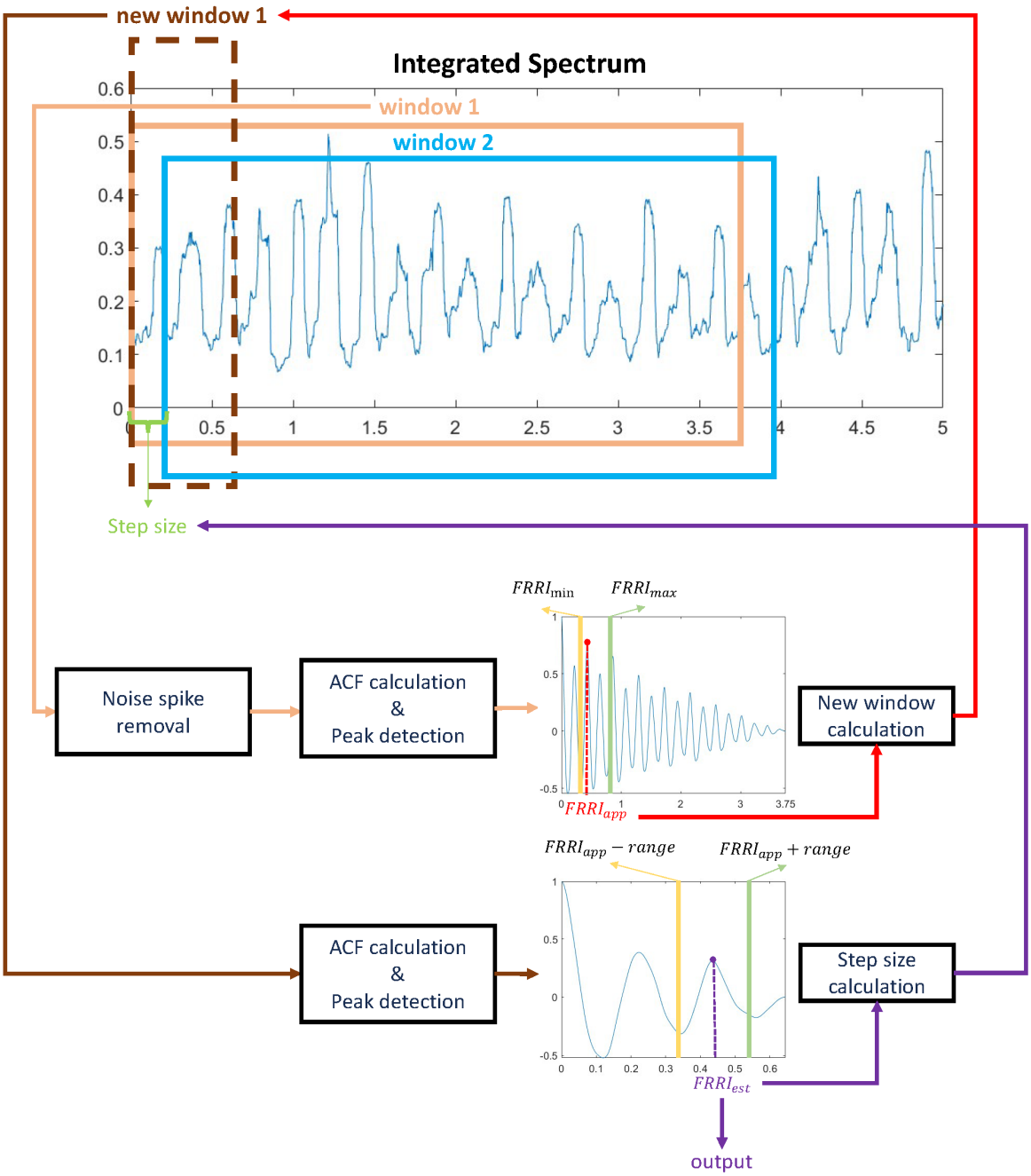


Figure 2.4. The workflow for FRRI estimation from DUS signals.

Figure 2.4 demonstrates the workflow for FRRI estimation from integrated spectrum data. This figure only presents the workflow for estimating FRRI in window 1 and determining window 2. The FRRI estimation procedure for each following window follows this workflow. To calculate the FRRI from the integrated spectrum data, the ACF of an integrated spectrum segment is calculated using a window of length S, which is set to 3.75 s in This

study. The selection of a 3.75 s window was based on its standard usage in the computerized analysis of fetal non-stress tests [3]. This window length has been widely adopted and proven to yield consistent and reliable results in related studies. Before calculating the ACF, it is necessary to minimize the effects of noise spikes. To achieve this, this study employ a spike removal algorithm, as introduced in [93], which effectively eliminates these undesirable noise spikes. In the calculated ACF, the maximum peak is detected from the range $[FRRI_{min}, FRRI_{max}]$, which is assumed to be a reasonable range of FRRIs. To ensure that the peaks are not harmonic, an additional procedure is performed if there are more than two distinct peaks within a window. The value of $FRRI_{min}$ and $FRRI_{max}$ is set to be the same as the optimized values in [3]. This study also use the algorithm proposed by [3] to identify whether a peak is harmonic and to determine the prominent peak. The time (the horizontal axis of the ACF) of the detected peak of the ACF can be considered as an approximate estimate $FRRI_{app}$ in the window of length W. To determine the FRRI for each heartbeat, this study use the approximate estimate $FRRI_{app}$ to adjust the window width for the second calculation of the ACF. The adjusted window length S_{new} is defined as shown in Equation (2.3), which includes two heartbeats.

$$S_{new} = 2 \times FRRI_{app} + \Delta, \tag{2.3}$$

where Δ in this case is set to $1/2 \times FRRI_{app}$. Using the adjusted window, this study apply the ACF again and use the same procedure as before to detect the peak in the ACF to derive the FRRI. Different from the peak detection of the first ACF, in the second ACF, the maximum peak is detected within the range $[FRRI_{app} - range, FRRI_{app} + range]$. The detected peak of the second ACF is the estimated FRRI in the new window, called $FRRI_{est}$. As the objective of the second ACF calculation and peak detection is to further refine the approximate estimate $FRRI_{app}$, it is expected that the value of $FRRI_{est}$ should be in close proximity to $FRRI_{app}$. After conducting several experiments, a value of 0.1 s is chosen as the value of $range$ to ensure that $FRRI_{est}$ remains around $FRRI_{app}$. After estimating the $FRRI_{est}$ in this window, the window length S is initialized as 3.75 s for the

next window. In addition, the window is shifted for half the time of the previous $FRRI_{est}$. Each subsequent window goes through the same operation to estimate FRRIs.

2.4.3 SQA

As shown in Figure 2.2, there are two procedures, including representation learning based on VAE and combined SQI generation based on SOM. In the proposed work, representation learning from the integrated spectrum is on the basis of an FCN-based VAE neural network. The integrated spectrum data are slid through a 1.2 s window (average value of resized windows after the first ACF for FRRI estimation). Each time the window is slid, the window is moved to align the start time stamps of the resized window recorded during FRRI estimation. As shown in Figure 2.5, the input vector $\boldsymbol{x} = [x_1, x_2, ..., x_L]$ is an integrated spectrum segment that has been resampled to 1024 samples, while the output vector $\hat{\boldsymbol{x}} = [\hat{x}_1, \hat{x}_2, ..., \hat{x}_L]$ represents a reconstructed integrated spectrum segment. In addition, the VAE neural network is a slight modification of an AE neural network used for noise reduction in ECG signals [94], where there are several FCN layers in both the encoder and decoder parts. Every FCN layer comprises three key components: (1) convolutional layer, (2) batch normalization layer, and (3) activation layer based on the exponential linear unit (ELU) function. Together, these components form the building blocks of each FCN layer, enabling effective feature extraction and non-linear transformations within the network. Furthermore, two dense layers and a sampling function are used to link the encoder and decoder. The mean and variance of the latent vector $\boldsymbol{z}$, called $\boldsymbol{\mu_z}$ and $\boldsymbol{\sigma_z^2}$, are generated by the dense layers. The sampling function is then used to generate $\boldsymbol{z}$ from $\boldsymbol{\mu_z}$ and $\boldsymbol{\sigma_z^2}$. In the encoder neural network, after five convolutional layers with a stride of 2, the input vector $\boldsymbol{x}$ is compressed to a vector of dimension 32×40. Using the dense layers and the sampling function, $\boldsymbol{\mu_z}$, σ_z^2, $\boldsymbol{z}$ are all of dimension 32×1. The VAE decoder takes the latent representation vector $\boldsymbol{z}$ as input and generates the reconstructed vector $\hat{\boldsymbol{x}}$ as output. This study uses transposed convolutional layers throughout the decoder neural network for upsampling. All convolutional and transposed convolutional layers have a kernel size 16, and all batch normalization layers have a batch size 64.

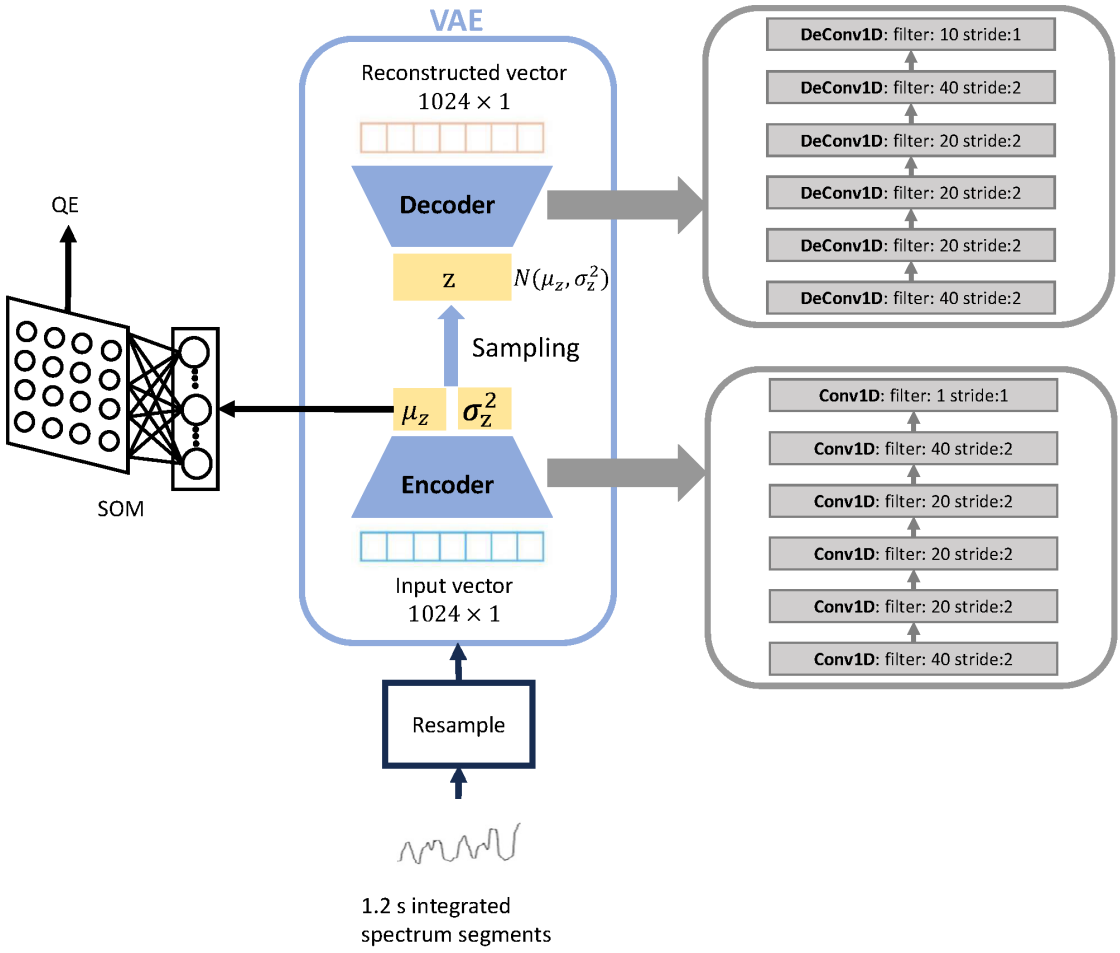


Figure 2.5. The proposed VAE and SOM-based SQA method for DUS signals.

Since the purpose of VAE training is to minimize reconstruction errors and enable $\boldsymbol{z}$ to approximate standard normal distribution $\mathcal{N}(0, \boldsymbol{I})$ as much as possible, the VAE loss function is given as Equation (2.4):

$$Loss = \frac{1}{L}\sum_{l=1}^{L}(\boldsymbol{x}_l - \hat{\boldsymbol{x}}_l)^2 + KLD_{\mu_z,\sigma_z^2}, \tag{2.4}$$

where L is the length of the input vector $\boldsymbol{x}$ and the reconstructed vector $\hat{\boldsymbol{x}}$, KLD_{μ_z,σ_z^2} is KLD between $\mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z^2)$ and $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, which is given as follows:

$$KLD_{\mu_z,\sigma_z^2} = \frac{1}{2}\sum_{i=1}^{d}(\boldsymbol{\mu}_{z(i)}^2 + \boldsymbol{\sigma}_{z(i)}^2 - \log \boldsymbol{\sigma}_{z(i)}^2 - 1), \tag{2.5}$$

where the dimension of $\boldsymbol{z}$ is d. After 100 training epochs, the latent representation vector $\boldsymbol{z}$ captures significant information about fetal cardiac activities, facilitating precise reconstruction of the input data.

By the trained FCN-based VAE neural network, the integrated spectrum segments are processed to the representation vectors with the dimension of 32×1. $\boldsymbol{\mu}_z$ is input to the SOM to compute the combined SQI (means combining a 32×1 vector to a single value), which is introduced as follows. Considering that high-quality DUS signal segments make up more than 80% of all segments in the proposed dataset, it is important to note that there is a significant imbalance between these two types of low- and high-quality segments. This imbalance allows the identification of low-quality segments as a task for outlier detection. Within the trained SOM neural network, the outliers exhibit considerable distance from their corresponding winning neuron, in stark contrast to the majority of samples which are positioned in close proximity to their respective winning neurons. Therefore, for detecting outliers, this study use QE as the output of the SOM neural network and calculate the combined SQI based on QE.

The combined SQI is defined based on the QE. A QE value can be computed for each integrated spectrum segment by feeding it into the trained SOM. Then, the QEs are normalized in the range [0, 1] based on the min–max normalization algorithm. The combined SQI based on the normalized QE is defined as follows:

$$SQI(\boldsymbol{x}) = \begin{cases} 1, & QE(\boldsymbol{\mu}_z(\boldsymbol{x})) <= th, \\ 1 - \frac{QE(\boldsymbol{\mu}_z(\boldsymbol{x})) - th}{1 - th}, & QE(\boldsymbol{\mu}_z(\boldsymbol{x})) > th, \end{cases} \tag{2.6}$$

where $SQI(\boldsymbol{x})$ is the combined SQI generated from the input vector of VAE $\boldsymbol{x}$, th is a threshold used to determine the quality of data that can be used to generate acceptably accurate FHR estimates. th is specifically set to the 9-th decile of all SQI values. This choice is based on the observation that approximately 90% of the segments within the proposed dataset exhibit high quality. For different datasets, th can be set to different

values. The combined SQI has a range of [0, 1], and the larger value corresponds to the higher quality.

2.4.4 FRRI Estimation Refinement

KF is a common and effective tool for estimating the optimal state of stochastic signals [95]. Inspired by previous studies [83, 84], a KF is applied to the rough FRRI estimates generated from the FRRI estimation block explained in Section 2.4.2 to further rectify the FRRI estimation. The proposed model is defined as a first-order autoregressive process, where the time series estimates are determined by the linear regression of the preceding measurement, establishing a relationship between the current estimate and its previous measurement. In this process, there are two noise metrics: process noise $\boldsymbol{W} \in \mathbb{R}^n$ and measurement noise $\boldsymbol{V} \in \mathbb{R}^n$, where $\boldsymbol{W} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$ and $\boldsymbol{V} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$. In addition, this study set the coefficient state transition metrics to be unitary. To merge the combined SQI into the KF, this study modify the measurement noise covariance metric $\boldsymbol{R}_m$ using the non-linear weighting function [83] based on the combined SQI, as shown in Equation (2.7). The measurement $\boldsymbol{Z}_m$ should be less trusted when the combined SQI is low.

$$\boldsymbol{R}_m = \boldsymbol{R}_0 \cdot e^{\frac{1}{SQI_m^2}-1}, \tag{2.7}$$

where SQI_m refers to the m-th combined SQI, while $\boldsymbol{R}0$ represents the initial value of the measurement noise covariance. As $SQIm$ approaches 0, indicating low quality, $\boldsymbol{R}m$ tends towards infinity. This adjustment serves to decrease the Kalman gain $\boldsymbol{K}m$, leading to reduced reliance on the current measurement compared to the previous measurement. Conversely, $\boldsymbol{R}_m$ approaches $\boldsymbol{R}_0$ while SQI_m approaches 1, and $\boldsymbol{K}_m$ is increased to reflect greater reliance on the current measurement. By incorporating this adaptive mechanism, the algorithm assigns less trust to measurements with lower quality, prioritizing the more reliable information for improved estimation accuracy.

This study first use a conventional KF with fixed $\boldsymbol{R}_m$ and $\boldsymbol{Q}_m$ to smooth the rough FRRI estimates, where this study calibrate $\boldsymbol{R}_0 = 1$ and $\boldsymbol{Q}_0 = 0.1$. In addition, a KF with changing $\boldsymbol{R}_m$ based on Equation (2.7) is applied to refine the FRRIs with relatively large

errors which are estimated from low-quality DUS segments. Through calibration, this study assign $\boldsymbol{R}_0 = 1$ and $\boldsymbol{Q}_0 = 1$ for the second KF. In addition, this study then eliminate the FRRIs from where the low-quality segments are continuous since it is difficult to correct FRRIs generated from such signals. To detect low-quality segments, a combined SQI threshold is established, specifically the 8th decile of all combined SQI values. Any FHR estimates obtained from more than three consecutive segments of low quality are discarded.

2.5 Results and Discussion

2.5.1 Experimental Setup

The DUS recordings and invasive fetal ECG recordings of ten subjects are collected simultaneously by Atom Medical Corporation. The fetal R-peaks are annotated by the experts to obtain the ground truth of the FRRIs and FHRs. The relationship between an FHR and an FRRI is given by

$$FHR_i = \frac{60 \cdot F_s}{FRRI_i} \tag{2.8}$$

where F_s is the sampling frequency of the DUS signal recordings. For each subject, there is a DUS recording of 60 s with a sampling frequency of 1 kHz. As for VAE, this study used the Adam optimizer and trained for 100 epochs. In addition, to achieve optimal performance of the SOM training, this study experimented with different values of each parameter while keeping the other parameters constant. A 30×30 output space and the Gaussian neighborhood function are used to train the SOM neural network for 15,000 iterations. The number of training samples for VAE and SOM neural network is 2466, where each training sample is a 1.2 s integrated spectrum, as mentioned in Section 2.4.3.

To evaluate the performance of the proposed robust FHR estimation method, three performance metrics are considered, which are given below:

- Root mean square error (RMSE): RMSE is calculated between the estimated value and the ground truth value of FRRI, which is calculated as follows:

$$RMSE = \sqrt{\frac{1}{\Gamma}\sum_{i=1}^{\Gamma} \|F\hat{R}RI_i - FRRI_i\|_2^2}, \tag{2.9}$$

- Averaged absolute error (AAE): AAE is calculated between the estimated value and the ground truth value of FHR, which is calculated as follows:

$$\begin{aligned} AAE &= \frac{1}{\Gamma}\sum_{i=1}^{\Gamma} \|F\hat{H}R_i - FHR_i\| \\ &= \frac{1}{\Gamma}\sum_{i=1}^{\Gamma} \|\frac{60 \cdot F_s}{F\hat{R}RI_i} - \frac{60 \cdot F_s}{FRRI_i}\|, \end{aligned} \tag{2.10}$$

- Coverage: Since some rough FRRI estimates may be removed based on the results of the SQA, the ratio of reserved FRRI estimates to all FRRI estimates is a critical indicator of whether as many FRRI estimates as possible have been retained.

This study compared these performance metrics in the following five scenarios:

- Use the method proposed by Valderrama et al. [3].

- FHR estimation only: The study performed only two steps to obtain rough FRRI estimates, including preprocessing and FRRI estimation (described in Section 2.4.1 and Section 2.4.2).

- Remove unreliable FRRIs: The study eliminate the unreliable rough FRRIs if the corresponding combined SQI is less than the 8th decile of all combined SQI values.

- Use conventional KF: A conventional KF with a fixed noise covariance metric was applied to the rough FRRI estimates.

- Proposed method: The four steps described in Section 2.4 are all used to estimate and refine FRRIs.

2.5.2 Experimental Results

The changing pattern of training loss throughout the training process of the FCN-based VAE (which is introduced in Section 2.4.3.) is illustrated in Figure 2.6. In this thesis, the proposed model operates as a reconstruction model employing the principles of a VAE. The experimental results reveal a compelling trajectory in the training loss throughout 100 epochs. From 1 to 100 epochs, the training loss consistently decreases, underscoring the proficiency of the model in learning and capturing intricate patterns within the dataset. As a reconstruction model based on VAE, this observed reduction in loss attests to the success of the model in encoding and subsequently reconstructing input data.

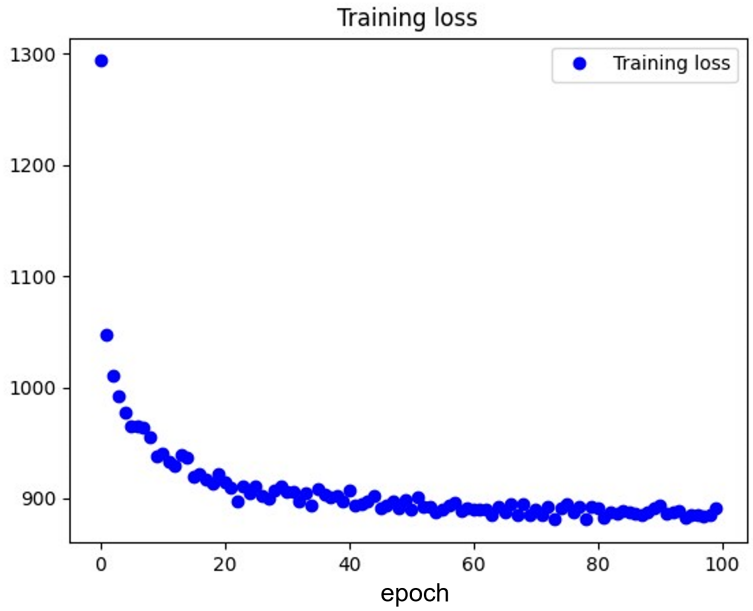


Figure 2.6. The training loss of the FCN-based VAE training process for DUS SQA.

To demonstrate the performance of the combined SQI, inspired by [96], this study defined a normalized absolute error (NAE) calculated as Equation (2.11):

$$NAE = \frac{|FRRI_{est} - FRRI_{true}|}{FRRI_{true}}, \tag{2.11}$$

where $FRRI_{true}$ and $FRRI_{est}$ represent the estimated FRRI and the ground truth of FRRI. By establishing a threshold, this study broadly categorize all integrated spectrum segments into high-quality and low-quality. In the proposed experiments, by manually observing low-quality and high-quality segments using different thresholds, this study selected the most appropriate threshold of 0.03. It is important to note that the classification into high and low quality is not based on expert annotations but rather roughly annotated based on the assumption that "if the signal quality is low, the FRRI prediction is inaccurate". Figure 2.7 illustrates the combined SQI corresponding to high-quality and low-quality segments. Since the number of low-quality segments is significantly less than that of high-quality segments, this study randomly selected an equivalent number of low-quality segments from high-quality signals for illustration. From the figures, it is evident that the majority of combined SQI values for high-quality segments are equal to or close to 1, while some low-quality segments exhibit considerably lower combined SQI values. Consequently, this study can affirm that the proposed proposed combined SQI is effective in distinguishing between high-quality and low-quality signals.

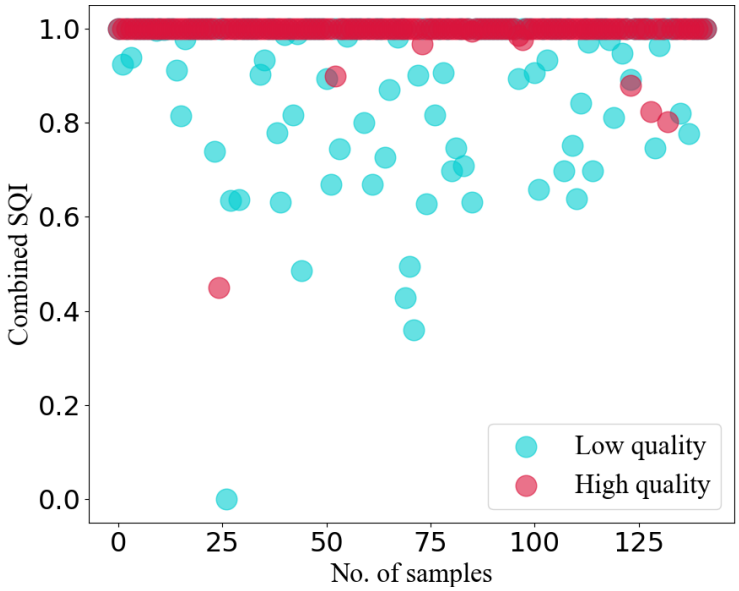


Figure 2.7. The combined SQI values of the low- and high-quality integrated spectrum segments processed from raw DUS signals.

In Figure 2.8, this study present an example of the pre-processed DUS signal segment, and the combined SQI generated from this segment (20–25 s of subject 9). The figure illustrates the presence of noise and relatively high amplitudes in both the filtered DUS

signal and the integrated spectrum within the time range of 22.5–23.5 s. Consequently, the estimated FRRI based on such a signal segment is highly unreliable, leading to significantly lower composite SQIs compared to the pre- and post-periods. This observation strongly suggests that the combined SQI is closely associated with the quality level of the DUS signals. The figure highlights the importance of signal quality assessment in accurately estimating the fetal heart rate and emphasizes the role of the combined SQI as an indicator of DUS signal reliability.

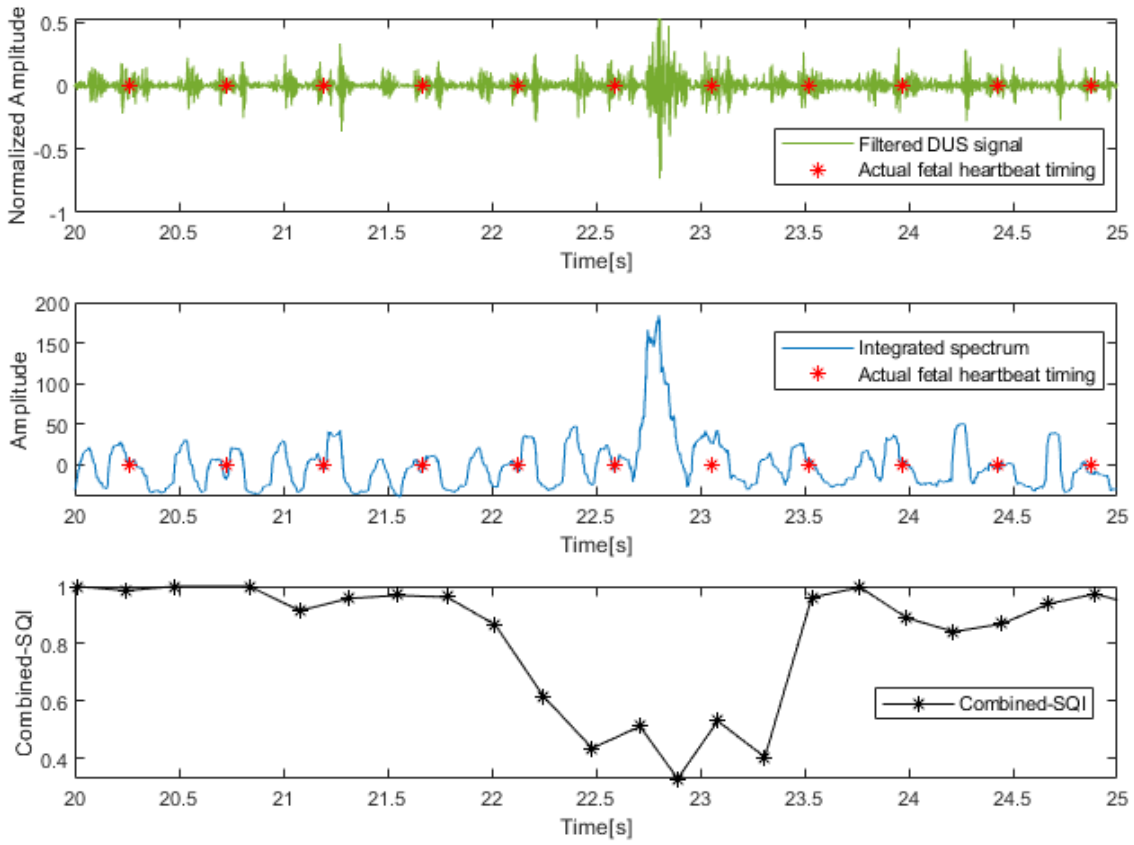
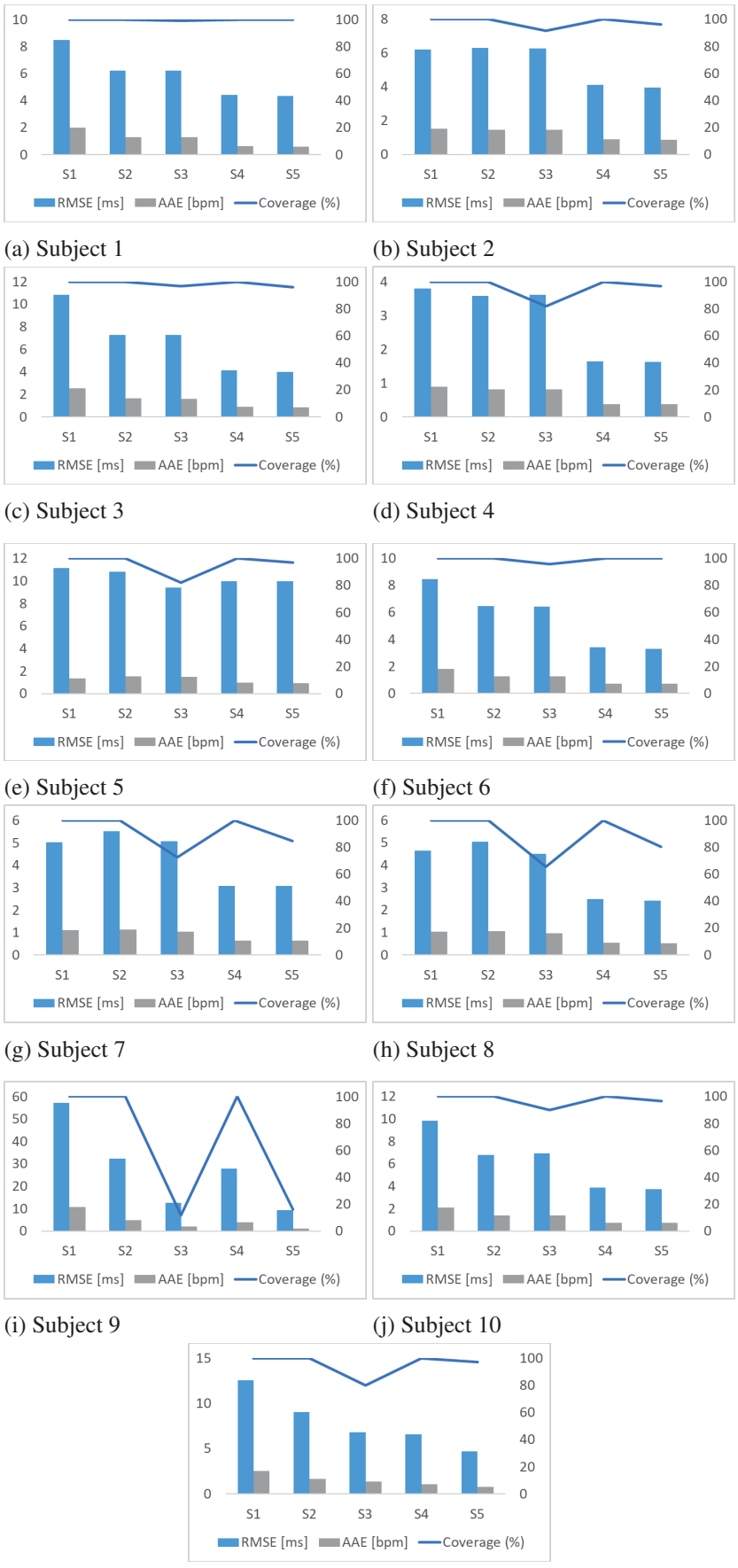


Figure 2.8. An example: The filtered DUS signal, the integrated spectrum, and the combined SQI values (from top to bottom) of a 5 s signal segment (20 s–25 s of subject 9).

Table 2.1 and Figure 2.9 provide insights into the RMSE of FRRI, AAE of FHR, and coverage in five scenarios: S1: Use the method proposed by Valderrama et al. [3]; S2: FHR estimation only; S3: Remove unreliable FRRIs; S4: Use conventional KF; S5: Proposed method. In each sub-figure, the y-axis on the left corresponds to the values of RMSE and MAE, and the one on the right corresponds to the values of coverage. It is evident that without SQA, the proposed FRRI estimation method outperforms the method

proposed by Valderrama et al. [3] for the majority of subjects. However, there are a few exceptions (e.g., subjects 7 and 8) where the proposed method may not yield better results. The other three scenarios considered in the analysis also contribute to the reduction of estimation errors to varying extents. In the "Remove unreliable FRRIs" scenario, the RMSE and AAE generally decrease, except for subjects 1, 4, and 10. However, it is important to note that this scenario leads to a significant reduction in coverage, especially for subject 9. In the "Use conventional KF" scenario, the FHR estimation errors are reduced while maintaining 100% coverage. This is achieved by applying a conventional KF that helps in smoothing out the rough FRRI estimates and handling sharp changes in the data. Comparing all the scenarios, the proposed proposed robust FHR estimation method demonstrates the lowest RMSE and AAE. Additionally, the proposed method maintains higher coverage for each subject compared to the conventional approach of removing unreliable FRRIs, which is commonly used for signal quality assessment tasks. Moreover, the experimental results validate the effectiveness of the proposed proposed technique in enhancing FHR prediction across various scenarios. Notably, the proposed method showcases significant improvements for recordings encompassing a substantial proportion of low-quality data (e.g., subject 9), as well as for recordings containing a smaller fraction of low-quality data (e.g., subject 1). These results highlight the robustness of the proposed approach, showcasing its ability to deliver reliable FHR predictions across recordings with varying levels of data quality.



(a) Subject 1

(b) Subject 2

(c) Subject 3

(d) Subject 4

(e) Subject 5

(f) Subject 6

(g) Subject 7

(h) Subject 8

(i) Subject 9

(j) Subject 10

(k) Average

Figure 2.9. Comparison of the DUS-based FHR estimation performances of five scenarios for each subject.

Table 2.1. Comparison of the DUS-based FHR estimation performances of five scenarios for each subject.

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Average**
Valderrama et al. [3]	RMSE [ms]	8.46	6.21	10.81	3.80	11.12	8.46	5.02	4.64	57.10	9.83	12.54
	AAE [bpm]	1.98	1.51	2.54	0.90	1.34	1.80	1.11	1.04	10.79	2.09	2.51
	Coverage (%)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
FHR estimation only	RMSE [ms]	6.20	6.30	7.26	3.58	10.83	6.48	5.52	5.03	32.36	6.78	9.03
	AAE [bpm]	1.27	1.47	1.63	0.82	1.55	1.27	1.14	1.06	4.89	1.39	1.65
	Coverage (%)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Remove unreliable FRRIs	RMSE [ms]	6.22	6.27	7.25	3.62	9.42	6.43	5.06	4.50	12.62	6.90	6.83
	AAE [bpm]	1.27	1.45	1.61	0.82	1.48	1.26	1.04	0.97	2.06	1.40	1.34
	Coverage (%)	99.23	91.41	96.87	81.93	95.18	95.53	72.54	65.52	11.95	89.75	79.99
Use conventional KF	RMSE [ms]	4.42	4.12	5.15	1.64	9.96	3.40	3.07	2.49	27.77	3.87	6.59
	AAE [bpm]	0.63	0.90	1.01	0.37	0.96	0.73	0.64	0.54	3.83	0.74	1.03
	Coverage (%)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Proposed method	RMSE [ms]	4.35	3.97	5.00	1.62	9.95	3.31	3.07	2.41	9.22	3.75	4.67
	AAE [bpm]	0.59	0.86	0.98	0.37	0.94	0.72	0.63	0.52	1.07	0.72	0.74
	Coverage (%)	100.00	96.06	100.00	96.77	100.00	100.00	84.71	80.35	16.00	96.69	87.06

2.5.3 Limitation and Future Works

Several limitations still exist, although the proposed proposed FHR estimation approach can accurately estimate FHR from DUS signals and maintain robustness under different signal qualities. To train VAE and SOM neural networks for the DUS SQA part, this study use all signal segments for training without a selection procedure. In fact, it is worth mentioning that this strategy only works if the signal segments are mostly of high quality, in other words, low-quality segments can be considered as outliers. If low-quality signal segments account for more than 40% of a dataset, it is better to use some simple quality indicators to roughly select high-quality data for training (e.g., entropy, skewness, or kurtosis). In addition, this study specifically developed and tested the proposed methods using only one dataset consisting of a small number of subjects, optimizing parameters for this dataset without retaining data for validation, so it is not possible to verify that the developed algorithms would be generalizable to the larger population of subjects studied. For the proposed future work, some larger datasets are needed to further verify the robustness.

2.6 Conclusions

This study present a robust DUS-based FHR estimation method assisted by DUS SQA based on unsupervised representation learning. By using STFT to generate integrated spectrum data, information on DUS signals in both time and frequency domains is extracted simultaneously. To achieve accurate FRRI estimation, this study employ two ACF calculations on the integrated spectrum data. The first ACF is utilized for approximate FRRI estimation and window resizing, while the second ACF provides a more precise FRRI estimation. The length of the FRRI estimation window is adjusted based on the previously estimated FRRI, enabling adaptability to varying heart rates. In addition, a method is proposed a DUS SQA method that uses an FCN-based VAE to obtain learning representations and a SOM neural network for combined SQI generation. The combined SQI is used to modify the measurement noise covariance, and a KF is then used to refine the FRRIs estimated from low-quality signals. Through extensive experiments, it is demonstrated the superiority of the proposed proposed approach in terms of accuracy and robustness for DUS-based FHR estimation, surpassing conventional methods. Moreover, the proposed SQA approach can be extended to address various SQA challenges involving different types of signals, as it is built entirely on unsupervised learning techniques.

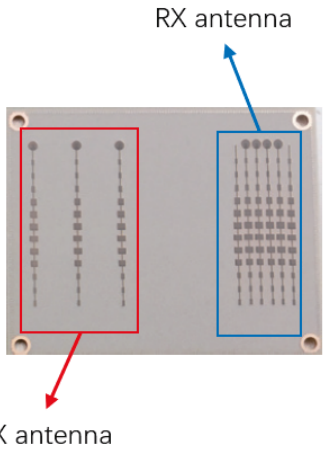
Chapter 3

Rough-to-Fine Model-based Non-contact Heart Rate Estimation using MIMO FMCW Radar

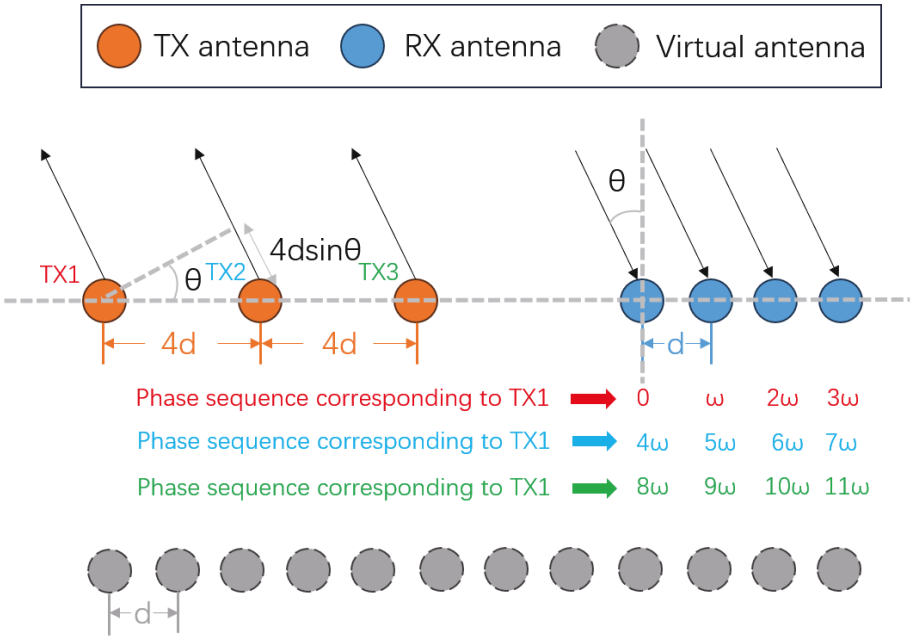
3.1 Introduction

Heart rate (HR) estimation is critical for assessing the health conditions of the patients and the old, where persistence and accuracy are the two most important metrics [97]. HR can be accurately estimated from signals like electrocardiograms (ECG) [31–36, 98] and photoplethysmography (PPG) [99–102]. However, both ECG and conventional PPG techniques rely on contact-based sensors, electrodes or photodetectors, which are less conducive to long-term monitoring and lack persistence. Thus, non-contact HR monitoring techniques have been widely used in recent years, including camera-based systems [97], laser vibrometry-based systems [103–105], Doppler radar-based systems [106–109], and Frequency Modulated Continuous Wave (FMCW) Radar-based systems [61–64, 71, 72, 110–112]. Compared to radar-based systems, the camera-based systems require clear visibility and address privacy and lighting concerns, while laser vibrometry-based systems may suffer from obstructions or complex surface vibrations. FMCW Radar identifies the human location for precise heartbeat signal extraction, contrasting with Doppler Radar, which captures both signal and background noise, affecting signal quality. FMCW Radar offers superior range resolution and accuracy, capable of measuring range and velocity, making it ideal for vital sign detection in diverse environments.

FMCW radar systems are differentiated by their configurations of antennas as Single-Input Single-Output (SISO), Single-Input Multiple-Output (SIMO), and Multiple-Input Multiple-Output (MIMO). The MIMO FMCW radar employs multiple transmitting (TX) and receiving (RX) antennas to enhance range resolution and angle estimation. This multi-antenna setup supports the creation of virtual arrays that can improve target and vital sign detection [113]. The proposed research utilizes a 3TX and 4RX MIMO FMCW Radar, detailed in Figure 3.1, where the signal from each TX antenna induces a unique phase sequence across the RX antennas, effectively generating a system with one TX and twelve virtual RX antennas. Regarding the FMCW Radar-based systems for heart rate estimation mentioned earlier, the research [61, 62, 72, 110] employed SISO FMCW Radar. Conversely, the studies in [63, 64, 71, 111, 112] utilized MIMO FMCW Radar.



(a) MIMO FMCW mmWave radar device (S-Takaya T14_01120112_2D) used for indoor HR estimation.



(b) Principle of MIMO FMCW Radar (4 RX, 3 TX)

Figure 3.1. The MIMO FMCW Radar used in the proposed experiments and the principle of MIMO FMCW Radar.

In MIMO FMCW radar applications for HR estimation, there are two significant steps: (i) person position estimation and (ii) HR estimation. Conventional person position estimation methods, which rely on range-angle maps and signal power thresholds using the CFAR algorithm [114], can suffer from inaccuracies due to radar positioning, environmental factors, and multipath interference. To overcome these limitations, K. Han et al. [115] proposed the Curve-Length (CL) method, analyzing unique motion patterns in IQ trajectories to better distinguish human presence from static objects. K. Endo et al. [116]

further improved detection by assessing the correlation of IQ signal curve lengths across the radar coverage area, thereby enhancing accuracy by effectively minimizing noise and interference effects in human location detection.

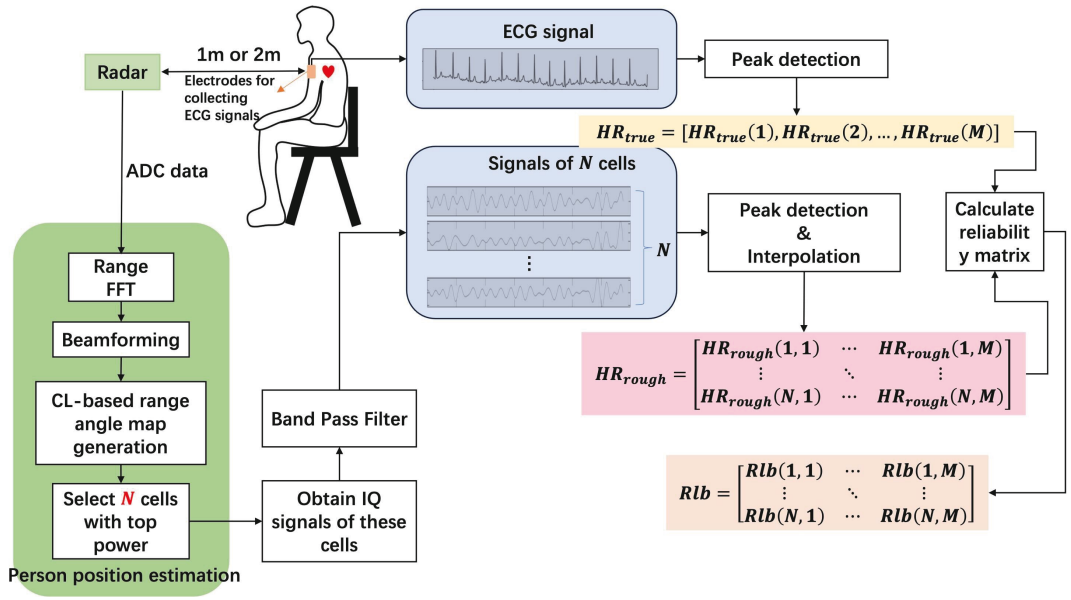
As for HR estimation methods using FMCW Radar, the papers [71, 110, 111] conventionally extract heartbeat signal directly from the phase difference signal. L. Liu et al. [63] propose a signal decomposition module applied to the phase difference signal using wavelet packet decomposition to isolate respiration and heartbeat signals, reducing respiration harmonics and heartbeat signal noise. Moreover, they introduce a vital sign rate reconstruction module employing an adaptive filter and a zero attracting sign exponentially forgetting least mean square algorithm for sparse spectrum reconstruction. This approach enhances the resolution of the sparse spectrum, facilitating precise and dependable HR estimation. T. K. V. Dai et al. [112] obtain vital sign data including heartbeat signal through arctangent demodulation and maximal ratio combining, enhanced with an adapted-wavelet continuous wavelet transform. Additionally, there are two papers [62,64] leveraging variational mode decomposition (VMD)-based method for heartbeat signal extraction. K. J. Wu et al. [62] utilize conventional VMD-based method, while L. Qu et al. [64] introduced an improved adaptive parameter variational mode decomposition (IAPVMD) algorithm which can dynamically adjust the mode number and penalty coefficient for VMDusing the energy loss rate and mode discrimination results. The works proposed by [61, 72] integrate deep learning (DL)-based method assisting HR estimation. The paper [72] proposes a DL-aided Newtonized Orthogonal Matching Pursuit (NOMP) technique to tackle the issue that the effectiveness of NOMP significantly declines under low signal-to-noise ratio conditions [117]. Another DL-based method is introduced by [61], where a DL-aided method leverages the spatial features of the spectrum for quick and accurately satisfying vital sign estimation outcomes. In their paper, H.-Y. Chang et al. detect the human target, select five bins with the highest power from the range profile, and extract key features (magnitude, phase, frequency response) for input into a DL neural network. The goal is to accurately predict weights for these bins to achieve precise

heart rate estimates. Recent studies [63, 64] have demonstrated high accuracy in HR estimation, showing potential for real-world application. However, these approaches rely on lengthy time windows (exceeding 10 s) for HR calculation, compromising their real-time capabilities. Moreover, the sensitivity of algorithm parameters to conditions may affect the robustness of these systems in practical settings. The DL-based technique introduced by H.-Y. Chang et al. [61] addresses these challenges but is constrained by the use of SISO FMCW Radar, which limits range resolution. For accurate HR estimates, the subject must be positioned close and directly in front of the radar. Furthermore, although the method proposed in [61] can shorten the time window for HR estimation, the context information (the information from the previous and post windows) can not be utilized which is crucial for the time series model. To overcome these limitations, the study introduce an idea that combines a CL method and a DL rough-to-fine approach with MIMO FMCW Radar, aiming to improve the HR estimation accuracy and applicability of the previous DL-based model [61].

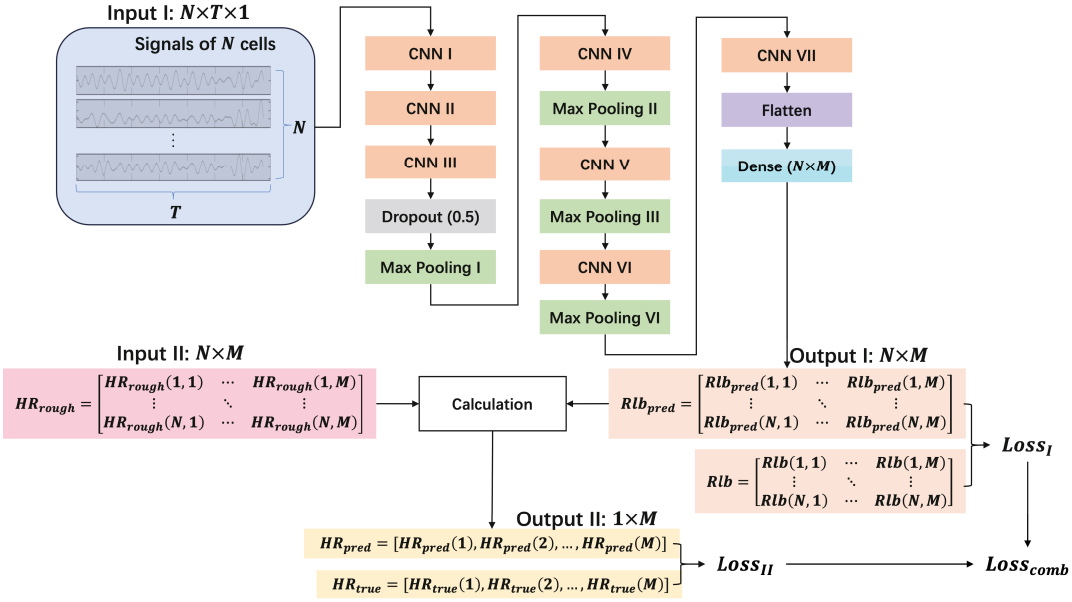
In this study, a method is proposed a rough-to-fine model-based non-contact HR estimation, employing MIMO FMCW Radar technology. As illustrated in Figure 3.2, the proposed process begins by detecting the location of human subject through a CL-based approach [116]. The study then select several cells within the CL-based range-angle to isolate heartbeat-related signals. The HR estimation comprises two phases: rough HR estimation and fine HR estimation. The rough estimation relies on straightforward peak detection, whereas the fine estimation employs a dual-input, dual-output CNN model emphasizing smaller sub-windows for a richer context. With a combination of feature extraction and novel use of dual loss functions focused on reliability and accuracy, the proposed model more precisely predicts HR. The proposed experiments demonstrate that the integrated approach not only achieves higher accuracy but also exhibits higher robustness than the conventional methods.

This chapter is structured as follows: The proposed method is introduced in Section 3.2. The experimental setup and the results obtained are detailed in Section 3.3. The chapter concludes with Section 3.4.

3.2 Proposed Method



(a) Person position estimation & Rough HR estimation



(b) CNN-based fine HR estimation

Figure 3.2. The overall structure of the proposed rough-to-fine HR estimation method using MIMO FMCW mmWave radar.

Figure 3.2 illustrates the comprehensive framework of the proposed proposed methodology, which is segmented into two principal phases as follows.

3.2.1 Person Position Estimation & Rough HR estimation

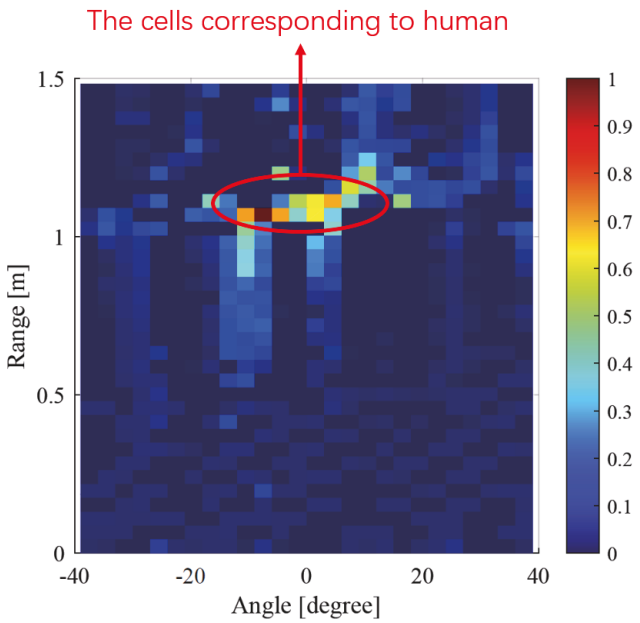


Figure 3.3. Range-angle map generated from raw radar ADC data by the CL method.

In the initial phase of human detection and localization, the estimation of human positions within the effective area of FMCW Radar is achieved by leveraging techniques from prior research [116]. The process begins with the application of the range Fast Fourier Transform (FFT) to the radar signals received from each beam direction. This is done in a sliding window with a step size of 2 s and a time length of 10 s or 15 s to ascertain the distance to the object. The decision on the length of the duration and the interval steps is made based on empirical evidence. The following procedures are all based on the 10 s or 15 s segments. Following this, beamforming is utilized to acquire IQ signals for a specified angular range. A range-angle map is subsequently formulated utilizing the CL method, wherein the signal power corresponds to the curve length of each IQ signal on the map. The CL method is designed to measure the variation length of the IQ signal across every identified range and angle within the Coherent Processing Interval (CPI), which is given by:

$$CL = \sqrt{\sum_{\tau=2}^{T_M} ((I[\tau] - I[\tau-1])^2 + (Q[\tau] - Q[\tau-1])^2)}, \tag{3.1}$$

where T_M denotes the CPI, and $I[\tau]$ and $Q[\tau]$ represent the in-phase and quadrature signals, respectively, for a specific range, angle, and sampling time τ.

Subsequently, as illustrated in Figure 3.3, the highest power N cells identified on the range-angle map via the CL method are selected as potential indicators of human presence, where $N = 10$ in the proposed proposed method is based on empirical insights. In scenarios involving a single individual, signals from the chosen cells are utilized for HR estimation. A Band-Pass Filter (BPF) is implemented on the IQ signal of top power N cells, and the study obtain the filtered signals of top power N cells $R = [R(1), R(2), ..., R(N)]$. The cutoff frequency range is established between 0.8 to 2.0 Hz, aligning with the typical frequency range of the human HR. For the signal of each cell, the study use a sliding sub-window to calculate HR from M sub-windows. The peaks in these filtered signals are detected, and the study calculate the rough HR estimates HR_{rough} based on the detected peaks, which is given by.

$$HR_{rough}(n, m) = \frac{60}{1/P_{n,m} \sum_{p=1}^{P_{n,m}} Itv_{n,m,p}}, \tag{3.2}$$

where $P_{n,m}$ is the number of peak intervals in the m-th sub-window of the n-th cell, $Itv_{n,m,p}$ is the p-th peak interval in the m-th sub-window of the n-th cell. In order to remove outliers, for the signal of each cell, the study first calculate the mean of peak intervals and replace the peak intervals below 0.6 times the mean and above 2 times the mean with the mean.

The reliability matrix Rlb for each segment is generated based on the error of the rough HR estimates HR_{rough}. The study assume that the cell with the lowest error of HR estimation signifies the optimal cell in the m-th sub-window. The errors can be roughly calculated by using HR_{rough}. Rlb is initially set to an all-zero matrix. In the m-th sub-window, if the rough HR estimate of n-th cell $HR_{rough}(n, m)$ achieves the lowest error referring to the ground truth $HR_{true}(m)$, the reliability indicator $Rlb(n, m)$ is assigned 1. The ground truth HR_{true} is generated from the ECG signals based on peak detection method. The reliability indicator $Rlb(n, m) \in \{0, 1\}$ denotes whether the n-th cell is the optimal channel within the

m-th sub-window for HR estimation among N cells. This indicates that $Rlb(n, m)$ is determined relative to other cells in the same sub-window. Consequently, the computation of *Rlb* is not influenced by environmental factors, ensuring that *Rlb* consistently identifies the optimal cell for HR estimation in each sub-window regardless of the environmental factors, which implies the robustness of the proposed method.

3.2.2 CNN-based Fine HR Estimation

Table 3.1. The parameters of the CNN-based neural network for fine HR estimation.

CNN	Filters	Kernel size	Activation	**Max pooling**	pool size
I	1	(10,100)	ReLU	**I**	(1,2)
II	16	(10,50)	Tanh	**II**	(1,4)
III	32	(3,10)	Tanh	**III**	(1,4)
IV	32	(5,10)	Tanh	**IV**	(1,4)
V	16	(5,80)	Tanh		
VI	5	(5,50)	Tanh		
VII	1	(5,100)	Tanh		

Drawing inspiration from Chang et al. [61], a method is proposed a CNN with a dual-input and dual-output configuration designed to refine rough HR estimates HR_{rough}. This network, depicted in Figure 3.2(b), processes two types of inputs: (1) **Input I**, the filtered signals of N cells: $R = [R_1, R_2, \ldots, R_N]$; and (2) **Input II**, the initial rough HR estimates HR_{rough}. It generates two outputs: (1) **Output I**, the predicted reliability matrix Rlb_{pred}; and (2) **Output II**, the refined HR estimates HR_{pred}. Different from [61], the proposed method employs smaller sub-windows within a larger window as the inputs of the DL neural network, rather than directly utilizing smaller windows as the inputs. This strategy allows for the extraction of contextual information from these smaller sub-windows, enhancing the time series model analysis.

To predict the reliability matrix Rlb_{pred} from the filtered signals, the proposed CNN utilizes seven CNN layers for feature extraction from Input I, alongside four pooling layers that preserve essential features by retaining larger numerical values. The configuration of the CNN and max pooling layers is detailed in Table 3.1. A dropout layer with a dropout rate of 0.5 is incorporated after the first three CNN layers to prevent overfitting. The extracted features are then flattened and passed through a dense layer to predict Output I. Furthermore, Output II, HR_{pred}, is obtained by combining Output I, Rlb_{pred}, with Input II, HR_{rough}, as follows:

$$HR_{pred}(m) = \frac{1}{N}\sum_{n=1}^{N} Rlb_{pred}(n,m) \times HR_{rough}(n,m), \tag{3.3}$$

where $HR_{pred}(m)$ represents the predicted HR estimate for the m-th sub-window, and $Rlb_{pred}(n,m)$ and $HR_{rough}(n,m)$ respectively indicate the predicted reliability indicator and the rough HR estimate for the m-th sub-window in the filtered signal of the n-th cell.

Unlike the approach suggested by Chang et al. [61], which employs a single loss function based on the mean square error of the predicted HR estimates, the study utilize two separate loss functions for the proposed outputs, denoted as $Loss_I$ and $Loss_{II}$, and combine them as follows:

$$\begin{aligned} Loss_I &= BCE(Rlb_{pred}, Rlb), \\ Loss_{II} &= MAE(HR_{pred}, HR_{true}), \\ Loss_{comb} &= Weight_I \times Loss_I + Weight_{II} \times Loss_{II}, \end{aligned} \tag{3.4}$$

where $BCE(Rlb_{pred}, Rlb)$ calculates the binary cross entropy between Rlb_{pred} and Rlb, and $MAE(HR_{pred}, HR_{true})$ assesses the mean absolute error (MAE) between HR_{pred} and HR_{true}. The weights for $Loss_I$ and $Loss_{II}$ are denoted as $Weight_I$ and $Weight_{II}$, respectively, $Weight_I + Weight_{II} = 1$. This method assumes that cells with lower errors in $HR_{rough}(n,m)$ for the m-th sub-window have higher reliability indicators, $Rlb(n,m)$. Thus, $Loss_I$, which corresponds to the loss of Rlb_{pred}, trains the model to predict higher reliability values for cells with lower HR estimation errors, assisting $Loss_{II}$ in reducing the

error of the predicted HR estimates HR_{pred}, during the training process. During the training phase, the study utilize the reliability matrix Rlb to compute $Loss_I$. However, during the testing phase, the reliability matrix Rlb is not required as input of the CNN-based model, since the study only focus on $Loss_{II}$ to evaluate the final heart rate estimation performance.

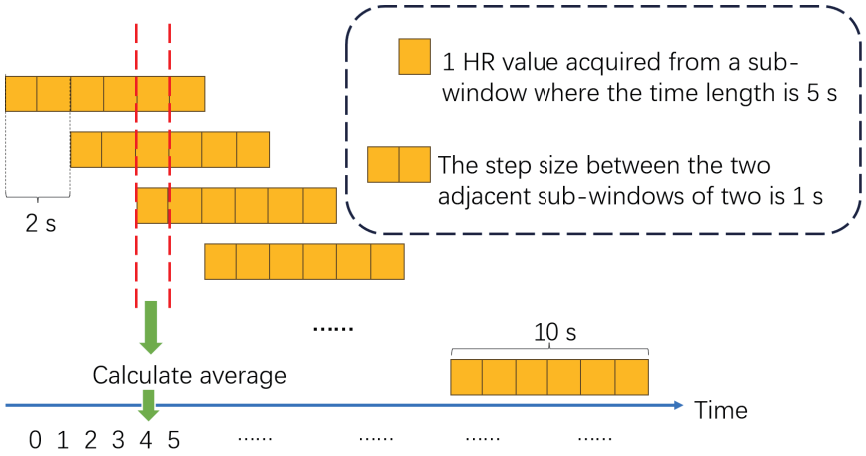


Figure 3.4. An example of how to calculate HR while using sliding windows and sub-windows.

Given that the windows used for predicting HR_{pred} overlap, it is necessary to average the HR estimates from different windows that correspond to the same time step within a recording to derive the final HR estimates. Figure 3.4 provides an illustration of this process, showing how to calculate the final HR estimates using a 10 s sliding window and a 5 s sliding sub-window, which is case 2 described in Table 3.3. In this table, "Inputs" represent "Input I" and "Input II" mentioned in Figure 3.2(b). "Outputs" represent final HR estimates.

3.3 Experiment and Results

3.3.1 Experimental Setup

MIMO FMCW Radar Settings

Table 3.2. The specifications of MIMO FMCW mmWave radar (S-Takaya T14_01120112_2D) utilized in the proposed experiment for HR estimation.

Item	Unit	Value
Module size	mm	$55 \times 45 \times 7.5$
Interfaces		USB2.0 (UART) USB3.0 (bus power)
Power supply voltage	V	5
Consumption Current	A	1.2
Frequency	GHz	$77 \sim 81$
Modulation bandwidth	GHz	3.5
Range	m	$0.04 \sim 40$
Angle	deg	± 45
Range resolution	cm	3.75
Chirp time	μs	60

Figure 3.1(a) shows the MIMO FMCW radar device T14_01120112_2D, produced by S-takaya Electronics Industry Co., Ltd., which is employed in the proposed experiment. The specifications of the MIMO FMCW radar used are detailed in Table 3.2. As illustrated in Figure 3.2(a), the radar unit is positioned in front of a seated subject at chest height. The setup maintains a distance of either 1 m or 2 m between the radar and the subject. In the proposed experiment, the study collected 18 recordings from the radar, each varying in time duration from 40 s to 180 s. These recordings are distinguished by four key variables: the subject involved, the room used for the experiment, the date of the experiment, and the distance between the subject and the radar. No two recordings share the exact same combination of these variables.

Baseline Method

Unlike the proposed method, the baseline approach involves a secondary refinement to select the top power N cells ($N = 10$), using the IQ signals from these cells for HR estimation. After identifying the top N cells, their received IQ signals are compared against the entire range-angle map to compute correlation values. Cells demonstrating a correlation above 0.5, indicative of human presence, are then highlighted in a correlation map. HR estimation cells are chosen based on these correlation measures and their variability. Specifically, cells associated with noise or multipath effects typically show standard deviations in the range of 0.2 to 0.5, while those indicating human presence have deviations between 0.02 to 0.1. To mitigate local noise impacts, the maximum correlation for each cell with its neighbors is also calculated, with human-associated cells showing correlations from 0.4 to 0.7. Cells meeting both criteria of a standard deviation below 0.2 and a maximum adjacent cell correlation above 0.4 are considered human-related. The selection of K cells for HR estimation follows, with K varying based on these criteria, where $0 <= K <= N$. The HR is not estimated in this window if $K = 0$. A BPF is applied to each IQ signal of each cell, peaks are detected, and HR estimates are calculated from these peak intervals. The final HR estimate, $HR_{baseline}$, is the average HR across these K signals, calculated as shown in Eq. (3.5).

$$HR_{baseline} = \frac{1}{K} \sum_{k=1}^{K} \frac{60}{1/P_k \sum_{p=1}^{P_k} Itv_{k,p}}, \tag{3.5}$$

where P_k is the number of peak intervals in the signal of the k-th cell, $Itv_{k,p}$ is the p-th peak interval in the signal of k-th cell. Moreover, the study use the same method described in Section 3.2.1 for removing the outliers of peak intervals for each signal.

VMD-based Method

Inspired by [64], as same as the baseline method, the study firstly obtain filtered signals of K cells. The signals undergo decomposition into ten Intrinsic Mode Functions (IMFs) via VMD, based on preliminary experimentation. The IMF that falls within the normal HR

frequency range is selected, and its energy is utilized to construct a weighted reconstructed signal $W(t)$ as follows:

$$W(t) = \frac{e_1 \times S_1(t)}{e_1 + \cdots + e_L} + \cdots + \frac{e_L \times S_L(t)}{e_1 + \cdots + e_L}, \tag{3.6}$$

where L denotes the number of selected IMF signals, $[S_1, S_2, \cdots, S_L]$ represents each IMF signal, and $[e_1, e_2, \cdots, e_L]$ represents their corresponding energies, calculated from the sum of squares of the amplitudes of each IMF signal. The HR estimation is performed using peak detection on the reconstructed signals.

Performance Evaluation

To assess the validity of the proposed proposed model, the study implemented a cross-validation strategy for performance evaluation. The study created 18 subsets for training and testing purposes, where each subset is composed of one recording as the testing set and the remaining recordings as the training set. The models were trained over 50 epochs for each subset. The accuracy of the HR estimation method is evaluated based on the MAE between the HR estimates and the actual ground truth values generated from the ECG signals. To conduct a thorough comparison between the proposed proposed method, the baseline method, and the VMD-based method, the study established four distinct cases, each with varying configurations for the inputs and outputs of the HR estimation algorithms. Since the baseline method and the VMD-based method cannot use a sub-window, they can only be evaluated in case 1 and case 3.

Table 3.3. The descriptions of 4 cases using for the MIMO FMCW mmWave radar-based HR estimation experiments.

	case	1	2	3	4
Settings of Inputs	T: Time length (s) of window	10	10	15	15
	M: Number of sub-windows in one window	1	6	1	11
Settings of Outputs	Time length (s) of sub-window	10	5	15	5
	Step size (s) of window	2	1	2	1

3.3.2 Results and Discussion

Selection of $Loss_I$

To determine the optimal weights for the two loss functions, the study evaluated the MAE across varying values of $Weight_I$. According to the results presented in Table 3.4 and Figure 3.5, the performance peaks when $Weight_I$ is set to 0.6. Consequently, the study adopt $Weight_I = 0.6$ and $Weight_{II} = 0.4$ for subsequent experiments.

Table 3.4. Comparision of the errors (MAE (bpm)) of HR estimation using MIMO FMCW mmWave radar while using different $Weight_I$.

	Weight of $Loss_I$					
	0.00	0.20	0.40	0.60	0.80	1.00
case 1	3.00	2.91	2.92	2.87	2.86	4.84
case 2	3.87	3.79	3.78	3.78	3.78	5.42
case 3	2.51	2.33	2.38	2.32	2.37	5.94
case 4	3.85	3.80	3.80	3.78	3.77	5.71

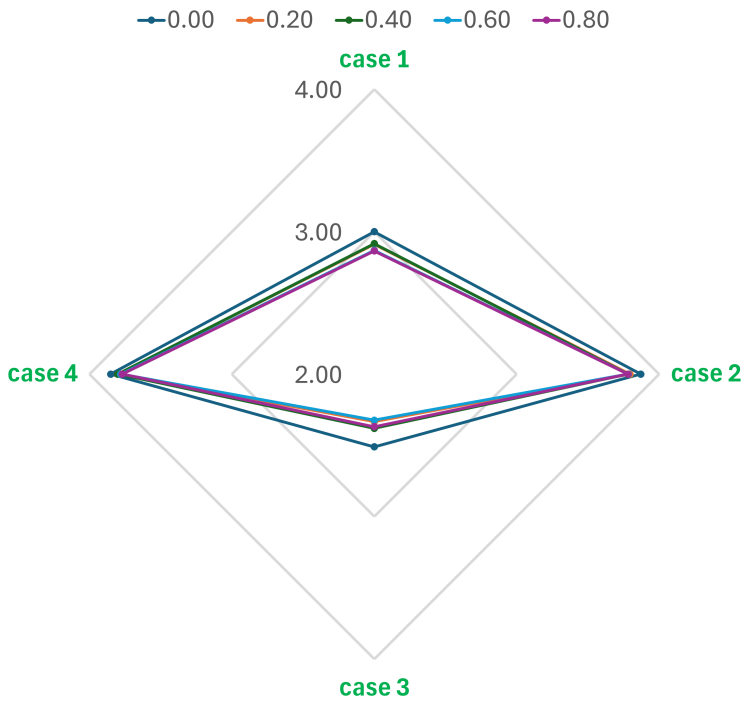


Figure 3.5. Comparison of estimation errors (MAE (bpm)) of HR estimation from MIMO FMCW mmWave radar using different weights of loss function I.

Results of HR Estimation

Table 3.5. The HR estimation errors (MAE (bpm)) using conventional methods and the proposed rough-to-fine method.

	Baseline		VMD-based		Proposed Method			
	case1	case3	case1	case3	case1	case2	case3	case4
sub1	8.84	8.96	4.38	4.96	2.89	3.34	2.59	3.62
sub2	7.31	6.87	2.67	2.31	3.73	3.43	2.95	3.81
sub3	5.18	4.08	2.50	1.04	2.54	3.66	2.41	3.79
sub4	6.31	6.64	2.48	1.15	2.82	3.43	1.96	2.99
sub5	4.08	3.39	1.67	1.98	2.63	3.08	2.18	4.11
sub6	7.91	7.30	8.57	3.73	2.54	3.56	2.15	4.02
sub7	8.09	6.80	6.82	1.57	3.04	4.23	2.24	4.18
sub8	4.27	1.61	1.49	2.50	2.20	3.30	1.04	3.37
sub9	3.38	2.62	2.83	3.22	2.54	4.03	2.76	3.56
sub10	3.60	3.68	2.89	2.21	2.65	4.05	2.40	3.74
sub11	5.94	6.96	1.25	2.71	2.13	3.31	1.63	3.66
sub12	6.09	4.87	3.02	1.91	2.77	3.83	2.16	3.60
sub13	3.79	3.53	3.29	2.00	2.30	3.48	1.59	2.88
sub14	2.28	3.38	4.84	4.39	1.73	3.21	1.06	3.13
sub15	9.20	9.11	5.27	3.10	3.53	3.82	3.47	4.17
sub16	10.05	10.13	3.84	4.61	3.72	4.73	2.49	3.96
sub17	8.54	8.48	3.39	4.81	4.24	4.57	3.21	4.43
sub18	12.27	11.94	3.10	5.17	3.65	5.01	3.52	4.99
Ave.	6.51	6.13	3.57	2.96	2.87	3.78	2.32	3.78

Table 3.5 and Figure 3.6 presents the MAE for HR estimation using the baseline method, the VMD-based method, and the proposed proposed method. In Table 3.5, “sub1” means the subset defining recording 1 as the testing set; “Ave.” means the average value of the results from 18 subsets. The proposed method outperforms both the baseline and VMD-based methods. The baseline method exhibits significant variability in MAE across the 18 subsets, with only some achieving satisfactory HR estimation performance, while others register MAEs too high for practical use. The VMD-based method shows that shortening the window duration notably impacts performance in certain subsets, such as subset 6 and

subset 7. Even with reduced window durations for HR estimation, the use of a 5 s sliding sub-window (cases 2 and 4) keeps the MAE values within an acceptable range, demonstrating the effectiveness of the proposed approach in optimizing the real-time capability for accurate HR estimation. Furthermore, to demonstrate the improvement provided by the proposed method over the approach in [61], specifically through the use of a sliding sub-window technique for context information extraction, the study analyzed the MAE outcomes in two conditions: (i) without sub-window, and (ii) with sub-window. Table 3.6 illustrates that incorporating a sub-window strategy yields a reduction in MAE by approximately 2 bpm. The settings for Table 3.6 are shown as follows: (i) Without sub-window: The time length of the window T is 5 s; (ii) With sub-window: case 2 described in Table 3.3. This demonstrates a notable improvement in the proposed proposed method relative to the technique outlined in [61].

Table 3.6. The HR estimation errors (MAE (bpm)) using the proposed rough-to-fine method with and without sub-window.

	Without sub-window	With sub-window
Average MAE (bpm)	5.63	3.78

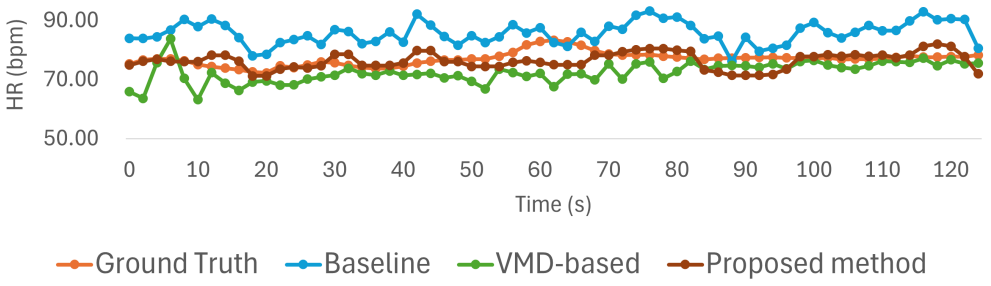


Figure 3.6. An example: the ground truth and estimation of HR of subset 1 using conventional methods and the proposed rough-to-fine method.

3.3.3 Limitations and Future Plan

The proposed research is currently in progress, and while the study have made notable strides, there remain certain limitations to the proposed findings. For instance, the study have yet to accomplish multi-person HR estimation using MIMO FMCW Radar technology. Moreover, the size of the proposed experimental dataset is insufficient to fully assess

the robustness of the proposed proposed method, and the better precision of the proposed rough HR estimation could benefit from further enhancement. The proposed future plans include successfully implementing multi-person HR estimation through MIMO FMCW Radar, amassing a larger and more comprehensive dataset to solidify the reliability of the proposed method, and improving the accuracy of the proposed rough HR estimation by other techniques.

3.4 Conclusion

The study introduce a pioneering approach to non-contact HR monitoring using MIMO FMCW Radar, which addresses and overcomes the limitations observed in previous HR estimation methodologies. The conventional methods have the high sensitivity of the algorithm parameters in different environments and the window length for HR estimation is long. The existing DL-based methods lack the consideration of contextual information for the training process. By integrating a CL method with a DL rough-to-fine model, the study have improved not only the accuracy but also the robustness of HR estimation. The proposed experimental results indicate a significant improvement over conventional methods. The proposed work lays the groundwork for future advancements in DL-based non-contact HR monitoring, promising to extend the applicability and precision.

Chapter 4

A Robust Multi-frame mmWave Radar Point Cloud-based Human Skeleton Estimation Approach with Point Cloud Reliability Assessment

4.1 Introduction

Millimeter-Wave (mmWave) radar-based skeleton estimation accurately detects and tracks human body movements [68], surpassing RGB-based [118–122], depth camera-based [123–125], and inertial sensor-based [126] approaches. It operates independently of lighting, overcomes limitations of depth camera-based methods such as reflections and occlusions, and provides precise real-time tracking, making it reliable and efficient for skeletal analysis.

In recent years, mmWave radar-based human skeleton estimation has gained significant attention, leading to notable studies [2, 65–67, 73, 127–129]. A pioneering method by A. Sengupta et al. [65] converts a 3D point cloud into two 2D images using projection and employs a forked Convolutional Neural Network (CNN) and Multi-layer Perceptron (MLP) network to estimate 25 skeletal joint coordinates. Another approach by Sengupta et al. [67] utilizes natural language processing (NLP) techniques for treating point cloud frames as paragraphs to estimate skeletal key points. To enhance performance and reduce complexity, S. An et al. [66] map 5D point cloud data to a lower dimension and utilize a CNN-based network to estimate 19 skeletal joints. S. An et al. [2] further improve this method by incorporating multi-frame fusion and meta-learning, resulting in improved efficiency and accuracy for mmWave radar-based skeleton estimation. S.-P. Lee et al. [128] collect data using both horizontal and vertical mmWave radars, where generated heatmaps are fed into a self-attention module for feature fusion and skeletal key point estimation. A point convolution-based method [73] employs point-by-point convolution to extract features from point cloud data, capturing more spatial information. Additionally, Z. Cao et al. [129] propose a two-branch learning model that independently extracts local and global features, followed by an attention-based fusion module for human skeleton estimation. Most existing point cloud-based methods primarily rely on the single-frame inputs [2, 65–67, 73, 127]. While an NLP-based method [67] incorporates multi-frame inputs, it introduces voxelization-related drawbacks, such as computational overhead and reduced spatial resolution. Notably, point cloud reliability assessment for improved robustness remains unaddressed.

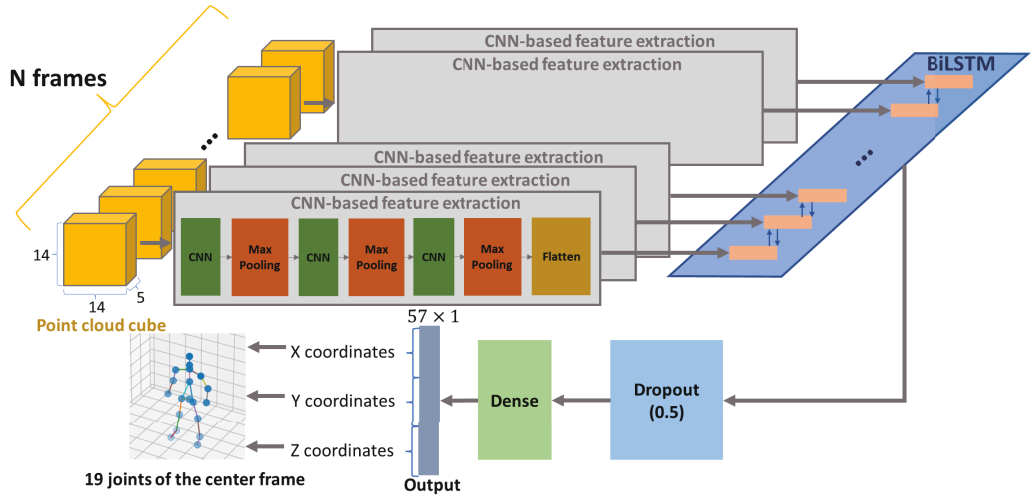


Figure 4.1. CNN and BiLSTM-based neural network for mmWave radar point cloud-based human skeleton estimation.

To address these limitations, this study presents a novel CNN and Bi-directional Long Short-Term Memory (BiLSTM)-based human skeleton estimation model that utilizes multi-frame point cloud data without the need for voxelization. Additionally, a Long Short-Term Memory (LSTM)-based neural network is introduced to assess the reliability of point cloud data segments, further enhancing the robustness of the proposed method. Through experimental evaluation, it is demonstrated that the proposed approach achieves improved accuracy and robustness in human skeleton estimation.

This chapter is structured as follows: Section 4.2 presents the proposed method, detailing its components. Section 4.3 describes the experimental setup and presents the corresponding results. Finally, Section 4.4 concludes the chapter.

4.2 Proposed Method

4.2.1 Point Cloud Processing

In the proposed research, the study use the public dataset provided by the paper [66] which generated high-quality point cloud data from only one TI IWR1443 76-81GHz radar. The maximum number of points detectable per frame is 64. To increase the point cloud density, the study apply the multi-frame fusion method used in [2] to increase the maximum number of points per frame to 192 by fusing three frames. For each point, there are 5

features including x coordinate, y coordinate, z coordinate, Doppler velocity, and signal intensity. After multi-frame fusion, the data dimension for each frame is 192×5 using zero paddings. As same as the data processing procedure proposed in [2], the point cloud data is arranged to a $14 \times 14 \times 5$, which is a 14×14 square with the depth of 5. Although multiple point cloud frames are merged during preprocessing, simple concatenation is applied rather than averaging. Our proposed approach retains the temporal variation between frames, allowing motion cues to be preserved. As a result, the effective frame rate is not compromised.

4.2.2 CNN and BiLSTM-based Skeleton Estimation

As depicted in Figure 4.1, inspired by a neural network proposed in the paper [68] for human activity recognition, a method is proposed a neural network architecture combining a time-distributed CNN and BiLSTM to estimate the locations of 19 human joints in the central frame using N frames of point cloud cubes. The proposed human skeleton estimation model employs a multi-frame-based approach. This choice of input allows us to leverage the temporal features captured by sequence models, as well as utilize the contextual information from the current frame. By incorporating both temporal information and contextual cues, the multi-frame-based model enhances the robustness of the skeleton estimation, yielding improved results. The time-distributed 2D CNN network is employed to extract a feature map for each frame. The CNN-based feature extraction structure comprises three 2D convolution layers, each connected to a max-pooling layer. The filters of three convolution layers are respectively 64, 32, and 16, and the kernel sizes are all 3×3. Each convolution layer is followed by a Rectified Linear Unit (ReLU) activation layer. These layers are followed by a flatten layer that generates 1D data as inputs for the BiLSTM neural network (units = 256). By utilizing the BiLSTM, the study can effectively capture contextual information and handle sequential tasks [130]. To mitigate overfitting, a Dropout layer with a dropout rate of 0.5 is applied. Finally, a Dense layer with an output vector of dimensions 57×1 is used to predict the locations of the 19 human joints.

4.2.3 LSTM-based Point Cloud Reliability Assessment

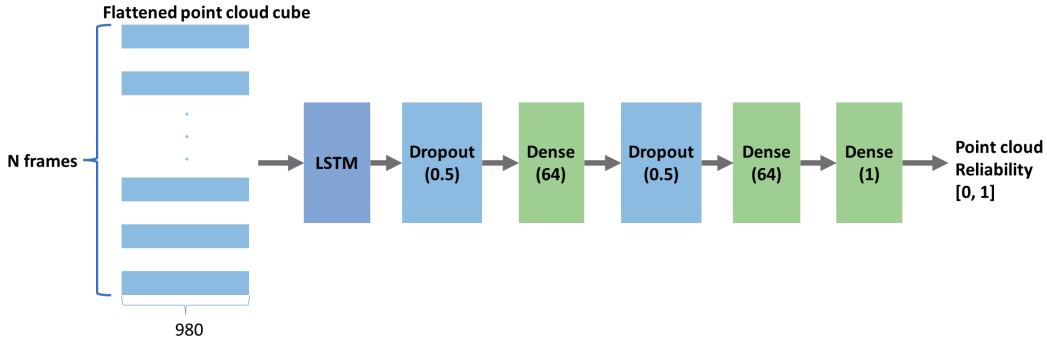


Figure 4.2. LSTM-based neural network for mmWave radar point cloud reliability assessment.

In the field of skeleton estimation, models are typically trained using relatively clean data. Consequently, when noisy data is fed into these trained models, unreliable results are often produced. As stated in Section 4.2.2, the proposed proposed multi-frame human skeleton estimation technique addresses the issue of unreliable frames that are discontinuous. However, it remains challenging to ensure robustness in scenarios where continuous unreliable frames are present. To address this issue, a method is proposed a novel approach for identifying unreliable frames of point cloud data that contain sufficient noise to impact the accuracy of human body point correspondences. By detecting and filtering out such frames, the study aim to enhance the robustness and reliability of skeleton estimation algorithms in the presence of noise. Thus, the study introduce an LSTM-based point cloud reliability assessment. Since the dataset [66] used in This study has minimal noise, the study intentionally contaminate a small portion of the dataset to evaluate the effectiveness of the proposed point cloud reliability assessment method. This assessment neural network helps identify contaminated segments including several frames and improves the robustness of skeleton estimation by removing unreliable data. As depicted in Figure 4.2, the point cloud reliability assessment neural network is constructed using LSTM. Initially, the N frames of point cloud cubes are flattened and combined to form the input vector for this LSTM-based neural network, as illustrated in Figure 4.2. The output of this neural network is a value within the range of $[0, 1]$, with values closer to 0 indicating reliability and values closer to 1 representing unreliability.

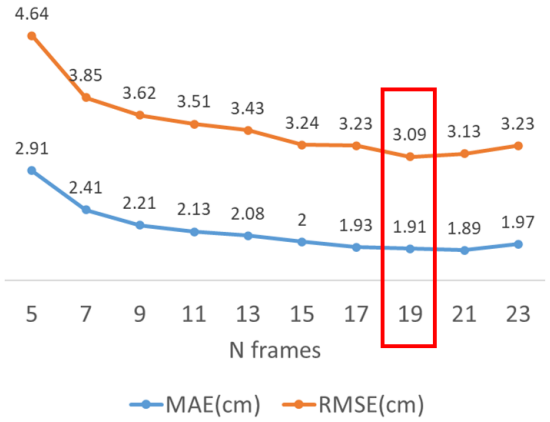


Figure 4.3. The estimation errors (MAE (cm) and RMSE (cm)) of estimated 19 joints location from mmWave radar point cloud when selecting different values of N.

Table 4.1. The estimation errors (MAE (cm) and RMSE (cm)) of estimated 19 joints location from mmWave radar point cloud using different methods

	x		y		z		Average		Number of parameters
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
mmPose [65]	6.80	10.21	4.79	6.67	6.94	9.86	6.18	8.91	2,281,739
MARS [66]	6.99	9.79	4.07	5.56	6.54	8.94	5.87	8.10	1,095,115
mRI [2]	5.08	7.52	2.71	3.73	5.00	7.10	4.26	6.12	3,247,481
Proposed method	2.22	3.62	1.63	2.42	1.87	3.22	1.91	3.09	799,465

Table 4.2. The estimation errors (MAE (cm) and RMSE (cm)) of estimated 19 joints location from mmWave radar point cloud in two scenarios: keeping or removing unreliable data detected by point cloud assessment

	Average MAE (cm)	Average RMSE (cm)
With unreliable data	3.42	5.58
Remove unreliable data	2.84	4.70

4.3 Experiment and Results

4.3.1 Experimental Setup

The study compared the performance of the proposed proposed method with conventional methods [2, 65, 66]. For the public dataset [66], the study followed the same division of

train, validate, and test sets as specified in the original paper [66]. To determine the optimal number of frames in each sample, denoted as N, the study conducted experiments with different values (As shown in Figure 4.3) and found that the best performance for the human skeleton estimation task was achieved when $N = 19$. Thus, the study subsequently set $N = 19$ for both the human skeleton detection neural network and the point cloud reliability assessment neural network. After applying the pre-processing procedure described in Section 4.2.1, the study obtained a total of 24,066 samples for training, 8,033 samples for validation, and 7,984 samples for testing. These datasets were utilized to train, validate, and test the proposed human skeleton estimation model.

As mentioned in Section 4.2.3, the study introduced random contamination into a small portion of the dataset to simulate noisy data commonly encountered in practical scenarios. Specifically, the study randomly selected several start timestamps (500 in the train set and 500 in the test set) and designated a segment beginning with each timestamp as a noisy segment, with the frames within the segment considered noisy frames. The length of each noisy segment was randomly set between 5 frames and 20 frames (equivalent to 0.5s - 2s). For each frame designated as a noisy frame, the study added points with random coordinate values (with both Doppler velocity and intensity set to 0). The number of noise points in each frame was randomly selected within the range of 0 to 100. Through these steps, the study created contaminated train and test sets. To label the segments with 19 frames in the contaminated train set as reliable or unreliable, the study calculated the mean absolute error (MAE) of skeleton estimation for each segment using the trained skeleton estimation model. Segments with an MAE exceeding 3 cm were labeled as unreliable. After training the point cloud reliability assessment model using the contaminated train set, the study employed the trained model to detect and remove unreliable segments in the contaminated test set, thereby evaluating whether the performance improved.

The evaluation of performance in this study is based on two metrics: mean square error (MAE) and root mean square error (RMSE) of the x, y, and z coordinates of 19 joints. Additionally, the average MAE and RMSE of the x, y, and z coordinates are calculated

to provide an overall performance measure. Furthermore, the number of trainable parameters used during model training is considered an evaluation indicator, which reflects the computational complexity associated with each model. This metric helps assess the efficiency and resource requirements of the models under consideration.

4.3.2 Results and Discussion

Table 4.1 displays the MAE and RMSE values for the x, y, and z coordinates of the estimated 19 joints, along with the average MAE and RMSE, and the number of trainable parameters. The results are compared between the proposed proposed method and the conventional methods [2, 65, 66]. As depicted in Table 4.1, the accuracy of estimating the human joints' locations exhibit a gradual improvement across the four methods: mmPose [65], MARS [66], mRI [2], and the proposed proposed method. In comparison to the paper that provided the public dataset the study utilized [66], the proposed proposed method achieved a significant improvement. Specifically, the proposed method reduced the MAE by approximately 3 cm, indicating a substantial enhancement in accuracy. By employing a multi-frame-based approach, the study are able to leverage temporal features and effectively incorporate contextual information for the current frame, resulting in improved robustness for the human skeleton estimation task. Therefore, in comparison to the conventional methods [2, 65, 66], the proposed proposed method excels in handling frames with noise and effectively reduces estimation errors. Furthermore, the proposed approach also demonstrated a reduction in the number of parameters by around 20%, resulting in improved computational efficiency. These advancements highlight the effectiveness and efficiency of the proposed proposed skeleton estimation method. Table 4.2 presents the average MAE and RMSE of 19 estimated joints under two scenarios: (1) including results derived from unreliable data, and (2) excluding results obtained from unreliable data detected through the point cloud assessment method. It reveals that the average MAE and RMSE are both reduced when employing the proposed proposed point cloud reliability assessment method by removing the detected unreliable segments. The proposed proposed point cloud assessment method can effectively eliminate the presence

of continuous noisy frames, thereby enhancing the robustness of human skeleton estimation.

4.4 Conclusion

In this study, the study present a novel CNN and BiLSTM-based model for human skeleton estimation using multi-frame point cloud data as inputs, eliminating the need for voxelization. To enhance the robustness of the proposed approach, the study introduce an LSTM-based neural network for assessing point cloud reliability. Experimental evaluation demonstrates that the proposed method achieves lower MAE, RMSE, and computational complexity than conventional single frame-based methods.

Chapter 5

mmPoint-Attention: A Unified Attention Framework for Human Pose and Activity Recognition from mmWave Radar Point Clouds

5.1 Introduction

Millimeter-Wave (mmWave) radar-based pose estimation and activity recognition have become increasingly acknowledged for their precision and effectiveness in detecting and tracking human movements. Compared to traditional Red-Green-Blue (RGB)-based methods [118–120], depth camera-based approaches [123], and inertial sensor systems [126, 131], mmWave radar offers significant advantages. Unlike RGB or depth cameras, mmWave radar operates independently of lighting conditions, making it suitable for use in bright and dark environments alike. Additionally, mmWave radar is resilient to environmental challenges such as reflections and occlusions that often hinder depth camera performance. This independence allows for robust, real-time pose tracking, capable of capturing fine skeletal movements and postures. These attributes make mmWave radar a reliable and efficient option for applications requiring detailed skeletal analysis and activity recognition, such as healthcare, sports training, and human-computer interaction, where continuous and accurate tracking is essential.

Recently, mmWave radar-based human pose estimation has gained considerable interest, leading to various approaches that explore innovative methods for both pose and activity recognition [2, 65–67, 73, 127–129]. While early works like that of Sengupta et al. [65] employed Convolutional Neural Network (CNN) and Multi-Layer Perceptron (MLP) networks to estimate skeletal joints from two-dimensional (2D) projections of three-dimensional (3D) point clouds, recent methods have enhanced efficiency by optimizing dimensionality (*MARS* [66]) or utilizing multi-frame fusion (*mRI* [2]). Additionally, techniques such as Point Transformers [1] and attention-based fusion [128, 129] have been introduced to further refine pose estimation accuracy.

In the realm of activity recognition, previous studies [68, 132] have mainly focused on detecting basic daily activities using point clouds from mmWave radar. For instance, *RadHAR* [68] applies CNNs and Long Short-Term Memory (LSTM) networks to voxelized point clouds, while *m-Activity* [132] uses noise filtering and a sliding window to accumulate activity-relevant data. Recently, the *mRI* dataset [2] introduced a more complex set of daily and rehabilitation activities, representing one of the first applications of

activity recognition on mmWave radar point cloud data that extends beyond basic activities. However, the two-step approach used in [2], where pose estimation feeds into an activity classification network depends heavily on pose estimation accuracy and may lose important context by focusing only on joint locations instead of the full point cloud.

To address these limitations, a method is proposed *mmPoint-Attention*, a deep learning model that unifies attention and Transformer encoder layers for both human pose estimation and activity recognition, directly utilizing multi-frame point cloud data without the need for voxelization. The proposed method consists of multiple stages: point cloud data pre-processing, feature extraction, context information processing, and dual-task predictions for both pose and activity recognition. By leveraging temporal information across frames and integrating attention mechanisms in the feature extraction and context information processing phases, the proposed model achieves better performance than conventional methods [1, 2, 74] in estimating 3D joint positions and classifying complex activities. Additionally, a post-processing step smooths activity predictions, enhancing robustness in continuous activity sequences. To address the performance degradation caused by domain shifts between training and testing data, the study evaluate lightweight model adaptation via fine-tuning with a small number of target domain samples. This unified framework not only achieves high recognition accuracy for a broad range of activities, including complex rehabilitation movements, but also maintains computational efficiency.

5.2 Related Work

5.2.1 Pose Estimation

Recently, mmWave radar-based human pose estimation has gained growing attention, spurring numerous innovative studies [2, 65–67, 73, 127–129]. A notable method by A. Sengupta et al. [65] introduced a process where a 3D point cloud is projected into two 2D images, then analyzed using a forked CNN and MLP network to estimate 25 joint coordinates of the human skeleton. Another approach by Sengupta et al. [67] leveraged Natural

Language Processing (NLP) techniques, treating frames from point clouds as sequences analogous to paragraphs to estimate skeletal key points effectively. To simplify processing, *MARS* [66] developed a CNN-based model that reduces five-dimensional (5D) point cloud data to a lower-dimensional representation, enabling the estimation of 19 skeletal joints with enhanced efficiency. Building upon this, *mRI* [2] incorporated multi-frame fusion to further improve efficiency and accuracy in mmWave radar-based pose estimation. In addition to [2], J. Yang et al. [1] utilized Point Transformer [133] to enhance the deep learning model. S.-P. Lee et al. [128] contributed by employing both horizontal and vertical mmWave radars, generating heatmaps that are processed with a self-attention module for feature fusion and accurate keypoint estimation. Another method, based on point convolution [73], performed point-by-point convolution on point cloud data, extracting fine spatial details for enhanced pose estimation. Additionally, Z. Cao et al. [129] introduced a two-branch learning model that separately captures local and global features, integrating them via an attention-based fusion module for accurate human pose estimation.

Current point cloud-based pose estimation methods predominantly rely on single-frame inputs [2, 65–67, 73, 127], which are unable to use context information and have low robustness for some occasionally noisy frames. While the NLP-based method by Sengupta et al. [67] incorporates multi-frame inputs, it faces challenges related to 3D voxelization, such as increased computational cost from processing high-resolution voxel grids, and loss of spatial fidelity due to the quantization of point cloud data into discrete 3D volumes, which can blur fine-grained geometric details and reduce overall accuracy [134].

Addressing the limitations of previous works, a method is proposed a novel attention and Transformer-based model for human skeleton estimation that directly leverages multi-frame point cloud data as input, eliminating the need for voxelization. By integrating an attention mechanism and Transformer encoder layers, the proposed model effectively captures contextual information across frames, enhancing both the accuracy and robustness of human pose estimation. This design allows the proposed approach to better handle complex poses and adapt to variations in point cloud data.

5.2.2 Activity Recognition

mmWave radar is valued for its ability to capture detailed spatial and motion information, enabling accurate human activity recognition in real-world environments. A key challenge lies in identifying features that are robust to environmental variations. Researchers have explored a variety of features, such as channel impulse response (CIR) frequency response, micro-Doppler spectrum, and point cloud data, along with domain adaptation techniques to minimize environmental influences during feature extraction and classification.

In [135], W. Jiang et al. developed a deep learning framework for device-free human activity recognition, where activity features are extracted from radar data using a CNN, with environment-independent learning enhanced by unsupervised domain adversarial training. Y. Wang et al. [136] employed Doppler-time patterns from FMCW radar and Bi-directional LSTM (BiLSTM) networks to capture continuous activity transitions through micro-Doppler features. Methods like *RadHAR* [68] and *m-Activity* [132] applied CNNs and LSTMs on voxelized radar point clouds, addressing data sparsity through techniques like sliding windows and noise filtering, to enable accurate, real-time classification in noisy settings. Unlike these single-step approaches, mRI [2] adopted a two-step pipeline by first estimating poses using a simple CNN that operates on single-frame point clouds, and then classifying activities with ActionFormer [137]. While this approach provides detailed pose information, it depends heavily on the accuracy of the per-frame pose estimation. Consequently, in the presence of sudden noise, an inaccurate pose from any single frame can propagate errors to the subsequent activity recognition step, thereby reducing overall classification accuracy. In addition, the two-stage design requires running two separate models sequentially—each incurring its own computational cost—and the CNN-based pose estimation model used in the paper has a relatively large number of parameters. As a result, the real-time performance of the system remains limited and requires further optimization.

Earlier human activity recognition studies with mmWave radar point clouds mostly targeted a small set of simple activities. However, recent datasets like *MM-Fi* and *mRI* introduced complex rehabilitation activities, marking significant progress in human activity recognition applications. Notably, *mRI* was the first to include both daily and rehabilitation activities in mmWave radar point cloud data. Nevertheless, its two-step method faces challenges, such as heavy reliance on accurate pose estimation, potentially overlooking valuable information in the full point cloud. To overcome these limitations, a method is proposed a single-step deep neural network with an attention mechanism and Transformer encoder layers for human activity recognition, directly utilizing full point cloud data from mmWave radar. This model maintains the rich contextual information of the entire point cloud, enhancing robustness and improving activity recognition accuracy for both daily and rehabilitation activities.

5.3 Preliminaries

5.3.1 mmWave Radar

As shown in Figure 5.1, the generation of a point cloud from mmWave radar involves a sequence of processing steps that transform raw radar signals into a detailed 5D representation of objects in the environment [66]. The process starts with the radar emitting electromagnetic waves, which travel through the environment, reflect off objects (such as people), and return to the radar sensor. The first stage in processing is interference mitigation. This step reduces noise, enhancing the accuracy of subsequent calculations. Next, the radar performs a 1st Dimension Fast Fourier Transform (FFT) on the processed signal to extract range information, commonly referred to as range FFT [138]. This transformation allows the radar to measure the distance to each detected object by analyzing the time it takes for the transmitted signal to return. Following range detection, the radar conducts a 2nd Dimension FFT (Doppler FFT) to calculate the Doppler velocity of the detected objects [138]. This step leverages the Doppler effect, where the frequency shift in the reflected signal indicates whether an object is moving toward or away from the radar,

as well as its velocity. The velocity information is key for tracking moving objects and differentiating between stationary and dynamic elements in the environment. Once range and velocity data are obtained, a technique named Constant False Alarm Rate (CFAR) detection [139] is used to identify and filter valid targets. CFAR detection dynamically adjusts the threshold for distinguishing real objects from background noise or clutter, ensuring that only relevant points are kept in the final data. This stage helps improve the reliability of object detection, especially in changing environments. Finally, a 3rd Dimension FFT (angle FFT) is applied to determine the angle of arrival (AoA) of the reflected signals [138]. By analyzing the phase differences of the signal received across multiple antennas, the radar calculates the angle at which each object is located relative to its own position. This angle information, combined with range and velocity, enables the radar to create a comprehensive point cloud.

The resulting point cloud consists of a collection of data points, each representing the location, Doppler velocity, and reflection intensity of an object in 3D space. This point cloud data can then be used in applications such as pose estimation and activity recognition, where the spatial and dynamic attributes of detected objects help reconstruct body postures or recognize specific actions based on movement patterns [140].

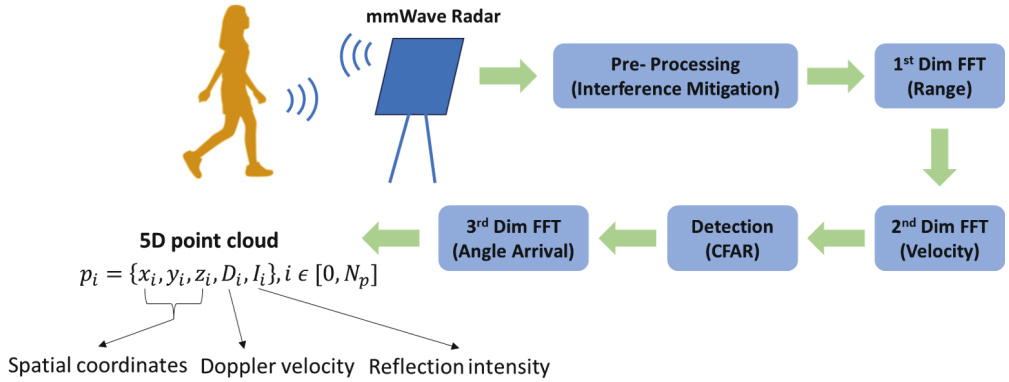


Figure 5.1. Point cloud generation from mmWave Radar.

5.3.2 Attention Mechanism

There are two kinds of attention layers employed in *mmPoint-Attention*: Efficient Channel Attention (ECA) [141] and Sequence-weighted Attention [142].

ECA [141]

ECA is a lightweight and efficient attention mechanism designed to enhance the representational power of CNNs by focusing on essential channel-wise features without significantly increasing model complexity. In the ECA module, the input feature map x with shape $[N, C, H, W]$ (N is the batch size, C is the number of channels, H and W represent the height and width of the feature map) undergoes the following steps:

First, global average pooling is applied along the spatial dimensions to produce a channel descriptor $y = [y_1, y_2, ..., y_C]$ with shape $[N, C, 1, 1]$:

$$y_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{cij}, \tag{5.1}$$

where x_{cij} represents the value at the location (c, i, j) in the feature map of a sample. Next, the squeeze and transpose operations reshape y to $[N, 1, C]$, preparing it for 1D convolution. A 1D convolution with adaptively selected kernel size k captures cross-channel dependencies. The kernel size k is calculated based on the number of channels C:

$$k = \begin{cases} \frac{|\log_2(C)+b|}{\gamma}, & \text{if } \frac{|\log_2(C)+b|}{\gamma} \text{ is odd} \\ \frac{|\log_2(C)+b|}{\gamma} + 1, & \text{otherwise} \end{cases} \tag{5.2}$$

where b and γ are two parameters to control the mapping. The 1D convolution output is passed through a sigmoid activation to obtain channel attention weights. Finally, these weights are applied to x through element-wise multiplication, resulting in the recalibrated feature map. This process allows the model to focus on informative channels, enhancing feature expressiveness.

Sequence-weighted Attention [142]

This attention mechanism enables the model to assess the significance of each frame in relation to the prediction task, assigning weights to frames when building a representation of the segment. For instance, a frame that captures crucial details may be particularly informative for understanding the overall content of the segment, so it should be weighted accordingly. Sequence-weighted attention is a simplified attention approach using a single

parameter in the attention layer:

$$v = \sum_{j=1}^{T} \frac{\exp(h_j \cdot w_a)}{\sum_{k=1}^{T} \exp(h_k \cdot w_a)} h_j. \tag{5.3}$$

Here, T represents the length of the segment, h_j denotes the hidden state or embedding of frame j, and w_a is the attention weight vector. The softmax function normalizes the scores into a probability distribution, enabling the weighted summation of frame embeddings to produce the final segment representation v, which is used by the classifier.

5.4 Proposed Method

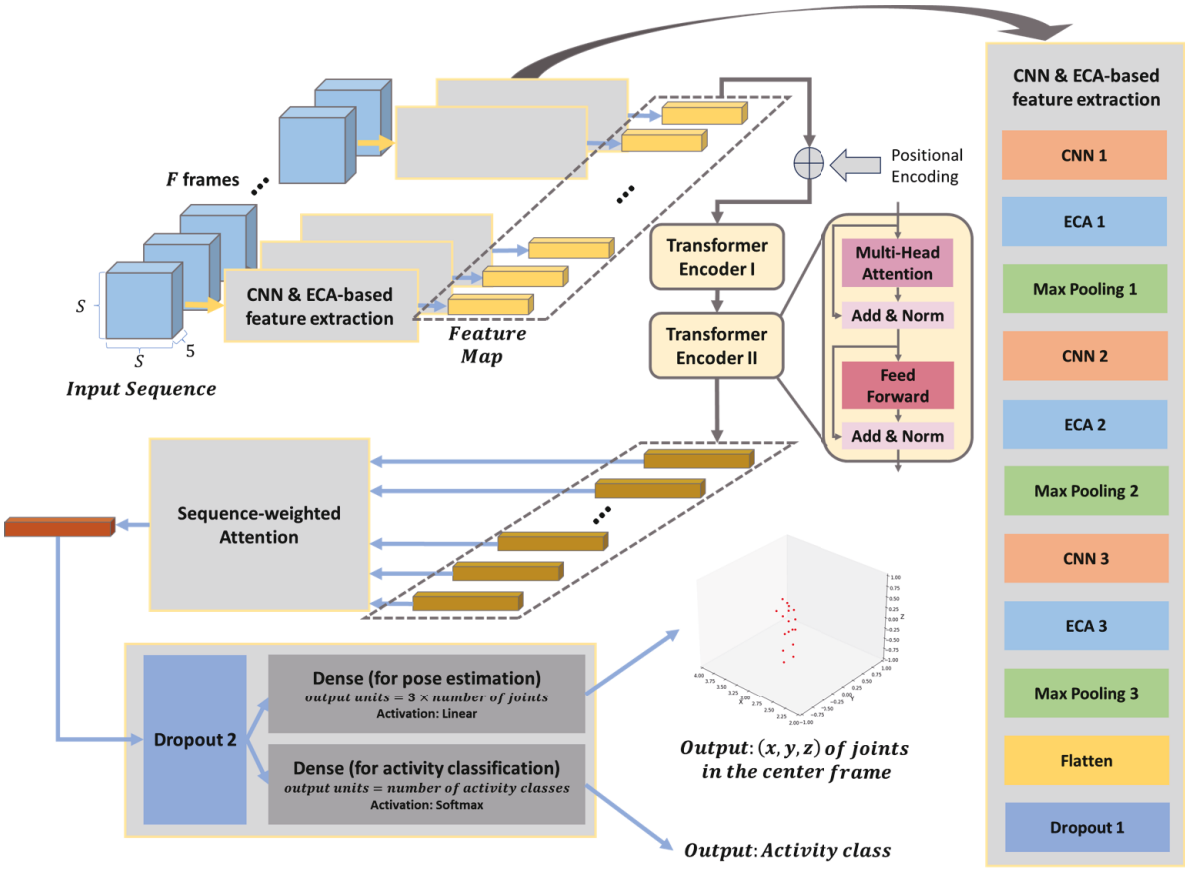


Figure 5.2. Attention and Transformer-based deep neural network for human pose estimation and activity recognition from mmWave Radar point clouds.

Five components make up the framework of *mmPoint-Attention*: (i) Point cloud data pre-processing, (ii) Human-related feature extraction, (iii) Context information extraction, (iv) Pose estimation & Activity recognition, (v) Domain-specific fine-tuning. These five

blocks are described separately as follows. As illustrated in Figure 5.2, the study present an Attention- and Transformer-based deep neural network architecture designed for human pose estimation and activity recognition. This architecture combines a human-related feature extraction block, a context information extraction block, a pose estimation block, and an activity recognition block to predict the 3D positions of 17 human joints in the central frame, as well as the activity class for the segment, using a sequence of N frames of point cloud data. This multi-frame approach allows the network to capture temporal dynamics across frames, while leveraging contextual information from the central frame, resulting in more robust and accurate skeleton estimation.

5.4.1 Point Cloud Data Pre-processing

In the proposed research, the study utilize two public datasets: *mRI* [2] and *MM-Fi* [1], which provide high-quality point cloud data collected with the Texas Instruments (TI) IWR1443 Boost (76-81 GHz) and TI IWR6843 (60-64 GHz) mmWave radars, respectively.

As shown in Figure 5.3, the study adopt a multi-frame fusion technique from [2] to increase point cloud density. Specifically, the study fuse three consecutive frames to create a denser representation for each activity, enhancing both data continuity and feature richness. Each individual point P_i within the point cloud contains five essential features: X_i, Y_i, and Z_i coordinates (representing spatial position), Doppler velocity D_i (indicating movement speed), and signal intensity I_i. After multi-frame fusion, each frame reaches a higher density, with a maximum number of points per frame set to S^2. To maintain a consistent shape across frames, if the number of points exceeds S^2, the study randomly select S^2 points. If the number of points is less than S^2, the study apply zero-padding to ensure a uniform shape of $S^2 \times 5$ for each frame.

Following the data processing approach in [2], the study reshape and structure the point cloud data into a 3D array with dimensions $S \times S \times 5$, where $S \times S$ forms a 2D spatial grid, and the depth dimension (5) holds the feature values. This transformation produces a compact, structured input format well-suited for further processing in neural

networks, capturing spatial, velocity, and intensity information across a well-defined grid. For the *mRI* dataset [2], the study set $S = 14$, and for the *MM-Fi* dataset [1], the study set $S = 11$. The selection of the value of S is based on empirical observations, ensuring that the number of points in almost all frames is less than S^2, while allowing for a very small number of exceptional frames where the point count slightly exceeds S^2.

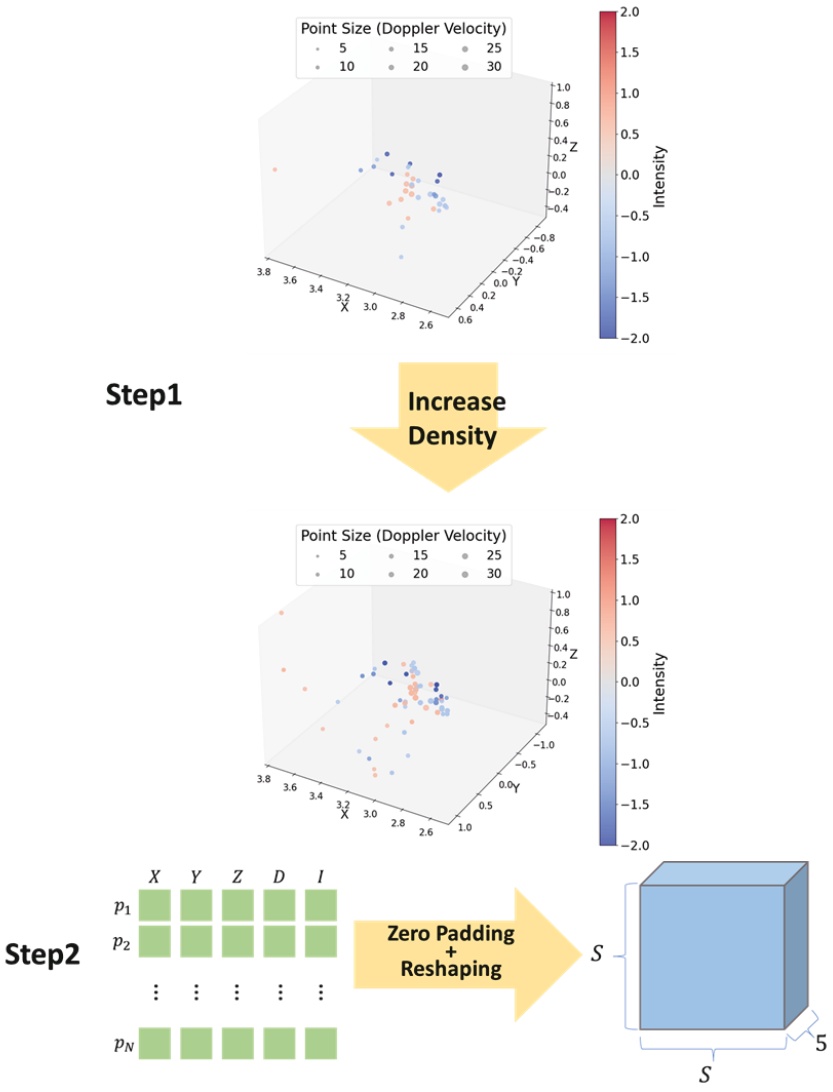


Figure 5.3. Data pre-processing method for mmWave Radar point clouds.

5.4.2 Human-related Feature Extraction

Table 5.1. The parameters of CNN and max-pooling layers in human-related feature extraction block.

	filters	**kernel size**	**strides**	**activation**
CNN 1	64			
CNN 2	32	(3,3)	(1,1)	Relu
CNN 3	16			
	strides	**pool size**		
Max pooling 1	(1,2)			
Max pooling 2	(1,2)	(2,2)		
Max pooling 3	(2,2)			

The input sequence for the deep neural network model in *mmPoint-Attention* consists of F frames of pre-processed point cloud data. Based on multiple experimental comparisons and experience, the study set $F = 13$. Each frame is processed individually through a time-distributed feature extraction block, which extracts a human-related feature map for each frame. The feature extraction block comprises three 2D CNN layers, three ECA layers [141], three max-pooling layers, a flatten layer, and a dropout layer. Each 2D CNN layer is followed by an ECA layer and a max-pooling layer to focus on informative channels, downsample, and emphasize essential features. To introduce non-linearity and enhance learning capacity, each CNN layer also includes a Rectified Linear Unit (ReLU) activation function. As shown in TABLE 5.1, different parameters are applied to the CNN and max-pooling layers. For all the ECA layers, the study set $b = 1$ and $\gamma = 2$ as specified in Eq. (5.2). After the convolutional operations, a flatten layer converts the 2D feature maps into 1D vectors, preparing them as sequential inputs for the context information extraction block. Additionally, the dropout rate for the dropout layer is set to 0.5. Through this CNN and ECA-based feature extraction block, a 1D feature map is generated for each input frame.

5.4.3 Context Information Extraction

The 1D feature maps from the N frames are first linearly projected into a 64-dimensional embedding space through a Dense layer. Then, positional encodings are added to preserve temporal order information. These enriched representations are subsequently passed through two stacked Transformer encoder layers [143]. Each Transformer encoder layer uses a multi-head self-attention mechanism with 4 attention heads, followed by a feed-forward network with a hidden dimension of 256. Dropout with a rate of 0.1 is applied after both the self-attention and feed-forward sub-layers to prevent overfitting. The embedding dimension is set to 64 throughout, and all weights are shared across frames. This design enables the model to effectively capture long-range dependencies and temporal dynamics across the entire sequence of point cloud frames. Unlike recurrent neural network models, the Transformer encoder allows each frame to directly attend to all others in the sequence, enabling efficient and flexible context modeling. This design enhances the ability of the network to capture temporal relationships, thereby improving the accuracy of joint position estimation and activity classification. Following the Transformer encoder, a sequence-weighted attention mechanism is applied to evaluate the importance of each frame for the prediction task, whether estimating the joint positions in the central frame or classifying the activity for the N-frame segment. Additionally, to mitigate overfitting, a Dropout layer with a dropout rate of 0.5 is applied after the sequence-weighted attention layer. The output of the context information extraction block is a 1D vector of length 64.

5.4.4 Pose Estimation & Activity Recognition

Finally, the deep neural network model uses two Dense layers to perform two different prediction tasks, respectively, after the context information block. For human pose estimation, a Dense layer with a linear activation outputs a vector with dimensions 51×1, representing the predicted x, y, and z coordinates for each of the 17 joints in the center frame.

For activity recognition, a Dense layer with softmax activation outputs a vector whose size is $N_{\text{classes}} \times 1$, ($N_{\text{classes}}$ is the number of activity classes) representing the prediction

probabilities for each activity class. Based on these probabilities, the study predict the activity label for each N-frame segment. Unlike *ActionFormer* [137], it is challenging to directly estimate the start and end timestamps of a continuous activity using the output of the proposed model, as it predicts the activity label for each segment independently. When consecutive segments in the ground truth correspond to the same activity, the proposed predictions may occasionally introduce an anomalous label within this sequence. For example, in the ground truth, a continuous period of activity "A01" with a start time of 5 and an end time of 50 implies that all frames in this interval should be labeled "A01". However, the proposed predictions for this time range may sporadically include other labels, such as "A02" or "A03", causing inconsistency within the continuous segment. These sporadic incorrect predictions can be considered as anomalies in an otherwise continuous period.

To address this issue, the study apply a post-processing procedure, illustrated in Figure 5.4, to eliminate anomalous predictions. A sliding window with increasing lengths, from 5 frames up to 320 frames, is used to filter out these inconsistencies. Starting with the initial predicted activity labels A, the study apply a sliding window centered around each frame, referred to as W_{center}, with a specified window length, W_{length}. Within each window, the most frequently occurring activity label, A_{freq}, is identified as the predominant activity within that window. If A_{freq} appears in more than 50% of the frames within the window, it is assigned as the activity label of the center frame $A(W_{\text{center}})$, thereby smoothing the predictions and reducing noise. This post-processing method is applied as the window slides across the entire sequence, ensuring that the final predictions contain minimal anomalous labels. It is important to note that the original frames are used for all sliding windows, and the refined values are not propagated to subsequent windows. The result is a refined and more stable set of activity labels across the sequence.

5.4.5 Domain-specific Fine-tuning

In real-world applications, it is common for the training and testing datasets to involve entirely different subjects or environments. That is, the testing set may contain previously

unseen subjects or environmental conditions that were not present during training. This domain shift typically leads to a significant drop in the performance of pose estimation and activity recognition models. To address this issue, the study investigate the effectiveness of lightweight model adaptation using a small amount of data from the target domain. Specifically, in the proposed experiments, the study construct training and testing sets that are disjoint in terms of either subjects or environments. The study then fine-tune the pre-trained model using only 1,200 samples randomly selected from the testing set (typically less than 1/15 of its total size) for 20 epochs.

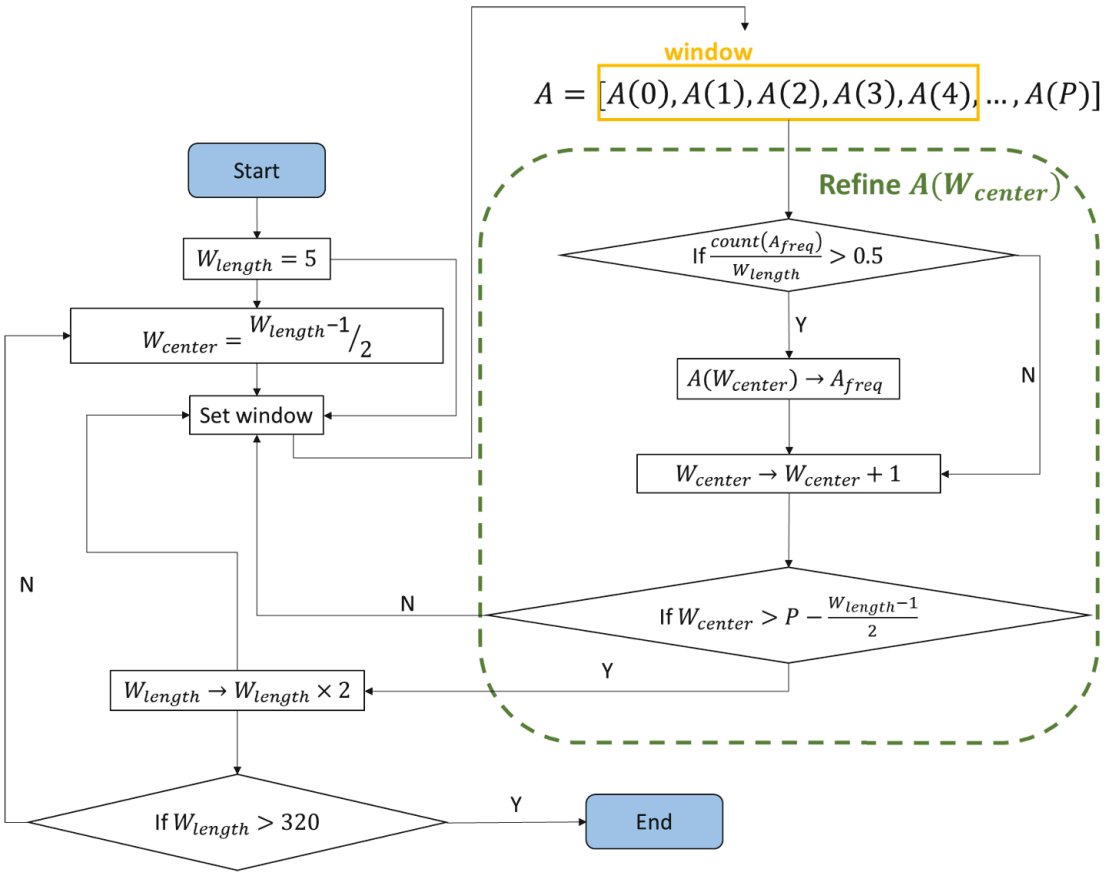


Figure 5.4. The post-processing workflow for activity labels estimated from attention and Transformer-based deep neural network.

5.5 Experiments and Discussion

5.5.1 Experimental Settings

Data Description

In the proposed research, the study utilize two publicly available mmWave radar-based human sensing datasets: *mRI* [2] and *MM-Fi* [1]. These datasets provide synchronized radar signals and 3D human pose annotations, enabling the evaluation of radar-based pose estimation and activity recognition algorithms. The *mRI* dataset was collected in a controlled indoor environment, where the radar device was placed at a fixed distance of 2.4 meters from the human subject. Ground-truth 3D poses were obtained using a dual-camera system calibrated with the radar to ensure spatial alignment. This dataset focuses on a variety of rehabilitation exercises and standard activities of daily living, making it suitable for healthcare-related applications. Similarly, the *MM-Fi* dataset was recorded with the radar sensor positioned 3 meters away from the subjects. It also employs a calibrated dual-camera setup to generate accurate 3D pose annotations. Compared to *mRI*, *MM-Fi* includes a broader set of daily activity classes and features more diverse subjects and environmental settings, such as different room layouts and clothing conditions, thereby introducing greater variability for cross-domain evaluations. In both datasets, the 3D human pose is represented by the coordinates of 17 body joints, following the standard skeleton format used in popular motion capture benchmarks. Additionally, each frame is annotated with an activity label, allowing for both pose estimation and activity classification tasks. Detailed information on activity classes is summarized in TABLE 5.2 and TABLE 5.3, respectively.

Table 5.2. The activity classes in *MM-Fi* [1] dataset

Rehabilitation Activity		**Daily Activity**	
ID	Description	ID	Description
A01	Stretching and relaxing	A02	Chest expansion (horizontal)
A06	Mark time	A03	Chest expansion (vertical)
A07	Limb extension (left)	A04	Twist (left)
A08	Limb extension (right)	A05	Twist (right)
A09	Lunge (toward left-front)	A13	Raising hand (left)
A10	Lunge (toward right-front)	A14	Raising hand (right)
A11	Limb extension (both)	A17	Waving hand (left)
A12	Squat	A18	Waving hand (right)
A15	Lunge (toward left side)	A19	Picking up things
A16	Lunge (toward right side)	A20	Throwing (toward left side)
A24	Body extension (left)	A21	Throwing (toward right side)
A25	Body extension (right)	A22	Kicking (toward left side)
A26	Jumping up	A23	Kicking (toward right side)
		A27	Bowing

Table 5.3. The activity labels in *mRI* [2] dataset

Rehabilitation Activity		**Daily Activity**	
ID	Description	ID	Description
B01	Left upper limb extension	B11	Stretching and relaxing
B02	Right upper limb extension	B12	Walking in a straight line
B03	Both upper limb extension		
B04	Left front lunge		
B05	Right front lunge		
B06	Squat		
B07	Left side lunge		
B08	Right side lunge		
B09	Left limb extension		
B10	Right limb extension		

Noise-Level-Based Data Augmentation for Robustness Evaluation

Table 5.4. Multi-level noise parameter settings for noise-level-based data augmentation.

Level	**Additive Points**		**Noisy Frames**		**Std**	**Perturb**
	Min	Max	Min	Max	σ(m)	Ratio
0	0	0	0	0	0.00	0.00
1	3	5	1	3	0.05	0.20
2	5	10	3	5	0.08	0.25
3	10	15	5	7	0.10	0.30

To validate the robustness of learning models in complex sensing environments, a method is proposed a multi-level noise augmentation framework for point clouds acquired by mmWave radar. In real-world scenarios, point cloud data is often affected by background clutter, multipath reflections, and sensor-induced measurement errors, leading to unstable perception outcomes. To simulate these real-world degradations during training and testing, the proposed method injects two types of synthetic noise into the data: additive noise, which inserts random pseudo-points in the global coordinate space to simulate false targets or clutter, and perturbative noise, which applies Gaussian perturbations to a subset of original points in position, velocity, or intensity to mimic sensor jitter or precision limitations. Noise is injected into continuous or dispersed segments of frames, with the decision of injection being probabilistically determined on a per-frame basis. To systematically control the complexity of the noise, the study define four levels (0 to 3), each specifying the number of additive points, the duration of noisy segments, and the magnitude of perturbation, as shown in Table 5.4. The meaning of the parameters in Table 5.4 are shown as following. Additive Points: number of pseudo-points per noisy frame; Noisy Frames: number of continuous frames containing noise; Std σ (m): standard deviation of Gaussian perturbation; Perturb: ratio of original points to be perturbed. This mechanism allows us to generate multiple noise levels, ranging from original point cloud sequences (Level 0) to strongly corrupted sequences (Level 3), supporting the evaluation under varying degrees of noise.

Performance Evaluation

Aligned with the research conducted in *MM-Fi* [1] and *mRI* [2], the study use different train-test split settings and different activity selection protocols for evaluation. The details of these split settings and protocols are shown in TABLE 5.5. **S1**, **S2**, and **S3** are train-test split settings. **P1**, **P2**, and **P3** are activity selection protocols. Since *mRI* dataset has no environment labels, the study only evaluate the performance using **S1** and **S2** settings.

To evaluate the performance of pose estimation, the study use two widely used metrics: Mean Per Joint Position Error (MPJPE), and Procrustes Analysis MPJPE (PA-MPJPE) [144]. MPJPE calculates the average Euclidean distance between the predicted joint positions and the ground truth joint positions. Given $\mathbf{Q}_{\text{gt}}$ as the ground truth 3D joint positions, $\mathbf{Q}_{\text{pred}}$ as the predicted 3D joint positions, and $M = 17$ as the total number of joints, MPJPE is computed as

$$\text{MPJPE} = \frac{1}{M} \sum_{i=1}^{M} \left\| \mathbf{Q}_{\text{gt}}^{(i)} - \mathbf{Q}_{\text{pred}}^{(i)} \right\|_2, \tag{5.4}$$

where $\| \cdot \|_2$ denotes the Euclidean (L2) norm. The PA-MPJPE, also known as "MPJPE after rigid alignment", applies a Procrustes alignment before computing the MPJPE. This alignment involves scaling, translation, and rotation to minimize the distance between the predicted and ground truth joint positions. The equation for PA-MPJPE is similar to that for MPJPE but includes a transformation step. The study compare the pose estimation results of the proposed proposed *mmPoint-Attention* with those of previous works [1, 2, 74]. The results for the Single frame-based method [2] and the Multi frame-based method [74] were obtained through the proposed own implementations based on the descriptions provided in the respective papers. Since [1] does not provide specific details on how to apply the Point Transformer within their model, the study rely on the reported results from that paper for evaluations on the *MM-Fi* dataset.

For activity recognition performance evaluation, the study apply different metrics for the *MM-Fi* and *mRI* datasets. In the MM-Fi dataset, each activity is recorded individually, with only one activity per recording. Both the training and testing data contain

no segments with multiple activities in a single recording. Thus, the study evaluate the performance using standard classification metrics: precision, recall, and F1 score. In contrast, the mRI dataset includes recordings with multiple activities, where the sequence, start, and end times of each activity are randomly varied. As a result, a single segment can contain multiple activities. To evaluate performance, the study use average precision (AP) with various temporal intersection over union (tIoU) thresholds, consistent with the original mRI paper [2]. tIoU is a metric commonly used in activity recognition, especially in video analysis, to measure the overlap between two temporal segments, typically between a predicted activity segment and the ground truth activity segment. Specifically, the study select tIoU = 0.5, tIoU = 0.75, and tIoU = 0.95, and also compute the mean AP over tIoU values from 0.5 to 0.95 in increments of 0.05. For activity recognition on this dataset, the study only use the **S2** train-test split setting for evaluation.

Table 5.5. Train-test split settings and activity selection protocols for performance evaluation of mmWave Radar point cloud-based pose estimation and activity recognition.

	MM-Fi [1]	*mRI* [2]
S1	Random Split (3:1)	Random Split (3:1)
S2	32 subjects for training and 8 subjects for testing	15 subjects for training and 5 subjects for testing
S3	3 environments for training and 1 environment for testing	
P1	14 daily activities	13 activities including daily & rehabilitation activities
P2	13 rehabilitation activities	10 rehabilitation activities
P3	27 activities including daily & rehabilitation activities	

5.5.2 Experimental Results

Pose Estimation

Tables 5.6 and 5.7 present the pose estimation results on the *MM-Fi* and *mRI* datasets (The units of MPJPE and PA-MPJPE are both meter), comparing the proposed proposed *mmPoint-Attention* with previous works [1, 2, 74]. As shown in Table 5.6, under **S1**, **S2** and **S3** settings, the proposed *mmPoint-Attention* achieves the best performance, yielding the lowest MPJPE and PA-MPJPE values across the protocols **P1**, **P2**, and **P3**. However, as shown in Table 5.6, before domain-specific fine-tuning, with the **S3** setting, the method proposed in [1] surpasses the proposed approach. This difference may arise because, unlike the single-frame-based methods [1, 2], multi-frame-based approaches like mmPoint-Attention use data segments for training, which often overlap and consequently contain many highly similar samples. This overlap can lead to a tendency towards slight overfitting in the proposed method compared to the single-frame approach of [1]. Expanding the diversity of data collection environments and increasing the overall data volume could likely improve the performance of the proposed approach by reducing overfitting and enhancing generalization. In the proposed method, domain-specific fine-tuning is applied only under the train-test split settings **S2** and **S3**, as setting **S1** does not involve any new subjects or environments in the test set. After domain-specific fine-tuning, both MPJPE and PA-MPJPE are significantly reduced, demonstrating that fine-tuning can adapt the pretrained model to better accommodate unseen subjects or environments. The proposed experiments also show that increasing the number of samples used for fine-tuning further improves pose estimation accuracy. However, in practical applications, collecting large amounts of ground-truth pose data (e.g., via cameras) is often impractical and time-consuming. Based on the proposed results using 1,200 samples for fine-tuning, the study find that a few minutes of data collection is sufficient for domain adaptation when deploying the system in a new environment or for a new user. It is recommended, however, that the collected data includes as diverse a set of activities as possible to maximize the capability of fine-tuning.

As shown in Table 5.7, before domain-specific fine-tuning, the proposed method consistently achieves the best performance for **S1** on the *mRI* dataset across all activity selection protocols, while exhibiting comparable performance to [74] for **S2**. These results demonstrate the robustness and accuracy of the proposed method on this dataset. After domain-specific fine-tuning, the performance under setting **S2** is further enhanced, indicating the benefit of adaptation to new subjects.

Table 5.6. The performance comparison of mmWave Radar point cloud-based pose estimation using *MM-Fi* dataset under conventional methods and the proposed mmPoint-Attention.

Single frame-based method [2]						
	P1		**P2**		**P3**	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
S1	0.1105	0.0602	0.1229	0.0624	0.1290	0.0682
S2	0.1465	0.0682	0.1732	0.0744	0.1650	0.0767
S3	0.2057	0.0826	0.2140	0.0938	0.2161	0.0917
Single frame-based method + Point Transformer [1]						
	P1		**P2**		**P3**	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
S1	0.1098	0.0556	0.1243	0.0553	0.117	0.0573
S2	0.1284	0.0587	0.1422	0.0574	0.1297	0.06
S3	0.1662	0.0739	0.168	0.073	0.1616	0.0737
Multi frame-based method [74]						
	P1		**P2**		**P3**	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
S1	0.0839	0.0490	0.0805	0.0463	0.0924	0.0536
S2	0.1322	0.0584	0.1327	0.0609	0.1313	0.0634
S3	0.1897	0.0817	0.1858	0.0838	0.1860	0.0876
mmPoint-Attention (w/o fine-tuning)						
	P1		**P2**		**P3**	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
S1	0.0735	0.0502	0.0784	0.0487	0.0873	0.0527
S2	0.1297	0.0584	0.1348	0.0618	0.1300	0.0631
S3	0.1793	0.0724	0.1985	0.0874	0.1819	0.0826
mmPoint-Attention (w/ fine-tuning)						
	P1		**P2**		**P3**	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
S1	0.0735	0.0502	0.0784	0.0487	0.0873	0.0527
S2	0.0833	0.0508	0.0952	0.0533	0.0939	0.0562
S3	0.0963	0.0555	0.1105	0.0587	0.0985	0.0571

Table 5.7. The performance comparison of mmWave Radar point cloud-based pose estimation using *mRI* dataset under conventional methods and the proposed mmPoint-Attention.

Single frame-based method [2]				
	P1		**P2**	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
S1	0.1669	0.1337	0.1396	0.0719
S2	0.2024	0.0990	0.1685	0.0737
Multi frame-based method [74]				
	P1		**P2**	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
S1	0.0944	0.0714	0.0689	0.0442
S2	0.1352	0.0845	0.1287	0.0617
mmPoint-Attention (w/o fine-tuning)				
	P1		**P2**	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
S1	0.0773	0.0680	0.0625	0.0422
S2	0.1394	0.0833	0.1299	0.0631
mmPoint-Attention (w/ fine-tuning)				
	P1		**P2**	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
S1	0.0773	0.0680	0.0625	0.0422
S2	0.0922	0.0650	0.0690	0.0460

Furthermore, as shown in Table 5.8, the proposed proposed model requires significantly fewer trainable parameters compared to the single-frame-based methods [1, 2]. While the number of trainable parameters has increased compared to the proposed previous work [74], it still remains within the range of a relatively small model, and the performance has also improved. This significant reduction in model size not only shortens training time but also moves us closer to real-time application capabilities. The smaller number of parameters reduces computational load, making the proposed model more efficient and practical for deployment in resource-constrained environments, such as mobile devices or embedded systems. This efficiency improvement without sacrificing performance makes the proposed approach highly suitable for real-time activity recognition tasks.

Table 5.8. Number of trainable parameters for mmWave Radar point cloud-based pose estimation using conventional methods and the proposed mmPoint-Attention.

Method	**Number of Trainable Parameters**
Single frame-based method [2]	2,015,603
Single frame-based method + Point Transformer [1]	≈ 500,000
Multi frame-based method [74]	58,235
mmPoint-Attention	233,086

As shown in Figure 5.5, which includes the MPJPE (m) of pose estimation using train-test split setting S1 and protocol P1, P2, and P3, the proposed evaluation of pose estimation under increasing noise levels demonstrates that the Transformer-based model exhibits superior robustness and accuracy, consistently maintaining lower MPJPE compared to LSTM-based and BiLSTM-based models across all noise conditions. This indicates that the Transformer-based model is more accurate at extracting meaningful contextual information from temporal sequences of point cloud frames. Furthermore, the results show that incorporating attention mechanisms leads to a reduction in MPJPE, highlighting the effectiveness of both the ECA and sequential attention components in enhancing pose estimation performance.

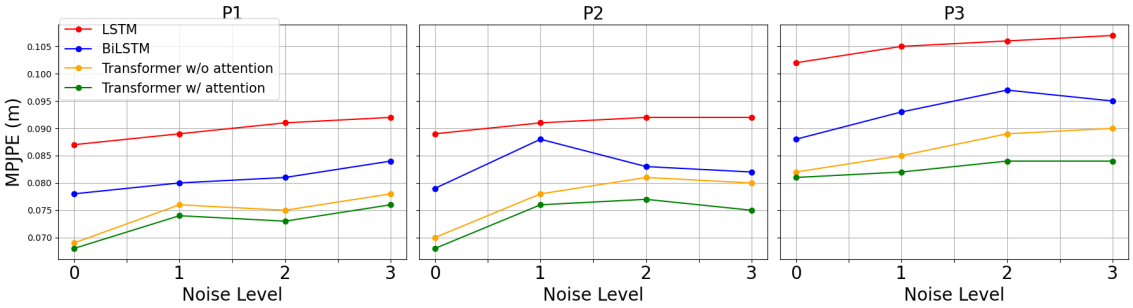
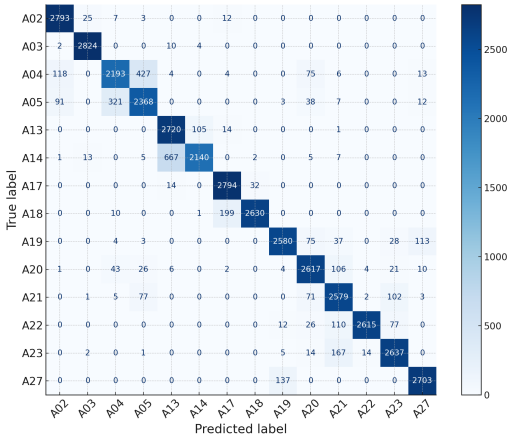
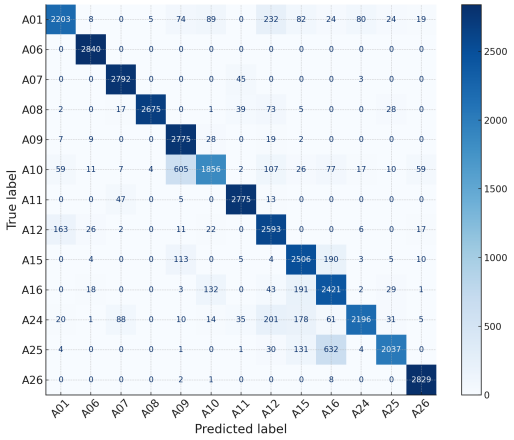


Figure 5.5. The estimation errors (MPJPE (m)) of LSTM, BiLSTM, Transformer-based (include w/o and w/ attention mode)-based model for pos estimation using mmWave Radar point cloud under different noise levels.

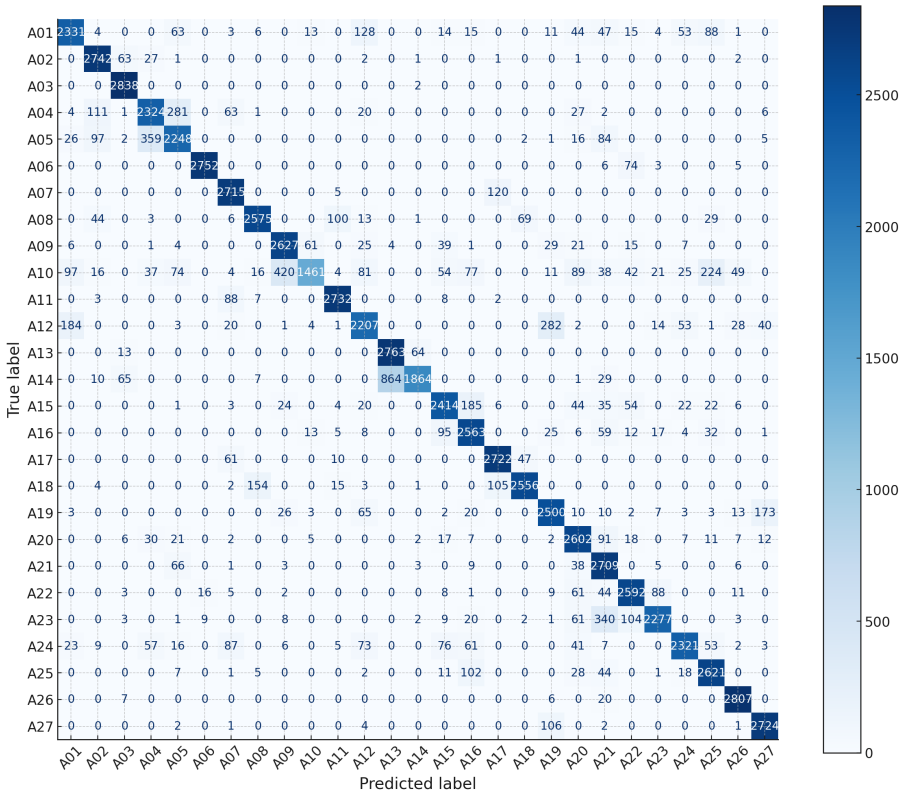
Activity Recognition



(a) Confusion matrix using protocol **P1**



(b) Confusion matrix using protocol **P2**



(c) Classification performance using protocol **P3**.

Figure 5.6. mmPoint-Attention-based activity recognition classification confusion matrix using *MM-Fi* dataset.

Figure 5.6 presents the activity recognition performance on the *MM-Fi* dataset (**S2** setting). As shown in Figure 5.6(a), activities such as A02, A03, A17, and A18 achieve higher performance than the others. This is likely because these activities, including chest

expansion and waving hands, are more standardized; there is less variation in how different subjects perform these activities, resulting in higher recognition accuracy. Similarly, as shown in Figure 5.6(b), A06, A07, A08, and A26 exhibit better performance than other rehabilitation activities for the same reason. In contrast, activities A04, A05, and A10 perform significantly worse than others across protocols **P1** and **P2**. The lower accuracy for A04 and A05, which involve twisting, can be attributed to the high dependence of these activities on subject-specific variations. Each subject performs the twist with differing ranges and angles, increasing variability and reducing recognition accuracy. For A10, the reduced accuracy is likely to be due to the presence of outliers and biases in the dataset for this activity class.

Table 5.9. Performance evaluation (AP) of mmWave Radar point cloud-based activity recognition using mRI dataset [2].

	P1				**P2**			
	AP_{50}	AP_{75}	AP_{95}	AP_{mean}	AP_{50}	AP_{75}	AP_{95}	AP_{mean}
mRI [2]	**98.22**	**97.59**	29.02	87.04	**99.00**	**97.21**	31.11	87.55
mmPoint-Attention								
(w/o attention & w/o fintune)	85.60	79.86	66.98	78.84	67.49	64.32	41.74	59.83
(w/ attention & w/o fintune)	94.70	91.13	78.61	89.65	93.87	89.75	78.11	87.94
(w/ attention & w/ fintune)	96.74	95.24	**86.08**	**94.01**	97.97	96.94	**91.46**	**96.17**

Table 5.9 compares the activity recognition results of the proposed proposed *mmPoint-Attention* with the two-step method used in [2]. In this Table, AP_{50} means the AP value while tIoU = 0.5. AP_{75} means the AP value while tIoU = 0.75. AP_{95} means the AP value while tIoU = 0.95. AP_{mean} means the average AP of the $tIoU$ = [0.5 : 0.05 : 0.95]. "w/o attention" means without attention layers; "w/ attention" means with attention layers; "w/o fine-tuning" means without domain specific fine-tuning; "w/ fine-tuning" means with domain specific fine-tuning. As shown in Table 5.9, the proposed *mmPoint-Attention* achieves the highest AP_{mean} and AP_{95}, while the two-step method introduced by [2] maintains the highest AP_{50} and AP_{75}. These results indicate that the proposed activity recognition approach generally outperforms the method in [2], achieving robust accuracy even at high tIoU thresholds (tIoU > 0.95). However, since the method in [2] is based on

ActionFormer [137], directly predicting the start and end times of activities, it is less affected by occasional outlier predictions within continuous segments, leading to slightly better performance at lower tIoU thresholds. However, overall, the proposed approach demonstrates superior performance. Moreover, through ablation experiments, the study compared the performance of the proposed method with and without the attention mechanism. The results clearly demonstrate that the inclusion of attention layers enhances the performance of the model in the activity recognition task. This confirms the effectiveness of attention mechanism in capturing crucial channel-related and contextual information essential for precise activity recognition. Moreover, domain-specific fine-tuning enhances the performance of activity recognition, demonstrating its capability in adapting to new subjects.

5.5.3 Discussion

The proposed proposed mmPoint-Attention framework introduces a multi-frame-based approach that employs attention mechanisms and Transformer encoder layers for human pose estimation and activity recognition. This unified model directly utilizes multi-frame point cloud data, eliminating the need for voxelization, which is a key advantage over previous single-frame-based methods and voxelization-based methods.

For Unseen Scenarios

As shown in Tables 5.6 and 5.7, despite the limited data, this fine-tuning process results in a model that is better adapted to the target domain, yielding improved accuracy in both pose estimation and activity recognition tasks. These 1,200 samples can be considered analogous to collecting a few minutes of motion data from a new subject or within a new environment in real-world settings. This demonstrates that even minimal additional data can significantly enhance model adaptability and performance in previously unseen scenarios.

Merits of mmPoint-Attention

One of the most notable advantages of the proposed method is its ability to leverage temporal context in multi-frame point clouds. Unlike single-frame-based approaches such as the CNN-based method [2] and the Point Transformer-based method [1], which process individual frames in isolation, mmPoint-Attention captures motion continuity. In contrast, mmPoint-Attention captures motion continuity and spatial dependencies across multiple frames using Transformer encoder layers. This significantly enhances the robustness of pose estimation by mitigating the impact of noisy frames and improving activity recognition through better sequence consistency.

Additionally, the proposed method employs two types of attention mechanisms:

- ECA, which enhances feature expressiveness while keeping computational overhead low.

- Sequence-weighted Attention, which assigns adaptive importance to different frames based on their contribution to pose estimation and activity recognition.

This combination improves feature selection and robustness, leading to superior performance compared to both single-frame-based and point-wise Transformer-based models.

Furthermore, compared to the Point Transformer [133], which applies self-attention on point features, the proposed method requires significantly fewer trainable parameters while achieving better performance in multi-frame scenarios. Point Transformer focuses on capturing long-range dependencies across point cloud data using self-attention. While this approach is effective for extracting fine-grained spatial features, it does not inherently model temporal dependencies, which are crucial for human pose estimation and activity recognition. The proposed Transformer-based approach overcomes this limitation by explicitly encoding time-series dependencies, leading to improved pose tracking and continuous activity classification. Additionally, Point Transformer-based methods tend to be computationally expensive due to the quadratic complexity of self-attention over all points. In contrast, the proposed model achieves a favorable balance between efficiency

and accuracy by utilizing lightweight attention mechanisms while maintaining state-of-the-art performance in pose estimation and activity recognition.

Limitations and Future Works

While the proposed experiments demonstrate that the proposed method enhances pose estimation and activity recognition accuracy compared to previous approaches, this study does have some limitations. First, the results indicate that when different environments are used for training and testing, the pose estimation accuracy drops considerably without fine-tuning. This highlights that, in human pose estimation using mmWave radar point clouds, the environment in which training data is collected significantly impacts the performance of the model in real-world applications. Given the vast variability of real-world environments, applying this algorithm in practical settings would require either a more robust approach to generalize across environments or a more efficient extraction of human-specific features that are independent of environmental factors. Addressing this limitation is a key focus for the proposed future work. Second, the current algorithm is limited to single-person scenarios and cannot yet be applied in multi-person contexts. In multi-person scenarios, accurately segmenting sparse point clouds and managing occlusions between individuals present substantial challenges. Developing methods to reliably separate individuals in crowded or complex environments will be essential as the study advance This study. As a potential future direction, the study envision integrating a pre-processing module for subject separation, such as clustering-based instance segmentation of the point cloud, followed by individual pose estimation pipelines per subject. For occlusion handling, incorporating temporal consistency checks and occlusion-aware attention mechanisms could help preserve identity and structural continuity across frames, even when partial information is missing. Finally, while the study evaluated the proposed *mmPoint-Attention* model on publicly available datasets to establish its feasibility, the proposed next steps will involve collecting the proposed own dataset and implementing this model in real-time applications. This progression will not only test the practicality

of the model but also allow us to optimize it for real-time use, ultimately enhancing its robustness and adaptability in diverse settings.

5.6 Conclusion

In conclusion, the proposed *mmPoint-Attention* framework makes significant contributions to mmWave radar-based pose estimation and activity recognition. By integrating attention mechanisms and Transformer encoder layers, the proposed approach directly utilizes multi-frame point cloud data, achieving superior performance on the *MM-Fi* and *mRI* datasets. Notably, the framework reduces model size compared to prior methods, enabling efficient computation and paving the way for real-time applications. Additionally, the experimental results highlight its robustness in capturing both pose and activity details, outperforming state-of-the-art methods in most evaluation protocols. This study addresses key limitations of existing approaches by eliminating the reliance on voxelization and improving temporal context modeling. Future research will focus on enhancing generalization across environments, extending to multi-person scenarios, and validating the model in real-world applications. Overall, *mmPoint-Attention* provides a unified, efficient, and accurate solution for pose estimation and activity recognition, setting a strong foundation for further advancements in the field.

Chapter 6

Conclusions and Future Work

6.1 Summary of Contributions

This thesis investigates robust, intelligent, and contactless sensing techniques for physiological signal estimation and human activity understanding. Leveraging Doppler ultrasound (DUS) and millimeter-wave (mmWave) radar technologies, the research aims to overcome challenges in real-world non-contact sensing such as signal noise, occlusion, domain variation, and single-subject limitations. The four core contributions are summarized as follows, each corresponding to a dedicated chapter in the thesis:

- **Chapter 2 – A Robust Approach Assisted by Signal Quality Assessment for Fetal Heart Rate Estimation from Doppler Ultrasound Signal:**
  This chapter presents a fetal heart rate (FHR) estimation framework using DUS signals, enhanced by an unsupervised signal quality assessment (SQA) module. We propose an integrated spectral analysis method based on short-time Fourier transform (STFT) and dual-phase autocorrelation functions (ACF) to estimate fetal RR intervals. The SQA module, implemented using variational autoencoders (VAEs) and self-organizing maps (SOMs), identifies low-quality segments and integrates with a Kalman filter to improve final estimation accuracy. The method achieves robustness across variable signal conditions without relying on labeled training data for SQA.

- **Chapter 3 – Rough-to-Fine Model-based Non-contact Heart Rate Estimation using MIMO FMCW Radar:**
  This chapter presents a non-contact heart rate (HR) estimation system using MIMO FMCW mmWave radar. The system includes a Curve-Length (CL) based person localization stage followed by a deep learning-based rough-to-fine HR estimation model. The rough stage uses peak detection. The fine stage refines the HR estimate using a dual-input CNN, which takes both filtered In-phase and Quadrature (IQ) signals and the initial rough HR estimate as inputs. The model achieves high accuracy and low latency in short time windows, demonstrating significant improvement over conventional methods.

- **Chapter 4 – A Robust Multi-frame mmWave Radar Point Cloud-based Human Skeleton Estimation Approach with Point Cloud Reliability Assessment:**
  This chapter proposes a novel skeleton estimation method from multi-frame mmWave radar point cloud data. The method uses a Convolutional Neural Network (CNN)-Bidirectional Long Short-Term Memory (BiLSTM) architecture to capture spatial and temporal dependencies. In addition, it introduces a Long Short-Term Memory (LSTM)-based point cloud reliability assessment module to suppress unreliable frames. Unlike voxelization-based methods, the approach preserves the original point cloud structure and reduces computation. The model demonstrates improved accuracy and robustness over single-frame baselines.

- **Chapter 5 – mmPoint-Attention: A Unified Attention Framework for Human Pose and Activity Recognition from mmWave Radar Point Clouds:**
  This chapter introduces mmPoint-Attention, a unified framework that jointly performs human pose estimation and activity recognition using multi-frame mmWave radar point clouds. It serves as an extended and advanced version of the method presented in Chapter 4. The method replaces BiLSTM with Transformer encoders, incorporates attention mechanisms to extract human-related features, and eliminates the need for intermediate pose-based pipelines or voxelization. Experimental results on the MM-Fi and mRI datasets show that *mmPoint-Attention* outperforms existing baselines and enables real-time application with reduced model size.

These chapters provide a unified understanding of radar- and ultrasound-based sensing. When combined with modern AI models, these technologies support accurate and robust human-centered monitoring in real-world environments.

6.2 Future Work

While the proposed methods have demonstrated strong potential and yielded encouraging results, several avenues remain open for future investigation and development. This section outlines opportunities to further extend and enhance each component of the thesis, as well as a long-term vision for their integration into a unified multimodal sensing platform.

6.2.1 FHR Estimation using Doppler Ultrasound

The unsupervised signal quality assessment (SQA) framework was trained under the assumption that low-quality segments represent statistical outliers. While effective in most cases, this assumption could be further refined—especially in scenarios where noise may dominate the dataset. As a future step, we aim to integrate lightweight statistical methods (e.g., entropy, kurtosis) to improve training data selection and enhance model robustness.

Additionally, the current experimental validation was conducted on a relatively small dataset. To improve generalizability, future work will focus on validating the framework across larger and more diverse clinical datasets, potentially in collaboration with medical institutions.

6.2.2 Heart Rate Estimation using FMCW Radar

The proposed rough-to-fine HR estimation system is currently focused on stationary, single-subject scenarios. Future development will extend the capability of the system to support multiple subjects, accommodate a wider range of body orientations, and maintain performance under occlusion and motion. To achieve this, we plan to integrate spatial filtering and adaptive beamforming techniques for multi-target separation.

Additionally, we will investigate HR estimation in dynamic conditions where subjects are moving or transitioning between postures. This will enable a more realistic validation of the system in real-world settings such as elderly care or sleep monitoring in smart homes.

6.2.3 Pose Estimation and Activity Recognition

The CNN-BiLSTM architecture (Chapter 4) and the mmPoint-Attention framework (Chapter 5) have shown promising results in single-subject settings. Building upon this foundation, future research will expand these models to support multi-person scenarios. This will involve addressing challenges such as sparse point cloud occlusion, subject interactions, and cross-identity tracking. Approaches such as instance-level segmentation, identity-preserving temporal modeling, and attention-based disentanglement will be explored.

To enhance generalization across domains, we plan to introduce domain adaptation strategies, including contrastive learning. Moreover, practical implementation and evaluation on real-time embedded systems will allow us to assess computational feasibility, latency, and robustness in dynamic environments.

In addition, incorporating shared parameters and balancing losses dynamically could be an interesting future direction, especially in scenarios where pose and activity recognition are strongly correlated or when computational resources are limited.

6.3 Long-Term Vision and Integration

As a broader goal, we envision integrating the four lines of work into a cohesive multimodal sensing platform capable of:

- Monitoring multiple users' physiological and behavioral states simultaneously;
- Adapting dynamically to environmental noise and sensor limitations;
- Supporting edge-device deployment for privacy-sensitive applications (e.g., in-home health monitoring);
- Leveraging continual learning and federated learning to ensure secure personalization without centralized data storage.

Ultimately, this work contributes to the future of ambient intelligence systems where contactless, privacy-respecting, and AI-enhanced sensing provides seamless health and activity monitoring for individuals in natural living environments.

6.4 Closing Remarks

Through the combination of signal processing, representation learning, and attention-based modeling, this thesis delivers practical solutions to longstanding challenges in radar and DUS-based physiological and behavioral sensing. Each chapter advances the field in a focused direction while contributing to a broader vision of scalable, non-contact health monitoring and smart home applications.

The foundation established by this research can support a new generation of sensing systems that are not only accurate and intelligent but also ethical, adaptable, and usable in real-world settings. Future development and collaboration with clinical and industrial partners will be essential to bring these ideas into practice.

Appendix A

List of Author's Publications and Awards

A.1 Articles on periodicals

1. **X. Shi**, N. Niida, K. Yamamoto, T. Ohtsuki, Y. Matsui, and K. Owada, "A robust approach assisted by signal quality assessment for fetal heart rate estimation from Doppler ultrasound signal," *Sensors*, vol. 23, no. 24, p. 9698, Dec. 2023.

2. **X. Shi**, M. Bouazizi, and T. Ohtsuki, "mmPoint-Attention: A Unified Attention Framework for Human Pose and Activity Recognition from mmWave Radar Point Clouds," *IEEE Sensors Journal*, vol. 25, no. 14, pp. 26794-26805, 2025, doi: 10.1109/JSEN.2025.3577772.

3. **X. Shi**, K. Yamamoto, T. Ohtsuki, Y. Matsui, and K. Owada, "Unsupervised learning-based non-invasive fetal ECG muti-level signal quality assessment," *Bioengineering*, vol. 10, no. 1, p. 66, Jan. 2023.

A.2 Articles on international conference proceedings (reviewed full-length articles)

1. **X. Shi**, K. Yamamoto, T. Ohtsuki, Y. Matsui and K. Owada, "Non-invasive Fetal ECG Signal Quality Assessment based on Unsupervised Learning Approach," *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Glasgow, Scotland, United Kingdom, 2022, pp. 1296-1299.

2. **X. Shi**, K. Yamamoto, T. Ohtsuki, Y. Matsui and K. Owada, "Unsupervised Representation Learning-based Doppler Ultrasound Signal Quality Assessment," *2022 IEEE Global Communications Conference (GLOBECOM)*, Rio de Janeiro, Brazil, 2022, pp. 2254-2259.

3. **X. Shi** and T. Ohtsuki, "A Robust Multi-Frame mmWave Radar Point Cloud-Based Human Skeleton Estimation Approach with Point Cloud Reliability Assessment," *2023 IEEE SENSORS*, Vienna, Austria, 2023, pp. 1-4.

4. **X. Shi**, M. Bouazizi, T. Ohtsuki, " Rough-to-Fine Model-based Non-contact Heart Rate Estimation using MIMO FMCW Radar," *2024 IEEE Global Communications Conference (GLOBECOM)*, Cape Town, South Africa, 2024, pp. 920-925.

A.3 Presentations at domestic conferences

1. **X. Shi**, K. Yamamoto, T. Ohtsuki, Y. Matsui, and K. Owada, "Unsupervised Learning-based 1D-Doppler Ultrasound Signal Quality Assessment," *Institute of Electronics, Information and Communication Engineers (IEICE) General Conference*, BS-3-1, 2022.

2. **X. Shi**, K. Yamamoto, T. Ohtsuki, Y. Matsui, and K. Owada, "Doppler Ultrasound Signal Quality Assessment based on Unsupervised Representation Learning for Fetal Heart Rate Estimation," *Institute of Electronics, Information and Communication Engineers (IEICE) Society Conference*, B-19-1, 2022.

3. **X. Shi**, K. Yamamoto, T. Ohtsuki and Y. Matsui, "A Robust Method for Estimating Fetal Heart Rate Using Signal Quality-Assessed Doppler Ultrasound Signals," *Institute of Electronics, Information and Communication Engineers (IEICE) Society Conference*, B-19-8, 2024.

A.4 Presentations at domestic technical meetings

1. **X. Shi**, K. Yamamoto, T. Ohtsuki, Y. Matsui, and K. Owada, "Unsupervised Learning-based Non-invasive Fetal ECG Signal Quality Assessment," *Institute of Electronics, Information and Communication Engineers (IEICE) Technical Committee on Sensor Network and Mobile Intelligence (SeMI)*, vol. 122, no. 46, pp. 15–19, 2022.

2. **X. Shi**, T. Ohtsuki, "Enhancing Human Skeleton Estimation with Multi-Frame mmWave Radar Point Cloud-based Method", Institute of Electronics, *Information and Communication Engineers (IEICE) Technical Committee on Sensor Network and Mobile Intelligence (SeMI)*, vol. 123, no.345, pp.18-21.

A.5 Others (Co-authored papers)

1. M. Mutsukawa, **X. Shi**, and T. Ohtsuki, "Chirp Correction and Phase Accumulation-Linear Interpolation-Assisted ICEEMDAN-Based Heart Rate Estimation from MIMO FMCW Radar," *2024 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*, pp. 1–6, 2024.

2. M. A. Al, **X. Shi**, B. Mondher and T. Ohtsuki, "mmGAT: Pose Estimation by Graph Attention with Mutual Features from mmWave Radar Point Cloud," *2024 IEEE International Conference on Communication (ICC)*, pp.2161-2166, 2024.

3. P. H. Rantelinggi, **X. Shi**, M. Bouazizi, and T. Ohtsuki, "A novel deep learning model for human skeleton estimation using FMCW radar," *Sensors*, vol. 25, no. 13, p. 3909, 2025, doi: 10.3390/s25133909.

References

[1] J. Yang, H. Huang, Y. Zhou, X. Chen, Y. Xu, S. Yuan, H. Zou, C. X. Lu, and L. Xie, "Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[2] S. An, Y. Li, and U. Ogras, "mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors," 2022.

[3] C. E. Valderrama, L. Stroux, N. Katebi, E. Paljug, R. Hall-Clifford, P. Rohloff, F. Marzbanrad, and G. D. Clifford, "An open source autocorrelation-based method for fetal heart rate estimation from one-dimensional doppler ultrasound," *Physiological Measurement*, vol. 40, no. 2, p. 025005, 2019.

[4] S. Ahmed, S. Irfan, N. Kiran, N. Masood, N. Anjum, and N. Ramzan, "Remote health monitoring systems for elderly people: a survey," *Sensors*, vol. 23, no. 16, p. 7095, 2023.

[5] M. Hamim, S. Paul, S. I. Hoque, M. N. Rahman, and I.-A. Baqee, "Iot based remote health monitoring system for patients and elderly people," in *2019 International conference on robotics, electrical and signal processing techniques (ICREST)*, pp. 533–538, IEEE, 2019.

[6] M. Alshamrani, "Iot and artificial intelligence implementations for remote healthcare monitoring systems: A survey," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 8, pp. 4687–4701, 2022.

[7] A. M. Ghosh, D. Halder, and S. A. Hossain, "Remote health monitoring system through iot," in *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pp. 921–926, IEEE, 2016.

[8] A. Bansal, S. Kumar, A. Bajpai, V. N. Tiwari, M. Nayak, S. Venkatesan, and R. Narayanan, "Remote health monitoring system for detecting cardiac disorders," *IET Systems Biology*, vol. 9, no. 6, pp. 309–314, 2015.

[9] L. H. Goetz and N. J. Schork, "Personalized medicine: motivation, challenges, and progress," *Fertility and sterility*, vol. 109, no. 6, pp. 952–963, 2018.

[10] A. Alyass, M. Turcotte, and D. Meyre, "From big data analysis to personalized medicine for all: challenges and opportunities," *BMC medical genomics*, vol. 8, pp. 1–12, 2015.

[11] M. Gams, I. Y.-H. Gu, A. Härmä, A. Muñoz, and V. Tam, "Artificial intelligence and ambient intelligence," *Journal of Ambient Intelligence and Smart Environments*, vol. 11, no. 1, pp. 71–86, 2019.

[12] N. Shadbolt, "Ambient intelligence," *IEEE Intell. Syst.*, vol. 18, no. 4, pp. 2–3, 2003.

[13] R. Dunne, T. Morris, and S. Harper, "A survey of ambient intelligence," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–27, 2021.

[14] S. Guarducci, S. Jayousi, S. Caputo, and L. Mucchi, "Key fundamentals and examples of sensors for human health: Wearable, non-continuous, and non-contact monitoring devices," *Sensors*, vol. 25, no. 2, p. 556, 2025.

[15] Q. Liang, L. Xu, N. Bao, L. Qi, J. Shi, Y. Yang, and Y. Yao, "Research on non-contact monitoring system for human physiological signal and body movement," *Biosensors*, vol. 9, no. 2, p. 58, 2019.

[16] 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.

[17] U. Saeed, S. Y. Shah, J. Ahmad, M. A. Imran, Q. H. Abbasi, and S. A. Shah, "Machine learning empowered covid-19 patient monitoring using non-contact sensing: An extensive review," *Journal of pharmaceutical analysis*, vol. 12, no. 2, pp. 193–204, 2022.

[18] J. M. Fernandes, J. S. Silva, A. Rodrigues, and F. Boavida, "A survey of approaches to unobtrusive sensing of humans," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–28, 2022.

[19] W. Wang and X. Wang, *Contactless vital signs monitoring*. Academic Press, 2021.

[20] N. Molinaro, E. Schena, S. Silvestri, F. Bonotti, D. Aguzzi, E. Viola, F. Buccolini, and C. Massaroni, "Contactless vital signs monitoring from videos recorded with digital cameras: An overview," *Frontiers in physiology*, vol. 13, p. 801709, 2022.

[21] H. Wang, Y. Zhou, and A. El Saddik, "Vitasi: A real-time contactless vital signs estimation system," *Computers and Electrical Engineering*, vol. 95, p. 107392, 2021.

[22] A. Li, E. Bodanese, S. Poslad, P. Chen, J. Wang, Y. Fan, and T. Hou, "A contactless health monitoring system for vital signs monitoring, human activity recognition, and tracking," *IEEE Internet of Things Journal*, vol. 11, no. 18, pp. 29275–29286, 2023.

[23] T. Negishi, S. Abe, T. Matsui, H. Liu, M. Kurosawa, T. Kirimoto, and G. Sun, "Contactless vital signs measurement system using rgb-thermal image sensors and its clinical screening test on patients with seasonal influenza," *Sensors*, vol. 20, no. 8, p. 2171, 2020.

[24] A. Vorobyov, E. Daskalaki, E. Le Roux, J. Farserotu, and P. Dallemagne, "Contactless vital signs sensing: a survey, preliminary results and challenges," in *2020*

XXXIIIrd General Assembly and Scientific Symposium of the International Union of Radio Science, pp. 1–4, IEEE, 2020.

[25] M. A. R. Ahad, U. Mahbub, and T. Rahman, *Contactless human activity analysis*, vol. 200. Springer, 2021.

[26] J. Ma, H. Wang, D. Zhang, Y. Wang, and Y. Wang, "A survey on wi-fi based contactless activity recognition," in *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, pp. 1086–1091, IEEE, 2016.

[27] U. Saeed, S. A. Shah, M. Z. Khan, A. A. Alotaibi, T. Althobaiti, N. Ramzan, M. A. Imran, and Q. H. Abbasi, "Software-defined radio-based contactless localization for diverse human activity recognition," *IEEE Sensors Journal*, vol. 23, no. 11, pp. 12041–12048, 2023.

[28] F. F. Abir, M. A. A. Faisal, O. Shahid, and M. U. Ahmed, "Contactless human activity analysis: An overview of different modalities," *Contactless Human Activity Analysis*, pp. 83–112, 2021.

[29] Z. Hao, J. Niu, X. Dang, and Z. Qiao, "Wipg: contactless action recognition using ambient wi-fi signals," *Sensors*, vol. 22, no. 1, p. 402, 2022.

[30] Z. Hussain, M. Sheng, and W. E. Zhang, "Different approaches for human activity recognition: A survey," *arXiv preprint arXiv:1906.05074*, 2019.

[31] G. Da Poian, C. J. Rozell, R. Bernardini, R. Rinaldo, and G. D. Clifford, "Matched filtering for heart rate estimation on compressive sensing ecg measurements," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 6, pp. 1349–1358, 2017.

[32] N. Dia, J. Fontecave-Jallon, M. Resendiz, M.-C. Faisant, V. Equy, D. Riethmuller, P.-Y. Gumery, and B. Rivet, "Fetal heart rate estimation by non-invasive single

abdominal electrocardiography in real clinical conditions," *Biomedical Signal Processing and Control*, vol. 71, p. 103187, 2022.

[33] A. Schäfer and J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram," *International journal of cardiology*, vol. 166, no. 1, pp. 15–29, 2013.

[34] Y. Song, J. Chen, and R. Zhang, "Heart rate estimation from incomplete electrocardiography signals," *Sensors*, vol. 23, no. 2, p. 597, 2023.

[35] P. H. Charlton, D. A. Birrenkott, T. Bonnici, M. A. Pimentel, A. E. Johnson, J. Alastruey, L. Tarassenko, P. J. Watkinson, R. Beale, and D. A. Clifton, "Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review," *IEEE reviews in biomedical engineering*, vol. 11, pp. 2–20, 2017.

[36] C.-T. Yen, S.-N. Chang, and C.-H. Liao, "Estimation of beat-by-beat blood pressure and heart rate from ecg and ppg using a fine-tuned deep cnn model," *IEEE Access*, vol. 10, pp. 85459–85469, 2022.

[37] M. Hassan, A. Malik, D. Fofi, N. Saad, and F. Meriaudeau, "Novel health monitoring method using an rgb camera," *Biomedical optics express*, vol. 8, no. 11, pp. 4838–4854, 2017.

[38] C. Romano, E. Schena, S. Silvestri, and C. Massaroni, "Non-contact respiratory monitoring using an rgb camera for real-world applications," *Sensors*, vol. 21, no. 15, p. 5126, 2021.

[39] H. Ghanadian, M. Ghodratigohar, and H. Al Osman, "A machine learning method to improve non-contact heart rate monitoring using an rgb camera," *IEEE Access*, vol. 6, pp. 57085–57094, 2018.

[40] A. Dubois and F. Charpillet, "Human activities recognition with rgb-depth camera using hmm," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4666–4669, IEEE, 2013.

[41] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2834–2841, 2013.

[42] M. H. Siddiqi, N. Almashfi, A. Ali, M. Alruwaili, Y. Alhwaiti, S. Alanazi, and M. Kamruzzaman, "A unified approach for patient activity recognition in healthcare using depth camera," *IEEE Access*, vol. 9, pp. 92300–92317, 2021.

[43] Y. Jang, I. Jeong, M. Younesi Heravi, S. Sarkar, H. Shin, and Y. Ahn, "Multi-camera-based human activity recognition for human–robot collaboration in construction," *Sensors*, vol. 23, no. 15, p. 6997, 2023.

[44] W. Sousa Lima, E. Souto, K. El-Khatib, R. Jalali, and J. Gama, "Human activity recognition using inertial sensors in a smartphone: An overview," *Sensors*, vol. 19, no. 14, p. 3213, 2019.

[45] A. Wang, G. Chen, J. Yang, S. Zhao, and C.-Y. Chang, "A comparative study on human activity recognition using inertial sensors in a smartphone," *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4566–4578, 2016.

[46] Y. Celik, M. F. Aslan, K. Sabanci, S. Stuart, W. L. Woo, and A. Godfrey, "Improving inertial sensor-based activity recognition in neurological populations," *Sensors*, vol. 22, no. 24, p. 9891, 2022.

[47] C.-T. Yen, J.-X. Liao, and Y.-K. Huang, "Human daily activity recognition performed using wearable inertial sensors combined with deep learning algorithms," *IEEE Access*, vol. 8, p. 174105–174114, 2020.

[48] P. Hamelmann, R. Vullings, A. F. Kolen, J. W. M. Bergmans, J. O. E. H. van Laar, P. Tortoli, and M. Mischi, "Doppler ultrasound technology for fetal heart rate monitoring: A review," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, p. 226–238, Feb. 2020.

[49] J. Jezewski, D. Roj, J. Wrobel, and K. Horoba, "A novel technique for fetal heart rate estimation from doppler ultrasound signal," *BioMedical Engineering OnLine*, vol. 10, Oct. 2011.

[50] T. Kupka, A. Matonia, M. Jezewski, J. Jezewski, K. Horoba, J. Wrobel, R. Czabanski, and R. Martinek, "New method for beat-to-beat fetal heart rate measurement using doppler ultrasound signal," *Sensors*, vol. 20, p. 4079, July 2020.

[51] P. Hamelmann, R. Vullings, L. Schmitt, A. F. Kolen, M. Mischi, J. O. E. H. van Laar, and J. W. M. Bergmans, "Improved ultrasound transducer positioning by fetal heart location estimation during doppler based heart rate measurements," *Physiological Measurement*, vol. 38, p. 1821–1836, Sept. 2017.

[52] A. Soumya, C. Krishna Mohan, and L. R. Cenkeramaddi, "Recent advances in mmwave-radar-based sensing, its applications, and machine learning techniques: A review," *Sensors*, vol. 23, p. 8901, Nov. 2023.

[53] F. J. Abdu, Y. Zhang, M. Fu, Y. Li, and Z. Deng, "Application of deep learning on millimeter-wave radar signals: A review," *Sensors*, vol. 21, p. 1951, Mar. 2021.

[54] A. Venon, Y. Dupuis, P. Vasseur, and P. Merriaux, "Millimeter wave fmcw radars for perception, recognition and localization in automotive applications: A survey," *IEEE Transactions on Intelligent Vehicles*, vol. 7, p. 533–555, Sept. 2022.

[55] 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, p. 463–508, 2025.

[56] H. Xue, Y. Ju, C. Miao, Y. Wang, S. Wang, A. Zhang, and L. Su, "mmmesh: Towards 3d real-time dynamic human mesh construction using millimeter-wave," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 269–282, 2021.

[57] H. Xue, Q. Cao, Y. Ju, H. Hu, H. Wang, A. Zhang, and L. Su, "M4esh: mmwave-based 3d human mesh construction for multiple subjects," in *Proceedings of the*

20th ACM Conference on Embedded Networked Sensor Systems, pp. 391–406, 2022.

[58] R. M. Grivell, Z. Alfirevic, G. M. Gyte, and D. Devane, "Antenatal cardiotocography for fetal assessment," *The Cochrane Database of Systematic Reviews*, vol. 2015, no. 9, p. CD007863, 2015.

[59] J. Jezewski, D. Roj, J. Wrobel, and K. Horoba, "A novel technique for fetal heart rate estimation from doppler ultrasound signal," *Biomedical Engineering Online*, vol. 10, pp. 1–17, 2011.

[60] F. Marzbanrad, Y. Kimura, M. Endo, M. Palaniswami, and A. H. Khandoker, "Classification of doppler ultrasound signal quality for the application of fetal valve motion identification," in *2015 Computing in Cardiology Conference (CinC)*, pp. 365–368, 2015.

[61] H.-Y. Chang, C.-H. Hsu, and W.-H. Chung, "Fast acquisition and accurate vital sign estimation with deep learning-aided weighted scheme using fmcw radar," in *2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)*, pp. 1–6, IEEE, 2022.

[62] K. J. Wu and C.-L. Yang, "Heart rate extraction with vmd algorithm in non-stationary clutter environment based on fmcw radar systems," in *2021 IEEE International Symposium on Radio-Frequency Integration Technology (RFIT)*, pp. 1–3, IEEE, 2021.

[63] L. Liu, J. Zhang, Y. Qu, S. Zhang, and W. Xiao, "mmrh: Noncontact vital sign detection with an fmcw mm-wave radar," *IEEE Sensors Journal*, vol. 23, no. 8, pp. 8856–8866, 2023.

[64] L. Qu, C. Liu, T. Yang, and Y. Sun, "Vital sign detection of fmcw radar based on improved adaptive parameter variational mode decomposition," *IEEE Sensors Journal*, vol. 23, no. 20, pp. 25048–25060, 2023.

[65] 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. 10032–10044, 2020.

[66] S. An and U. Y. Ogras, "Mars: Mmwave-based assistive rehabilitation system for smart healthcare," *ACM Transactions on Embedded Computing Systems*, vol. 20, no. 5s, 2021.

[67] A. Sengupta and S. Cao, "mmpose-nlp: A natural language processing approach to precise skeletal pose estimation using mmwave radars," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8418–8429, 2022.

[68] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar," in *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, pp. 51–56, 2019.

[69] C. H. Peters, E. D. ten Broeke, P. Andriessen, B. Vermeulen, R. C. Berendsen, P. F. Wijn, and S. G. Oei, "Beat-to-beat detection of fetal heart rate: Doppler ultrasound cardiotocography compared to direct ecg cardiotocography in time and frequency domain," *Physiological Measurement*, vol. 25, no. 2, p. 585, 2004.

[70] C. E. Valderrama, F. Marzbanrad, L. Stroux, and G. D. Clifford, "Template-based quality assessment of the doppler ultrasound signal for fetal monitoring," *Frontiers in physiology*, vol. 8, p. 511, 2017.

[71] J. C. D. Cruz, F. A. C. Javier, C. C. Puyat, and J. C. Ibera, "Non-invasive heart and respiratory rates extraction using frequency-modulated continuous-wave radar," in *2021 IEEE 4th International Conference on Computer and Communication Engineering Technology (CCET)*, pp. 382–386, IEEE, 2021.

[72] H.-Y. Chang, C.-H. Lin, Y.-C. Lin, W.-H. Chung, and T.-S. Lee, "Dl-aided nomp: a deep learning-based vital sign estimating scheme using fmcw radar," in *2020 IEEE 91st vehicular technology conference (VTC2020-Spring)*, pp. 1–7, IEEE, 2020.

[73] J. Zhong, L. Jin, and R. Wang, "Point-convolution-based human skeletal pose estimation on millimetre wave frequency modulated continuous wave multiple-input multiple-output radar," *IET Biometrics*, vol. 11, no. 4, pp. 333–342, 2022.

[74] X. Shi and T. Ohtsuki, "A robust multi-frame mmwave radar point cloud-based human skeleton estimation approach with point cloud reliability assessment," in *2023 IEEE SENSORS*, pp. 1–4, IEEE, 2023.

[75] P. Bakker, G. Colenbrander, A. Verstraeten, and H. Van Geijn, "The quality of intrapartum fetal heart rate monitoring," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 116, no. 1, pp. 22–27, 2004.

[76] R. Sameni and G. D. Clifford, "A review of fetal ecg signal processing; issues and promising directions," *The Open Pacing, Electrophysiology & Therapy Journal*, vol. 3, pp. 4–20, 2010.

[77] P. Hamelmann, R. Vullings, A. F. Kolen, J. W. Bergmans, J. O. van Laar, P. Tortoli, and M. Mischi, "Doppler ultrasound technology for fetal heart rate monitoring: a review," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 2, pp. 226–238, 2019.

[78] D. Rouvre, D. Kouamé, F. Tranquart, and L. Pourcelot, "Empirical mode decomposition (emd) for multi-gate, multi-transducer ultrasound doppler fetal heart monitoring," in *the 5th IEEE International Symposium on Signal Processing and Information Technology*, pp. 208–212, 2005.

[79] H. M. Al-Angari, Y. Kimura, L. J. Hadjileontiadis, and A. H. Khandoker, "A hybrid emd-kurtosis method for estimating fetal heart rate from continuous doppler signals," *Frontiers in Physiology*, vol. 8, p. 641, 2017.

[80] N. Seeuws, M. De Vos, and A. Bertrand, "Electrocardiogram quality assessment using unsupervised deep learning," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 2, pp. 882–893, 2021.

[81] J. Pereira and M. Silveira, "Learning representations from healthcare time series data for unsupervised anomaly detection," in *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 1–7, IEEE, 2019.

[82] T. Kohonen, I. T. Nieminen, and T. Honkela, "On the quantization error in som vs. vq: A critical and systematic study," in *Advances in Self-Organizing Maps*, pp. 133–144, 2009.

[83] Q. Li, R. G. Mark, and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter," *Physiological Measurement*, vol. 29, no. 1, p. 15, 2007.

[84] F. Andreotti, F. Graesser, H. Malberg, and S. Zaunseder, "Non-invasive fetal ecg signal quality assessment for multichannel heart rate estimation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 12, pp. 2793–2802, 2017.

[85] X. Shi, K. Yamamoto, T. Ohtsuki, Y. Matsui, and K. Owada, "Unsupervised representation learning-based doppler ultrasound signal quality assessment," in *2022 IEEE Global Communications Conference*, pp. 2254–2259, 2022.

[86] M. S. Roy, R. Gupta, and K. D. Sharma, "Photoplethysmogram signal quality evaluation by unsupervised learning approach," in *2020 IEEE Applied Signal Processing Conference (ASPCON)*, pp. 6–10, IEEE, 2020.

[87] Q. Lei, J. Yi, R. Vaculin, L. Wu, and I. S. Dhillon, "Similarity preserving representation learning for time series clustering," *arXiv preprint arXiv:1702.03584*, 2017.

[88] N. A. Heckert and J. J. Filliben, "Nist/sematech e-handbook of statistical methods; chapter 1: Exploratory data analysis," 2003. Accessed July 21, 2025.

[89] T. Rui, S. Zhang, T. Ren, J. Tang, and J. Zou, "Data reconstruction based on supervised deep auto-encoder," in *Pacific Rim Conference on Multimedia*, pp. 869–879, 2018.

[90] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

[91] P. Ghosal, D. Sarkar, S. Kundu, S. Roy, A. Sinha, and S. Ganguli, "Ecg beat quality assessment using self organizing map," in *2017 4th International Conference on Opto-electronics and Applied Optics (Optronix)*, pp. 1–5, 2017.

[92] K. Yamamoto, K. Toyoda, and T. Ohtsuki, "Spectrogram-based non-contact rri estimation by accurate peak detection algorithm," *IEEE Access*, vol. 6, pp. 60369–60379, 2018.

[93] S. Schmidt, E. Toft, C. Holst-Hansen, C. Graff, and J. Struijk, "Segmentation of heart sound recordings from an electronic stethoscope by a duration dependent hidden-markov model," in *2008 Computers in Cardiology*, pp. 345–348, 2008.

[94] H.-T. Chiang, Y.-Y. Hsieh, S.-W. Fu, K.-H. Hung, Y. Tsao, and S.-Y. Chien, "Noise reduction in ecg signals using fully convolutional denoising autoencoders," *IEEE Access*, vol. 7, pp. 60806–60813, 2019.

[95] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, vol. 82, pp. 35–45, 03 1960.

[96] S.-H. Liu, R.-X. Li, J.-J. Wang, W. Chen, and C.-H. Su, "Classification of photoplethysmographic signal quality with deep convolution neural networks for accurate measurement of cardiac stroke volume," *Applied Sciences*, vol. 10, no. 13, p. 4612, 2020.

[97] H. Gao, X. Wu, J. Geng, and Y. Lv, "Remote heart rate estimation by signal quality attention network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2122–2129, 2022.

[98] J. R. Giudicessi, M. Schram, J. M. Bos, C. D. Galloway, J. B. Shreibati, P. W. Johnson, R. E. Carter, L. W. Disrud, R. Kleiman, Z. I. Attia, *et al.*, "Artificial intelligence–enabled assessment of the heart rate corrected qt interval using a mobile electrocardiogram device," *Circulation*, vol. 143, no. 13, pp. 1274–1286, 2021.

[99] A. Temko, "Accurate heart rate monitoring during physical exercises using ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2016–2024, 2017.

[100] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep ppg: Large-scale heart rate estimation with convolutional neural networks," *Sensors*, vol. 19, p. 3079, July 2019.

[101] D. Biswas, N. Simoes-Capela, C. Van Hoof, and N. Van Helleputte, "Heart rate estimation from wrist-worn photoplethysmography: A review," *IEEE Sensors Journal*, vol. 19, p. 6560–6570, Aug. 2019.

[102] M. Panwar, A. Gautam, D. Biswas, and A. Acharyya, "Pp-net: A deep learning framework for ppg-based blood pressure and heart rate estimation," *IEEE Sensors Journal*, vol. 20, p. 10000–10011, Sept. 2020.

[103] L. Antognoli, S. Moccia, L. Migliorelli, S. Casaccia, L. Scalise, and E. Frontoni, "Heartbeat detection by laser doppler vibrometry and machine learning," *Sensors*, vol. 20, no. 18, p. 5362, 2020.

[104] G. Cosoli, L. Casacanditella, E. Tomasini, and L. Scalise, "Evaluation of heart rate variability by means of laser doppler vibrometry measurements," *Journal of Physics: Conference Series*, vol. 658, p. 012002, Nov. 2015.

[105] E. J. Sirevaag, S. Casaccia, E. A. Richter, J. A. O'Sullivan, L. Scalise, and J. W. Rohrbaugh, "Cardiorespiratory interactions: Noncontact assessment using laser doppler vibrometry," *Psychophysiology*, vol. 53, p. 847–867, Mar. 2016.

[106] T. Kitagawa, K. Yamamoto, and T. Ohtsuki, "Non-contact heartbeat detection based on beam diversity using multibeam doppler sensor," in *2021 IEEE Global Communications Conference (GLOBECOM)*, pp. 01–06, IEEE, 2021.

[107] V. L. Petrovic, M. M. Jankovic, A. V. Lupsic, V. R. Mihajlovic, and J. S. Popovic-Bozovic, "High-accuracy real-time monitoring of heart rate variability using 24 ghz continuous-wave doppler radar," *IEEE Access*, vol. 7, p. 74721–74733, 2019.

[108] S. Yuan, S. Fan, Z. Deng, and P. Pan, "Heart rate variability monitoring based on doppler radar using deep learning," *Sensors*, vol. 24, p. 2026, Mar. 2024.

[109] K. Yamamoto, K. Toyoda, and T. Ohtsuki, "Spectrogram-based non-contact rri estimation by accurate peak detection algorithm," *IEEE Access*, vol. 6, pp. 60369–60379, 2018.

[110] E. Antide, M. Zarudniev, O. Michel, and M. Pelissier, "Comparative study of radar architectures for human vital signs measurement," in *2020 IEEE Radar Conference (RadarConf20)*, pp. 1–6, IEEE, 2020.

[111] Q. Huang, D. Lu, J. Hu, H. Fan, M. Liang, and J. Zhang, "Simultaneous location and parameter estimation of human vital sign with mimo-fmcw radar," in *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, pp. 1–4, IEEE, 2019.

[112] 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, 2021.

[113] S. Rao, "Mimo radar," Tech. Rep. SWRA554A, Texas Instruments Incorporated, United States, 2018.

[114] H. Rohling, "Radar cfar thresholding in clutter and multiple target situations," *IEEE Transactions on Aerospace and Electronic Systems*, no. 4, pp. 608–621, 1983.

[115] K. Han and S. Hong, "Detection and localization of multiple humans based on curve length of i/q signal trajectory using mimo fmcw radar," *IEEE Microwave and Wireless Components Letters*, vol. 31, no. 4, pp. 413–416, 2021.

[116] K. Endo, T. Ishikawa, K. Yamamoto, and T. Ohtsuki, "Multi-person position estimation based on correlation between received signals using mimo fmcw radar," *IEEE Access*, vol. 11, pp. 2610–2620, 2023.

[117] B. Mamandipoor, D. Ramasamy, and U. Madhow, "Newtonized orthogonal matching pursuit: Frequency estimation over the continuum," *IEEE Transactions on Signal Processing*, vol. 64, no. 19, pp. 5066–5081, 2016.

[118] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[119] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *Acm Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–14, 2017.

[120] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 11656–11665, 2021.

[121] R. Yue, Z. Tian, and S. Du, "Action recognition based on rgb and skeleton data sets: A survey," *Neurocomputing*, vol. 512, pp. 287–306, 2022.

[122] C. Wang and J. Yan, "A comprehensive survey of rgb-based and skeleton-based human action recognition," *IEEE Access*, vol. 11, pp. 53880–53898, 2023.

[123] L. C. Büker, F. Zuber, A. Hein, and S. Fudickar, "Hrdepthnet: Depth image-based marker-less tracking of body joints," *Sensors*, vol. 21, no. 4, 2021.

[124] B.-S. Lin, L.-Y. Wang, Y.-T. Hwang, P.-Y. Chiang, and W.-J. Chou, "Depth-camera-based system for estimating energy expenditure of physical activities in gyms," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 1086–1095, 2019.

[125] K. Anand, K. Senthil, N. George, and V. Talasila, *Precision Testing of a Single Depth Camera System for Skeletal Joint Detection and Tracking*, pp. 385–395. Springer Nature Singapore, 2022.

[126] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll, "Sparse inertial poser: Automatic 3d human pose estimation from sparse imus," *Computer Graphics Forum*, vol. 36, no. 2, p. 349–360, 2017.

[127] S. An and U. Y. Ogras, "Fast and scalable human pose estimation using MmWave point cloud," in *Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC '22)*, (New York, NY, USA), pp. 889–894, 2022.

[128] S.-P. Lee, N. P. Kini, W.-H. Peng, C.-W. Ma, and J.-N. Hwang, "Hupr: A benchmark for human pose estimation using millimeter wave radar," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5704–5713, 2023.

[129] Z. Cao, W. Ding, R. Chen, J. Zhang, X. Guo, and G. Wang, "A joint global–local network for human pose estimation with millimeter wave radar," *IEEE Internet of Things Journal*, vol. 10, no. 1, pp. 434–446, 2023.

[130] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[131] Y. Uotani, K. Yamamoto, C. Ye, M. Bouazizi, and T. Ohtsuki, "An lstm-based approach for fall detection using accelerometer-collected data," in *Proc. IEEE Asia Pacific Conf. Commun. (APCC)*, pp. 250–255, 2023.

[132] Y. Wang, H. Liu, K. Cui, A. Zhou, W. Li, and H. Ma, "m-activity: Accurate and real-time human activity recognition via millimeter wave radar," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8298–8302, IEEE, 2021.

[133] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16259–16268, 2021.

[134] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 922–928, IEEE, 2015.

[135] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, *et al.*, "Towards environment independent device free human activity recognition," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, pp. 289–304, 2018.

[136] 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. 13607–13619, 2020.

[137] C.-L. Zhang, J. Wu, and Y. Li, "Actionformer: Localizing moments of actions with transformers," in *European Conference on Computer Vision*, pp. 492–510, Springer, 2022.

[138] S. Rao, "Introduction to mmwave sensing: Fmcw radars," *Texas Instruments (TI) mmWave Training Series*, pp. 1–11, 2017.

[139] T.-T. V. Cao, J. Palmer, and P. E. Berry, "False alarm control of cfar algorithms with experimental bistatic radar data," in *2010 IEEE Radar Conference*, pp. 156–161, IEEE, 2010.

[140] 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*, vol. 25, no. 4, pp. 2052–2087, 2023.

[141] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11534–11542, 2020.

[142] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," *arXiv preprint arXiv:1708.00524*, 2017.

[143] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[144] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.