Contactless Sensing for Future Hearing Aid Device

Aisha Fatima

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

School of Engineering
College of Science and Engineering
University of Glasgow

University of Glasgow

February 2026

Abstract

Hearing impairment affects approximately 5% of the global population, and this number is expected to rise to nearly 700 million by 2050. In the United Kingdom alone, around 11 million individuals experience hearing difficulties with age related hearing loss. Existing technologies; i.e., vision-based methods, and wearable sensors are designed to support non-verbal communication. However, they have several limitations. Vision-based methods depend heavily on lighting and raise privacy concerns, and wearable sensors, such as those embedded in hearing aids, can be inconvenient, requiring regular maintenance and frequent charging.

To overcome these challenges, this thesis explores Multimodal (MM) sensing approach for hearing impairments, focusing on sign language recognition, facial expression analysis, hand and head movement detection, phrase recognition in British Sign Language (BSL) and common signs in American Sign Language (ASL) and BSL. The research investigates the use of Radio Frequency (RF) sensing through radar and camera fusion to develop contactless, privacy-preserving system that is capable of interpreting non-verbal communication.

Across three main studies, radar and video data were collected and analyzed using advanced Deep Learning (DL) model. The first study compared radar-only and video-only systems for fifteen signs that are common in ASL and BSL, where radar achieved 96% accuracy and video achieved 82%. The second study focused on recognising micro movements, i.e., head and hand movements, and facial expressions using radar sensing, achieving 94.2% accuracy. The final study used radar data, video data, and multimodal fusion to recognise phrases in BSL. Radar-only achieved 95.47%, video-only achieved 89.72% and MM sensing achieved 93% classification accuracy. The proposed MM technique successfully demonstrate that RF sensing and multimodal fusion can accurately recognise non-verbal communication without physical contact or visual intrusion. This work establishes a proof of concept for next-generation, contactless MM hearing aid systems designed to promote independence, privacy, and accessibility for Deaf and hard-of-hearing individuals.

Contents

List of Tables

List of Figures

List of Abbreviations

ANN	Artificial Neural Network
ASL	American Sign Language
ArSL	Arabic Sign Language
BSL	British Sign Language
CSI	Channel State Information
CNN	Convolutional Neural Networks
DL	Deep Learning
FFT	Fast Fourier Transform
FPS	Frame Per Second
FMCW	Frequency-Modulated Continuous Wave
GAN	Generative Adversarial Network
GHMM	Gaussian Hidden Markov Models
HA	Hearing Aids
IC	Integrated Circuit
IOT	Internet Of Things
ISL	Indian Sign Language
KNN	k-Nearest Neighbor
ML	Machine Learning
MFCC	Mel-Frequency Cepstral Coefficients
MTI	Moving Target Indicator

MM	Multi Modal
Rx	Receiver
RTI	Range-Time-Intensity
RF	Radio Frequency
RFID	Radio Frequency Identification
RBF	Radial Basis Function
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
SE	Speech Enhancement
SLR	Sign Language Recognition
Tx	Transmitter
TPR	True Positive Rate
UWB	Ultra Wide-band
UHF	Ultra High Frequency
ViT	Vision Transformer
ViViT	Video Vision Transformer
WHO	World Health Organization

List of Publications

During my Ph.D. studies, I have been involved in several publications as the first author and co-author. While some of these publications are directly related to my thesis topic. The following is a comprehensive list of all my publications during this period.

A. *Articles*

(1) **Fatima, A.**, Hameed, H., Imran, M. A., Abbasi, Q. H., and Abbas, H. (2025) *Utilizing Contactless Sensing Technology for the Identification of Hand and Head Movements in Conjunction with Facial Expressions*. In: *IEEE Sensors Journal*, 2025, doi: 10.1109/JSEN.2025.3552489.

(2) **Reay, M.**, Hameed, H., Saleemi, B., **Fatima, A.**, Imran, M. A., and Abbasi, Q. H. (2025) *Hybrid Next-Gen Language Recognition: Multimodal Dataset for Contactless Lip Reading*. Submitted to *IEEE Sensors Journal*, 2025.

(3) **Fatima, A.**, Hameed, H., Reay, M., Imran, M. A., Abbasi, Q. H., and Abbas, H. (2025) *Performance Comparison of Video and Radar for Common Signs Recognition in BSL and ASL* Submitted to *Transactions on Human Machine Systems*.

(4) **Fatima, A.**, Hameed, H., Reay, M., Mujtaba, M., Tariq, F., Imran, M. A., Abbasi, Q. H., and Abbas, H. (2025) *Multi Modal Sensing for Sign Language Phrase Recognition* Submitted to *IEEE Sensors Journal*, 2025.

B. *Conference Proceedings*

(1) **Fatima, A.**, Hameed, H., Saleemi, B., Imran, M. A., Abbasi, Q. H., and Abbas, H. (2025) *Contactless Body Gesture Recognition for Enhancing Non-Verbal Communication: A Deep Learning Approach Using RF Sensing*. Submitted to *IEEE EMBC*, 2026.

(2) **Fatima, A.**, Hameed, H., Saleemi, B., Imran, M. A., Abbasi, Q. H., and Abbas, H. (2025) *Contactless Body Gesture Recognition for Enhancing Non-Verbal Communication: A Deep Learning Approach Using RF Sensing*. In: 2nd International Conference

on Microwave, Antennas and Circuits (ICMAC 2025), Islamabad, Pakistan, 16–17 April 2025.

(3) **Fatima, A.**, Hameed, H., Reay, M., Imran, M. A., Abbasi, Q. H., and Abbas, H. (2025) *Contactless Sensing for Sign Language Phrase Recognition*. In: 2025 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting (IEEE AP-S/URSI 2025), Ottawa, Canada, 13–18 July 2025.

(4) **Fatima, A.**, Hameed, H., Imran, M. A., Abbasi, Q. H., and Abbas, H. (2024) *Contactless Sensing for Recognizing Common Signs in ASL and BSL*. In: 2024 IEEE International Symposium on Antennas and Propagation and ITNC-USNC-URSI Radio Science Meeting, Florence, Italy, 14–19 July 2024.

(5) Farooq, M., Shawky, M. A., **Fatima, A.**, Tahir, A., Khan, M. Z., Abbas, H., Imran, M., Abbasi, Q. H., and Taha, A. (2023) *Room-Level Activity Classification from Contextual Electricity Usage Data in a Residential Home*. In: International Telecommunications Conference (ITC-Egypt 2023), Alexandria, Egypt, 18–20 July 2023.

(6) Hameed, H., Ishabakaki, P., Farooq, M., **Fatima, A.**, Arshad, K., Assaleh, K., Imran, M., and Abbasi, Q. H. (2024) *BSLR: Bridging Communication Gaps with Wi-Fi Enabled British Sign Language Recognition*. In: 2024 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting, Florence, Italy, 14–19 July 2024.

(7) Zaidi, F., Hameed, H., Farooq, M., **Fatima, A.**, Arshad, K., Assaleh, K., and Abbasi, Q. H. (2024) *Privacy-Preserving Visual Cues Communication for Hearing-Impaired People Using Deep Learning*. In: IEEE ICIP 2024, Abu Dhabi, UAE, 30 October 2024.

(8) Hameed, H., **Fatima, A.**, Lubna, L., Liaqat, S., Arshad, K., Assaleh, K., Imran, M., and Abbasi, Q. H. (2025) *AI-Driven RF Sensing for Workplace Employee Health and Fitness Monitoring*. In: 2025 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting (IEEE AP-S/URSI 2025), Ottawa, Canada, 13–18 July 2025.

(9) Hameed, H., Lubna, L., Liaqat, S., **Fatima, A.**, Assaleh, K., Arshad, K., Abbasi, Q. H., and Imran, M. A. (2025) *Revolutionizing Activity Recognition Through Walls with Deep Learning*. In: 2nd International Conference on Microwave, Antennas and Circuits (ICMAC 2025), Islamabad, Pakistan, 16–17 April 2025.

(10) Liaqat, S., Elsayed, M., Akram, Z., Hameed, H., **Fatima, A.**, and Imran, M. A. (2025) *Contactless Non-Line of Sight Respiration Rate Monitoring Using FMCW Radar*. In: 2nd International Conference on Microwave, Antennas and Circuits (ICMAC 2025), Islamabad, Pakistan, 16–17 April 2025.

Acknowledgements

Throughout the course of my doctoral research, I was privileged to receive the invaluable support and mentorship of many esteemed individuals, to whom I extend my deepest gratitude. I am very thankful to the Mary Gib Dunlop Scholarship for the fully funded award. It covered my tuition fees for 3.5 years and also provided a stipend. This support allowed me to focus on my doctoral studies without financial stress. I also sincerely thank the James Watt School of Engineering for providing supportive and motivating research environment. I am grateful for the access to essential resources and the continuous support throughout my PhD journey.

First of all, I would like to sincerely thank my supervisors, Dr. Hasan Abbas and Prof. Qammer H. Abbasi. I am deeply grateful for their continuous guidance, patience, and support throughout my PhD journey. Their expertise and valuable feedback greatly strengthened my research work. They encouraged me to think critically and approach my work with confidence. Their mentorship has played a vital role in both my academic and personal development, and I would not have been able to complete this journey without their support. Furthermore, I am grateful to Dr. Ahmad Taha for his supervision and support during the first year of my PhD.

I would also like to thank Prof. Muhammad Imran, Head of the James Watt School of Engineering, for his leadership and support during my studies. His encouragement and commitment to academic excellence created a positive and motivating research environment.

I am thankful to my colleagues and friends at the Communication, Sensing, and Imaging (CSI) Lab, who made my time at the university both productive and enjoyable. I would especially like to thank Dr. Aziz Shah and Dr. Hira Hameed for their valuable discussions, constructive feedback, and continuous support and encouragement. Working with such dedicated and supportive individuals was a truly rewarding experience.

I am also grateful to the staff members and fellow researchers at the James Watt School of Engineering. Their kindness, cooperation, and willingness to help created a welcoming and supportive research environment throughout my PhD journey.

I am deeply grateful to my family for their unconditional love and support. I especially thank my parents and siblings for their prayers, encouragement, and belief in me. Their sacrifices and support gave me the strength to overcome every challenge.

My heartfelt gratitude goes to my husband, whose love, patience, and support made this journey possible. From the beginning of my PhD in October 2022, he stood by me through

every challenge. He encouraged me during difficult times and never stopped believing in me. When our son was born in December 2022, life became more demanding. Balancing research and motherhood was challenging, but my husband's selfless support helped me continue my studies. He took on many responsibilities with care and understanding and I am forever grateful for this.

Finally, I dedicate this work to my beloved son. He came into my life at the very start of my PhD journey and became my greatest source of love, strength, and inspiration. He was only one month old when I returned to my studies, and although he was too young to understand, he quietly endured my long hours away with the innocence of a child. There were times when I had to put my work before holding him, and that was the most difficult part of this journey. However, through every smile and every tear, he gave me purpose and strength. His laughter helped me forget my tiredness, and his presence reminded me why I needed to keep going. The moments I missed with him were never forgotten, they became my motivation to finish this work. My success belongs as much to him as it does to me, and I hope one day he will know that everything I achieved was inspired by his love.

Declaration

University of Glasgow

Appendix 2.4

University of Glasgow
College Identity

Statement of Originality to Accompany Thesis Submission

Name: Aisha Fatima

Registration Number: _____

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

The copyright of this thesis rests with the author. No quotation from it is permitted without full acknowledgement.

I declare that the thesis does not include work forming part of a thesis presented successfully for another degree [unless explicitly identified and as noted below].

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

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

Signature: _____

Date: 16/04/2026

This completed statement must be bound into the submitted copies of the soft-bound thesis.

Statement of Copyright

Chapter 1

Introduction

1.1 Background and Motivation

Typical hearing is characterised by the ability to perceive sounds at or above a 20 dB hearing level. Failure to discern sounds at this threshold may indicate hearing loss. Hearing impairment can be mild or severe. It affects how people communicate and learn. Many people with hearing loss struggle to understand speech, especially in noisy places. They may ask others to repeat words, increase the TV volume, or feel tired from listening [1]. At present, about 5% of the global population, or nearly 430 million people, live with some form of hearing impairment. This number is expected to increase significantly and may reach around 700 million by 2050 [2]. In the United Kingdom (UK), approximately 11 million people are affected by hearing loss. One of the most common causes is age-related hearing loss [3].

The future of hearing aids, particularly by 2050, will require significant improvements. Current systems that rely mainly on sign language and basic non-verbal cues have limitations. Future hearing aids must better support the deaf community. They will need Multimodal (MM) processing, which uses information from different sensory modalities. These include lip reading, sign language, head movements, and facial expressions. MM processing can make communication more clearer and understandable by using information from different sensors. This increases reliability and effectiveness of hearing aids.

In recent years, privacy-preserving non-verbal communication and sign language recognition have gained research attention. This is mainly because better communication support is needed for deaf and hard-of-hearing people. In the literature, numerous approaches are used to capture non-verbal communication and sign language. Generally, they are categorized into vision-based [4, 5], and wearable sensor-based methods [6, 7].

However, existing approaches have several limitations. Camera-based systems often raise privacy concerns because they record images or videos. This makes them less suitable for real-world use. Moreover, vision-based systems also depend on lighting; poor lighting can reduce their accuracy. Wearable sensor-based systems have some benefits. However, they also have

usability problems. They often need calibration and regular maintenance. This can make them difficult for users to manage. As a result, they are less accessible for many people. Contactless sensing is becoming an important technology in healthcare. It does not require people to wear or carry devices. It can support applications such as activity monitoring and assisted living. This makes it more comfortable and convenient for users. Unlike traditional methods, contactless sensing utilises existing infrastructure like Wi-Fi routers [8], Ultrawideband (UWB) radar systems [9], Bluetooth [10], and Radio Frequency Identification (RFID) [11]. They are more convenient and cost-effective as compared to traditional methods. Moreover, contactless sensing also offers better privacy protection. It does not capture images or videos. This reduces concerns about surveillance and personal privacy. These systems do not depend on lighting conditions. They can work well in different environments. Furthermore, contactless sensing respects privacy, as sign language communication remains confidential between individuals without disclosing private data. These systems are not affected by noise or acoustic conditions. This makes them more reliable and useful in daily life.

The number of people with hearing impairments is increasing [2]. This creates a clear need for hearing aids that work well in daily life. Current hearing aids mainly use vision-based systems or wearable sensors. These approaches offer some benefits but also have clear limitations. These limitations reduce their effectiveness in everyday situations. Therefore, better and more practical solutions are needed to support users in real-world environments.

1.2 Problem Description

The number of people with hearing impairments is increasing gradually [2]. So, there is a dire need for Hearing Aids (HA) that work well in daily life. Current HA mainly use vision-based systems or wearable sensors. These approaches offer some benefits but also have some drawbacks. These drawbacks reduce their effectiveness in everyday use. Therefore, better and more practical solutions are required to support users in real-world environments.

1. **Vision-Based:** Vision-based systems use cameras to capture visual information. This includes lip movements, facial expressions, and hand gestures. These systems support lip reading and sign language recognition. They can work well in controlled environments. However, they face many problems in real-world use. Camera-based systems depend on good lighting and a clear view of the face. Poor lighting, face masks, or obstacles may reduce accuracy [12]. Continuous video recording also raises privacy concerns. This is especially important in personal and healthcare settings. Because of these issues, vision-based systems are not suitable for continuous or private communication support.

2. **Wearable Sensors:** Wearable devices are electronic gadgets that can be worn on the body, e.g., smart watches, hearing bands, or sensor-based gloves. These devices are often used

to monitor body movements or health data. These can be useful for hearing-impaired individuals because they provide additional sensory input. However, it is uncomfortable to wear sensors all the time, specially in hot weather. The devices are also inconvenient to wear during daily activities and may also cause skin irritation. Moreover, most wearable devices need regular charging or battery replacement. This can also be a problem for elderly or busy users [13]. Aforementioned factors make wearable hearing systems difficult to use in real-life situations.

All existing approaches have certain strengths but also serious drawbacks. These problems show the need for a new type of hearing aid technology that is non-invasive, contactless, and privacy-preserving. For this purpose, Radio Frequency (RF)-based sensing is proposed as a promising solution that can detect small movements such as lip, hand, and head gestures without physical contact or cameras. RF systems can work in low light, do not rely on audio signals, and protect user privacy. Therefore, this research focuses on developing an RF-based MM sensing system that can accurately recognise gestures and facial expressions to support the next-generation of hearing aids.

1.3 Aims and Objectives

The aim of this thesis is to investigate and evaluate a contactless multimodal sensing framework that combines radar and video data. The framework is used to recognise micro-motion cues such as hand gestures, head movements, facial expressions, and sign language phrases. The study examines whether this approach can improve recognition accuracy, robustness, and real-time performance, especially in dynamic or low-visibility environments. The objectives of this research are as follows:

1. To investigate and evaluate a contactless sensing framework for recognising head movements, hand gestures, facial expressions, and sign language signs, while preserving users' privacy.

2. To study the use of radar-based and vision-based sensing methods for recognising micro-motion cues and sign language signs, including signs shared between BSL and ASL.

3. To evaluate DL based classification approaches for micro-motion and sign language recognition, and to compare the performance of different model architectures.

4. To analyse the independent and combined contributions of radar and video modalities, through unimodal and multimodal evaluations, in order to assess their impact on recognition accuracy and robustness.

5. To evaluate the performance of the proposed models using metrics such as Accuracy, Precision, Recall, and F1-score, and compare the robustness of radar, video, and MM fusion approaches under diverse environmental conditions.

1.4 Proposed Solutions

Contactless sensing and imaging technologies, such as radar sensors and cameras, offer a promising approach for recognising sign language and enabling non-verbal communication while preserving users' privacy. Previous studies in this domain have made significant progress but still face limitations, particularly due to the lack of diverse datasets that capture samples from users of varying ages and genders, as well as the inability to accurately detect a broad spectrum of human movements including head motions, hand gestures, facial expressions, and body postures. This work aims to address these challenges. It proposes a unified MM system. The system can recognise different non-verbal expressions and sign language gestures. It uses micro-Doppler signatures from radar sensors. It also uses visual data captured by cameras. By combining these inputs, the system can support more accurate and reliable recognition of signs.

The first stage of this thesis focuses on sign language recognition in everyday environments. Accurate and reliable recognition is important for real world use. Many existing systems rely only on video data. These systems are affected by lighting changes, background clutter, and privacy concerns. This stage explores whether sign language can be recognised using non visual sensing. It also compares this approach with video based methods. A proof-of-concept MM framework is developed. The framework recognises fifteen common signs from American Sign Language and British Sign Language. The results show that the radar-based approach performs very well. It achieves a classification accuracy of 96%. It also remains stable under different environmental conditions. In comparison, the video-based approach achieves 81% accuracy. It is more sensitive to changes in lighting and background.

Building on the previous stage, the next part of this thesis proposes a contactless and privacy-preserving framework for recognising human emotional expressions. Traditional camera-based systems and wearable sensors have several limitations. Therefore, this work explores a radar-based expression recognition system. The system detects human movements without capturing visual images. An UWB radar is used to sense micro movements of the head, hands, and face. These movements are related to emotional expressions. The radar captures micro-Doppler signatures from these movements. Spatiotemporal features are then extracted from the captured data. Several deep learning models are used for classification. These include GoogLeNet, SqueezeNet, VGG16, and VGG19. Lightweight models are tested for real-time use. Deeper models are tested for stronger learning ability. A dataset of 1,440 samples is used. It includes sixteen different emotional expressions. The results show that radar sensing can effectively recognise emotional expressions without using visual data. The highest accuracy of 94.2% is

achieved using the VGG16 model. This performance shows the feasibility of radar-based expression recognition. It also highlights the potential of this approach for assistive applications where privacy, robustness, and reliable performance in different environments are important.

The final stage of this thesis focuses on recognising complete British Sign Language phrases in real world environments. This is a difficult task because single-sensor systems often struggle with changes in the environment, occlusions, and differences in signing style. To address these challenges, this stage evaluates radar sensing, video sensing, and their combination for phrase level recognition. The results show that the radar-only system performs well. It achieves an accuracy of 96% and remains robust under different environmental conditions. The video only system achieves 90% accuracy. However, it is more affected by changes in lighting and background. The MM system combines radar and video data. It provides the best overall performance with higher accuracy and improved robustness. Training and validation results show stable learning behaviour. Both accuracy curves converge between 95% and 97%. They closely follow each other. This indicates good generalization and minimal overfitting.

1.5 Thesis Organisation

This thesis is organised into the following chapters:

Chapter 2 presents a detailed review of existing literature and current technologies related to contactless sensing and sign language recognition. It discusses vision-based, audio-based, wearable, and RF-based systems, highlighting their strengths, limitations, and applicability in real-world scenarios. Furthermore, Machine Learning (ML) and Deep Learning (DL) approaches are discussed in this chapter.

Chapter 3 Chapter 3 describes the methodology used in the three studies presented in this thesis. Although each study investigates a different recognition problem, they all follow a common and consistent process based on contactless sensing, data preprocessing, and deep learning. The chapter first explains how radar data were collected and processed into micro-Doppler spectrograms to recognise combined head, hand, and facial expressions, and how several pre-trained CNN models were fine-tuned and evaluated for classification. It then presents the methodology for recognising fifteen common signs in ASL and BSL using radar and video data, including the experimental setup, preprocessing steps for both modalities, and the ResNet-18 model used to compare radar only and video-only performance. Finally, the chapter describes the methodology for phrase level BSL recognition using synchronized radar and video sensing, where unimodal (radar only and video only) and multimodal fusion configurations were evaluated to improve robustness and accuracy. Overall, this chapter provides a clear description of the experimental design, data collection, signal processing, model training, and evaluation metrics used through-

out the thesis.

Chapter 4 Chapter 4 presents and discusses the results obtained from the three studies in this thesis. The chapter first reports the results of the radar-based system for recognising head, hand, and facial expressions, where the performance of different deep learning models is evaluated using accuracy, precision, recall, F1 score, and confusion matrices. It then discusses the results for recognising common signs in BSL and ASL, with a clear comparison between radar-only and video-only approaches to highlight their strengths and limitations in real-world conditions. Finally, the chapter presents the results of phrase-level BSL recognition using radar-only, video-only, and multimodal sensing, showing that combining radar and video data improves recognition accuracy, robustness, and consistency across different sign classes. Overall, this chapter explains the significance of the results and shows how each study supports the development of reliable, privacy-preserving, and practical sign language recognition systems.

Chapter 5 concludes the thesis by summarising the key findings and contributions across all studies.

Chapter 2

Literature Review

This chapter provides an overview of MM sensing techniques designed to support individuals with hearing impairments, which represents a vital research area within modern healthcare systems. With the continuous advancement of healthcare technologies, a growing number of sectors are expected to adopt MM solutions to enhance accessibility and communication [14, 15, 16]. One of the key aspects discussed in this chapter is motion detection, which involves the use of sensing technologies to analyze and interpret human movement patterns [17, 18, 19]. Small scale motion detection systems are particularly valuable for individuals with hearing impairments, as they enable the recognition of subtle physical cues, such as hand movement, head movement, body gestures and facial expressions that are essential for effective non-verbal communication. According to the World Health Organization (WHO), the global population affected by hearing impairments is projected to exceed 700 million by 2050, with approximately 11 million individuals in the UK currently experiencing some degree of hearing loss. Furthermore, age related hearing impairment continues to rise, posing significant social and healthcare challenges [2, 20]. MM assistive technologies are instrumental in empowering the deaf and hard of hearing community by facilitating communication across different environments and bridging the interaction gap with the hearing population [15, 21]. In the context of SLR, these technologies combine various sensing modalities to capture, interpret, and translate both verbal and non-verbal communication cues. The primary modalities explored in this chapter include vision-based, wearable sensor-based, and RF-based systems.

This chapter is organized as follows: Section 2.1 discusses vision-based approaches, including camera and image based recognition of hand gestures and facial movements. Section 2.2 examines wearable sensor-based systems that use gloves, motion sensors, and inertial measurement units for gesture detection. Section 2.3 explores RF-based methods, such as RF sensing, which provide contactless recognition under varying environmental conditions. In addition, this chapter presents an overview of ML and DL techniques used to classify and analyze data acquired from the discussed sensing modalities, as detailed in Section 2.4. Finally, Section 2.5.3 summarizes the key findings of this chapter and outlines the identified research gaps that inform

the proposed framework of this study.

2.1 Vision-based Sign Language Recognition (SLR)

In the vision-based approach to Sign Language Recognition (SLR), extensive research has been conducted to translate sign gestures into spoken or written language. Recent advancements in DL have significantly improved the accuracy of recognising hand gestures, facial expressions, and body movements from visual data. For instance, Wadhawan and Kumar [4] utilised Convolutional Neural Networks (CNNs) for ISL recognition, achieving 99.72% and 99.90% accuracy on color and grayscale images, respectively. Similarly, Balaha et al. [22] proposed a hybrid CNN–RNN model for continuous Arabic Sign Language (ArSL) recognition, attaining 98% accuracy on a dataset of 20 Arabic words and demonstrating strong generalization on the UCF-101 benchmark. Expanding on DL architectures, Kothadiya et al. [23] introduced SignFormer, a Vision Transformer (ViT) encoder that leverages self-attention for ISL recognition, achieving 99.29% accuracy. To bridge the gap between linguistic semantics and vision models, Zuo et al. [24] developed the Natural Language-Assisted Sign Language Recognition (NLA-SLR) framework, integrating semantic gloss information for improved contextual understanding. Likewise, Liu et al. [5] employed a Feature Pyramid Network (FPN) combined with a Detection Transformer (DETR), reaching 96.45% accuracy a 1.7% improvement over prior benchmarks. Overall, these works demonstrate that CNN and transformer based architectures can effectively interpret sign gestures from RGB or depth imagery, establishing the foundation for robust visual SLR systems [25, 26, 27].

Building on this foundation, Trabelsi et al. [28] provided a comprehensive survey of classical and DL approaches to SLR, encompassing SVMs, HMMs, CNNs, LSTMs, and Transformers across isolated and continuous sign tasks using RGB, RGB-D, and skeleton data. Their taxonomy highlights rapid progress in real-time MM recognition while underscoring challenges such as signer variability, occlusion, limited dataset diversity, and high computational demands. In addressing these limitations, Brettmann et al. [29] introduced a Video Vision Transformer (ViViT) framework for word level ASL recognition. Leveraging self-attention to capture long range spatial-temporal dependencies, their model achieved a top-1 accuracy of 75.6% on WLASL100 about 10% higher than traditional CNNs demonstrating the efficacy of transformer architectures for temporal gesture modeling. However, its computational intensity limits deployment in real-time environments.

To further enhance MM integration, Alkhoraif et al. [30] proposed an ensemble Swin transformer architecture that fuses appearance (RGB) and pose information via a landmark-to-image transformation. The model achieved 93.5%, 92.6%, and 86.5% accuracies on WLASL, MS-ASL, and ASL citizen datasets, respectively; while highlighting superior robustness to background complexity and signer variation. Nonetheless, its focus on isolated word recognition and

reliance on high end GPUs restrict scalability for conversational signing. Meanwhile, Song et al. [31] developed a Hand-Aware Graph Convolutional Network (HA-GCN) emphasizing local hand topology within the human skeletal graph. Using multi-stream fusion of joint, bone, and motion features, HA-GCN achieved 96.8% and 99.6% accuracies on AUTSL and INCLUDE datasets, respectively, outperforming ST-GCN and MS-G3D baselines. The model captures fine-grained hand articulations effectively, though it remains dependent on precise pose estimation and lacks RGB-based adaptability.

Expanding into large-scale pretraining, Luqman [32] introduced SignVLM, a CLIP-based large video model that learns visual-language correspondences through contrastive pretraining. Incorporating a Transformer decoder for temporal modeling, SignVLM achieved state-of-the-art results on KArSL, WLASL, LSA64, and AUTSL datasets, demonstrating exceptional generalization to low-resource sign languages. However, it inherits semantic bias from general-vision pretraining and requires considerable computational resources. Moving toward MM learning, Ren et al. [33] designed a CNN–LSTM framework that fuses RGB and skeleton modalities through late-fusion mechanisms to improve isolated SLR. The hybrid approach enhanced accuracy across diverse environments but struggled with signer-independent variability and real-time inference limitations. Complementing this, Zhou et al. [34] integrated 3D deformable convolutions with spatio-temporal attention in a MM fusion model for sensor-network-based SLR, achieving high robustness under varied conditions though still dependent on specialized sensor data.

In pursuit of efficient real-time solutions, [35] proposed a Hybrid Transformer–CNN architecture that unites local convolutional and global attention mechanisms. Tested on the ASL Alphabet dataset, it achieved 99.97% accuracy at 110 FPS, balancing efficiency and precision. Similarly, Hassan et al. [36] introduced the Sign Nevestro DenseNet Attention (SNDA) model, combining DenseNet and attention mechanisms to reach 99.76% accuracy on ASL datasets, with enhanced robustness to illumination and camera variations. Both methods perform exceptionally for static alphabet gestures but lack temporal modeling for continuous sign sequences. Focusing on sign language translation, Gan et al. [37] developed SANet, a skeleton-aware neural network for translating sign videos into text or speech. Incorporating skeleton-guided feature fusion and graph-based temporal weighting, SANet achieved superior results on several public SLT datasets and demonstrated mobile deployment capability. However, its performance depends heavily on reliable pose estimation and high computational resources.

To optimize temporal learning, Baihan et al. [38] trained Hybrid CNN–Self-Attention–LSTM model with a Hybrid Optimizer (HO) that combines adaptive and momentum-based updates. The framework improved recognition accuracy and convergence speed while mitigating overfitting, though at the cost of increased computational complexity. Similarly, the authors in [39] conducted a comparative study of CSLR techniques, demonstrating that transformer architectures excel at contextual understanding while CNN-RNN hybrids remain effective for limited-

resource conditions, emphasizing the importance of dataset diversity for model evaluation. Advancing real-time assistive systems, Ahmad et al. [40] presented Sign Assist, an LSTM-based ISLR and translation system using MediaPipe holistic keypoints and GPT-driven language translation. The system introduced the PSL20 dataset of dynamic gestures and achieved real-time translation accuracy, although its vocabulary size and signer dependence remain constraints. Also, Baghdadi et al. [41] proposed an interpretable Vision Transformer (ViT) with LIME integration for Arabic Sign Language, achieving 99.46% and 99.88% accuracy on two ArSL datasets while offering visual interpretability. Despite its explainability, it demands extensive labeled data and is confined to static gestures.

To further address MM fusion challenges, Zhou et al. [34] extended their earlier work by combining Multi-Stream Spatio-Temporal Graph Convolutional Networks (MSGCN) with 3D deformable ResNets (D-ResNet) for adaptive fusion of RGB and skeletal cues. Their model achieved state-of-the-art results on AUTSL and WLASL datasets but required multi-sensor setups, limiting scalability. Exploring efficiency on resource-limited platforms, Carneiro et al. [42] introduced a low-cost hybrid SLR framework combining CNN–RNN DL with handcrafted descriptors such as hand-centroid trajectories. The method improved AUTSL accuracy by 7.96% with minimal added parameters, demonstrating interpretability and efficiency, though its accuracy declines under occlusions or multi-signer scenes. Similarly, Zhao et al. [43] presented CorrNet, a correlation-based continuous SLR model that captures body trajectory similarity without relying on explicit keypoints or optical flow. CorrNet achieved state-of-the-art accuracy on PHOENIX14, CSL-Daily, and CSL datasets but remains computationally demanding.

Pushing beyond single-sentence translation, [44] proposed SCOPE, a context-aware vision-language framework integrating Large Language Models (LLMs) with SLR and SLT for dialogue-based contexts. By aligning motion embeddings with LLM-derived linguistic embeddings and fine-tuning with Q-LoRA, SCOPE achieved state-of-the-art results on PHOENIX-2014T, CSL-Daily, and a newly introduced 72-hour Chinese Sign Language dataset. Although transformative, the model's dependence on LLM resources and computational cost pose practical challenges. Focusing on translation fluency, Li et al. [45] developed a Multi-View Spatio-Temporal Graph Transformer (MV-STGT) that fuses pose graphs and RGB cues via cross-attention within a Transformer encoder–decoder. Evaluated on PHOENIX14T and CSL-Daily, it surpassed prior BLEU-4 and ROUGE-L scores, offering better semantic alignment yet requiring high-quality pose data and substantial computational power. Also, Zhang et al. [46] introduced the Depth-Aware Spatio-Temporal Transformer (DASTT), which integrates RGB and depth sequences for fine-grained gesture analysis. Combining 3D convolutions with temporal attention, DASTT improved top-1 accuracy by up to 5.8% on MS-ASL and AUTSL datasets and enhanced robustness to lighting variations. Nevertheless, its reliance on depth sensors and increased latency limit deployment on lightweight devices.

2.2 Wearable Sensor-based Sign Language Recognition (SLR)

Just as in the case of vision-based, many researchers have focused on developing wearable sensor-based systems for sign language translation. These approaches aim to capture fine grained motion dynamics of the hands and fingers using embedded sensors, enabling accurate recognition even under variable lighting or complex environments. For example, the authors in [6] proposed a durable and highly sensitive glove-based sensor system that successfully translated sign language into text across 21 different sign languages. Similarly, Sign-Glove system is developed in [47], combining bend and inertial sensors to improve gesture detection accuracy, achieving 85.21% recognition performance. In another study, Qahtan et al. [7] introduced the PFDOSM-PBM framework, which provides a comparative evaluation of real-time SLRSs based on wearable sensors, emphasizing flexibility and robustness in handling uncertain motion data.

The authors in [48, 49] mentioned that wearable-based methods generally employ multiple embedded sensors; such as flex sensors, accelerometers, and gyroscopes that are mounted on gloves, wrists, or forearms to capture detailed motion patterns. These systems convert physical gestures into measurable digital data, which are then processed by ML algorithms to translate sign language gestures into textual or auditory forms. Kudrinko et al. [50] conducted a large-scale review of 72 studies published between 1991 and 2019 on wearable sensor–based SLR. Their analysis covered technologies such as flex sensors, inertial measurement units, electromyography, and magnetometers embedded in wearable platforms. The review concluded that sensor-based approaches offer greater robustness against environmental conditions and superior mobility compared to vision-based systems. However, it identified recurring challenges, including the lack of standardized datasets, inconsistent evaluation protocols, user discomfort caused by bulky hardware, and limited research on continuous sentence-level recognition.

Similarly, the authors in [51] provided an in-depth review of glove-based SLR systems developed between 2007 and 2017. Their analysis included sensory gloves employing flex, accelerometer, gyroscope, tactile, proximity, and optical sensors to record finger bending, wrist orientation, and overall hand movement. While the study confirmed the effectiveness of these systems in translating static and dynamic gestures into text or speech, it also highlighted key drawbacks; i.e., high device costs, calibration complexity, and reduced user comfort. The authors proposed that future glove designs prioritize lighter materials and improved ergonomics to enhance usability. The authors in [52] reviewed 88 studies on SLR using surface electromyography sensors. Their work explored muscle activity as a non-visual input modality for recognising hand gestures. Various classifiers were evaluated with LSTM emerging as the most effective for modeling sequential EMG data. The review also revealed that fusing EMG with inertial data improved recognition accuracy up to 99.6%. Nonetheless, EMG-based systems face challenges from signal instability, electrode placement variation, and limited participant diversity. The authors emphasized the need for standardized datasets and advanced preprocessing techniques to improve reliability and generalization.

Taneja et al. [53] focused on accelerometer, gyroscope, and flex-sensor-based wearable systems such as gloves and wristbands. They compared glove-based designs that are capable of high precision finger tracking with inertial-only devices that provide better portability but lower precision. The paper reviewed common datasets (e.g., RWTH-BOSTON-50, Chalearn LAP IsoGD, HDM05) and concluded that wearable-based models outperform vision-based ones in environmental robustness and real-time performance. However, issues like high cost, reduced comfort, and lack of standardized evaluation methods remain major barriers. Building on these findings, [54] proposed WearSign, an end-to-end sign language translation system integrating a smartwatch and EMG armband. Unlike conventional models that recognise only isolated signs, WearSign translates continuous sign sequences into spoken text using a multi-task encoder decoder DL architecture. The model achieved a 4.7% word error rate for user-independent tests and 8.6% for unseen sentences, running in real time on a smartphone (<300 ms latency). Its main limitations include a small training dataset, limited vocabulary, and insufficient validation with expert signers.

The authors in [55] introduced a wearable ArSL recognition system using six IMUs: five on the fingers and one on the palm to capture dynamic hand motion. Their ANOVA-based feature selection method combined with an SVM classifier achieved 98.6% accuracy for user-dependent and 96% for user-independent testing. While the system performed well for alphabet-level recognition, scalability to continuous signing and comfort during extended wear remain key limitations. [56] developed a dual-arm wearable setup using Myo armbands equipped with both sEMG and IMU sensors for turkish sign language recognition. Testing across 80 words, the Random Forest algorithm achieved 99.875% accuracy, outperforming other classifiers. The dual-arm configuration captured bimanual gestures effectively, but reliance on commercial hardware and the lack of large-scale user trials limit broader applicability.

Liu et al. [57] proposed a lightweight chinese sign language system combining stretchable strain sensors and IMUs, with a CNN handling classification. Their wearable design, weighing under 20g, achieved 95.85% accuracy for isolated gestures and 84% for sentence-level recognition. Though comfortable and wireless, its accuracy declined with fast transitions and required user specific calibration. [58] designed a hybrid EMG–IMU framework for korean sign language recognition. Using a CNN–LSTM model, they achieved 98.1% accuracy overall and 94.3% in signer independent tests, demonstrating the value of MM fusion. However, high computational cost and electrode sensitivity hindered long-term usability. Similarly, the authors in [59] presented a fabric-integrated glove using piezoresistive textile sensors and IMUs for continuous ASL recognition. Their CNN–LSTM system attained 96.2% accuracy with 200 ms latency, offering comfort and flexibility.

The authors in [60] developed a CNN–LSTM model combining IMU and flex-sensor data for real-time ISL recognition. The system reached 97.8% accuracy for 30 words with 120 ms latency, confirming high temporal precision. Yet, its performance decreased across different

users due to limited dataset diversity. Chen et al. [61] introduced a hybrid EMG–IMU setup for continuous chinese sign language translation, achieving 96.3% signer-dependent and 91.2% signer-independent accuracy with an attention-based RNN. Though effective, it required careful electrode placement and had higher energy consumption in continuous operation. The authors in [62] proposed a low-power edge-based glove for ArSL, integrating flex sensors, an IMU, and an ESP32 microcontroller to perform on-device learning. Their quantized neural model achieved 95.7% accuracy with 50 ms latency and eliminated cloud dependency. However, the glove was limited to alphabet gestures and required individual calibration.

The authors in [63] presented a sensory glove for Moroccan Sign Language featuring five MPU6050 IMUs connected to a Raspberry Pi Pico running an embedded ANN. The system achieved 98% accuracy on 7,000 samples but focused only on alphabet level recognition. The authors in [64] designed a sensor-equipped glove for ArSL incorporating flex sensors, force-sensing resistors, and an MPU9250 IMU. Multiple algorithms were tested with extra trees achieving 100% accuracy on letters, numbers, and words. Despite strong results, the system lacked validation across users and was sensitive to glove fitting variations. Abdelrahman et al. [65] proposed a smart glove with flex sensors and an IMU for ASL alphabet translation using a CNN classifier. Achieving 97.2% accuracy and <150 ms latency, it proved effective for real-time use but was limited to single-user data and alphabetic gestures only.

The authors in [66] combined sEMG and IMU data for continuous ISL recognition using a CNN–LSTM model. The MM setup reached 96.8% accuracy for isolated gestures and 93.1% for continuous sequences, though power demands and calibration complexity limited practical use. The authors in [67] developed a low-cost Arduino-based glove for basic ASL recognition using flex, accelerometer, and gyroscope sensors. With a Random Forest classifier, it achieved 94.5% accuracy on a 10 word dataset. The system emphasized affordability and simplicity but was constrained by wired connections and limited vocabulary. Further, [68] presented an intelligent glove integrating flex sensors, motors, and vibration actuators for both recognition and learning of Korean Sign Language. Using an LSTM model trained on 20 signs, it achieved 85% accuracy while providing haptic feedback to guide learners. Despite its innovation in tactile learning, the glove's mechanical complexity and restricted vocabulary reduced scalability. Moreover, [69] further reviewed 88 EMG-based SLR studies, confirming the potential of EMG and IMU fusion for high-fidelity gesture capture. They identified LSTM as optimal for time-dependent EMG data and highlighted reproducibility challenges due to diverse datasets and sensor configurations.

2.3 RF-based Sign Language Recognition (SLR)

RF sensing has emerged as a promising modality for SLR and broader Human Computer Interaction (HCI). RF sensors operate in a contactless and privacy-preserving manner and are insensitive to lighting conditions, clothing variations, and visual clutter, which enables reliable

performance even in dark or crowded environments. Recent studies [70, 71] have shown that RF-based systems can recognise sign gestures and capture fine motion dynamics from a distance. Although RF sensors cannot directly observe hand shapes or facial expressions, they can precisely measure changes in range, angle, and velocity over time. By exploiting the micro-Doppler effect [72], RF systems extract motion signatures that describe detailed hand and finger trajectories during signing [73] and other gesture activities [74, 75], forming the basis for RF-based SLR. One of the earliest works in this domain is [76], which demonstrates that Wi-Fi signals can be used to recognise sign language gestures in a non-intrusive and device-free manner. Using three commercial Wi-Fi devices to record Channel State Information, the system applies Support Vector Machines and fuses decisions from multiple receivers to recognise distinct ASL gestures. WiSign achieves 93.8% accuracy with a 1.55% false-positive rate, confirming the feasibility of RF-based SLR without cameras or wearables. However, its evaluation on a small vocabulary under controlled indoor conditions and its reliance on multi device calibration limit its scalability to real-world, continuous signing.

Building on this idea, [77] significantly scales up the vocabulary to 276 ASL gestures using Channel State Information (CSI) from commodity Wi-Fi devices and a 9-layer CNN for end-to-end feature learning. With 8,280 samples across lab and home environments, SignFi reports accuracies above 94% in combined settings and clearly outperforms earlier Wi-Fi-based systems restricted to a few gestures. Nevertheless, the system's performance degrades when applied to unseen users, and the computational cost of the deep model poses challenges for real-time deployment on resource-constrained platforms. To better capture temporal dynamics in RF signals, Ahmed et al. [78] investigate LSTM networks on the SignFi dataset, comparing amplitude-only and joint amplitude phase CSI representations. Their results show that incorporating phase information with an LSTM architecture yields accuracies up to 99.8%, highlighting the benefit of sequence modeling for RF-based SLR. However, the study is limited to offline experiments on pre-collected data, without addressing robustness to environmental changes or user diversity in real deployments.

Extending RF-based SLR beyond ASL, DF-WiSLR [79] targets 49 ISL gestures using commercial Wi-Fi routers. The authors evaluate both classical ML models and an 8-layer CNN, and introduce Additive White Gaussian Noise (AWGN) augmentation to improve robustness despite a small dataset. Their approach achieves high accuracy, especially for static gestures, and demonstrates that Wi-Fi-based sensing can support both static and dynamic signs. However, the study involves a single participant and shows reduced performance for complex dynamic gestures, making generalization uncertain. While the above systems focus primarily on isolated gestures, WiSign [80] moves toward sentence-level recognition using Wi-Fi CSI. The framework combines Power Spectral Density (PSD)-based segmentation, Deep Belief Networks (DBNs), Hidden Markov Models (HMMs), and an N-gram language model to recognise continuous ASL sentences. The system achieves strong performance for user-specific models but shows a sub-

stantial drop in accuracy for generalized, multi-user scenarios. Its multi-stage pipeline is also computationally demanding and sensitive to environmental variations, which complicates practical deployment. Alternative RF technologies such as millimeter-wave are explored in mmASL [81], which uses 60 GHz mmWave signals and a multi-task DL model to recognise 50 ASL signs across different rooms and users. The system provides robust, device-free interaction and is less affected by lighting and background, making it attractive for smart assistant applications. However, its limited vocabulary, high computational cost, and inability to capture detailed facial or finger configurations highlight persistent challenges of mmWave-only SLR.

Gurbuz et al. [82] showed that RF sensing can understand details of sign language by using multi-frequency radar to capture small motion patterns (micro-Doppler signals) from native ASL signers. Their study found that RF signals can tell the difference between fluent and imitation signing, and can also recognise a small set of ASL words. This suggests that RF sensing could enable privacy-friendly technology designed for deaf users. However, their work used only a small set of words and a limited dataset, showing that larger and more diverse RF sign language datasets are still needed. Kurtoglu et al. [83] studied how to detect ASL trigger signs using Frequency-Modulated Continuous Wave (FMCW) radar while people perform different activities. They used a learning model that combines multiple types of input and tasks, achieving high accuracy in detecting triggers and recognising gestures. This shows that RF sensing could work as a "wake word" system for Deaf-friendly technology. However, their experiments were done in controlled settings with only a few signs, so it is still unclear how well the system would work in real-world use. To address the gap between directed lab data and natural signing, Kurtoglu et al. [84] propose ChessSIGN, an interactive radar- and video-based platform that collects unscripted ASL data during gameplay. By leveraging 77 GHz radar micro-Doppler signatures and physics-aware generative adversarial networks for data augmentation, the system substantially improves recognition of natural signs compared to models trained only on directed recordings. Even so, accuracy on natural signing remains moderate, and the absence of non-manual features such as facial expressions restricts full linguistic coverage.

Several works refine radar-based methods for both recognition and domain adaptation. Rahman et al. [85] use 77 GHz FMCW radar with a multi-branch GAN-based domain adaptation framework to jointly address word-level ASL recognition, sequence classification in continuous streams, and trigger detection. Their results show strong performance but rely on specialized radar hardware and relatively small datasets. In parallel, Kulhandjian et al. [86] use 24 GHz Doppler radar with CNNs on micro-Doppler spectrograms to recognise a small set of signs, achieving high accuracy and robustness to lighting but requiring careful calibration and offering limited vocabulary. Within the Wi-Fi domain, Wi-SignFi [87] introduces a multitask framework that recognises gestures while also identifying user and environment attributes using an efficient CNN–KNN hybrid and grayscale CSI representations. This design supports deployment on embedded devices and confirms that RF-based SLR can be integrated into IoT platforms;

however, it is still evaluated under controlled conditions and does not fully address continuous sentence recognition. Similarly, WiASL [88] focuses on micro-motion based ASL letter input using amplitude–phase fusion, micro-motion detection, and a Bi-LSTM–CNN–attention network. It delivers accurate and efficient contactless text entry but is sensitive to environmental noise, fine-grained gesture similarity, and potential user fatigue.

Beyond Wi-Fi and radar, RF-Sign, the authors in [89] explores passive RFID tags for position-independent finger-level SLR. By modeling hand position and using a reference tag for dynamic segmentation, RF-Sign reconstructs phase features that are robust to changes in hand location, achieving high recognition accuracy with low power consumption. However, the requirement for worn tags and residual sensitivity to multipath effects limit its practicality in fully device-free scenarios. Emerging systems such as mm-SLR [90] and WiBaSL [91] push RF-based SLR into more dynamic and diverse contexts. mm-SLR employs 77 GHz mmWave radar and temporal convolutional networks to recognise gestures reliably in cluttered environments, while WiBaSL provides the first Wi-Fi CSI dataset for Bangladeshi Sign Language with a CNN–LSTM model achieving strong performance. Both works broaden the scope of RF-based SLR but inherit challenges related to dataset size, generalization across environments, and the absence of full linguistic information.

Complementary DL approaches, such as the ensemble framework in [92], achieve near-perfect accuracy on benchmark CSI datasets using combinations of CNNs, LSTMs, and attention mechanisms. However, these models typically assume stable conditions and require substantial computational resources. In contrast, lighter-weight methods like the bispectrum and SVM-based approach in [93] offer efficient recognition using higher-order spectral features but do not scale well to continuous or highly varied signing. Finally, cross-domain frameworks such as CSI-Cro [94] employ dual-attention feature fusion and domain adaptation to maintain performance across different environments and hardware, addressing one of the key practical barriers, albeit with increased model complexity. Taken together, these studies show a clear progression: from early Wi-Fi CSI prototypes with small vocabularies, to large-vocabulary CNN-based systems, radar and mmWave solutions robust to lighting and occlusion, RFID-based fine-motion tracking, and cross-domain and natural-signing aware architectures. RF-based SLR now offers strong evidence of its potential as a contactless, privacy-preserving alternative or complement to vision-based methods. However, persistent challenges remain in scaling to continuous, large-vocabulary, signer-independent recognition and in incorporating non-manual linguistic cues that are essential for full sign language understanding.

Table 2.1 compares existing approaches in terms of modality, performance, and limitations. It shows that vision-based systems achieve high accuracy but are sensitive to lighting and privacy issues, while radar-based methods are more robust and privacy-friendly. Multimodal approaches combine the strengths of both to improve overall performance.

Table 2.1: Comparison of Sign Language Recognition Techniques

Technique	Advantages	Limitations	Relevance to This Work
Vision-based	Captures detailed spatial features such as hand shape and facial expressions; high recognition accuracy	Sensitive to lighting, occlusion, and background clutter; raises privacy concerns	Used to extract rich visual features of gestures
RF-based	Contactless and privacy-preserving; robust to lighting, occlusion, and clothing variations	Cannot capture fine details such as finger shape or facial expressions	Used to capture motion patterns via micro-Doppler signatures
Wearable-based	High precision and direct motion sensing; less affected by environmental conditions	Intrusive; requires users to wear devices; less practical for daily use	Not used due to lack of user convenience
Multimodal (Vision + RF)	Combines strengths of multiple modalities; improved robustness and accuracy	Increased system complexity and data fusion requirements	Adopted approach in this work

2.4 Machine Learning

ML refers to the capability of computers to automatically identify patterns within data and utilise these patterns to make predictions or informed decisions [95]. It has become a foundational technology in numerous modern applications, including autonomous driving, speech processing, and intelligent systems [96]. A wide range of ML algorithms has been developed, each tailored to address specific categories of problems [97]. These algorithms construct models by analysing training datasets, which provide representative examples that enable the system to recognise underlying structures and relationships within the data. Once trained, these models can be applied to previously unseen inputs, allowing them to generalise beyond the examples encountered during training. For instance, when an ML model is trained on sensor data corresponding to particular hand gestures, it can subsequently identify similar gestures in new or unfamiliar samples. In its more advanced form, ML encompasses DL, which employs ANNs inspired by the architecture and learning mechanisms of the human brain to achieve higher levels of abstraction and accuracy.

2.4.1 Advantages of Machine Learning

One of the major advantages of ML is that it removes the need to manually program explicit rules for recognising complex patterns. Instead, ML algorithms can automatically learn and extract

meaningful relationships directly from data through exposure to examples. This capability not only reduces the time and effort required for manual feature engineering [98], but also allows the detection of subtle or latent patterns that may be difficult or impossible for humans to identify through traditional analytical methods [99].

Moreover, ML systems continue to improve as they receive more data, which makes them highly adaptable in situations where conditions or rules change frequently [100]. ML models also show strong generalisation ability, meaning they can make accurate predictions on new data after being trained on suitable examples [101]. In addition, ML supports large-scale automation, allowing organisations to analyse large amounts of data efficiently and with less manual effort [102]. These advantages make ML a valuable tool in many fields, including healthcare, manufacturing, finance, and cybersecurity [103].

2.4.2 Machine Learning Approaches

Machine Learning (ML) techniques are commonly divided into three main categories: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, algorithms are trained using labelled datasets, where each input is associated with a corresponding output. The model learns from these examples to accurately classify or predict outcomes for new, unseen data [104]. In contrast, unsupervised learning deals with unlabelled data, where algorithms aim to uncover hidden structures or patterns within the dataset, often by grouping similar samples into clusters based on shared characteristics [105]. Meanwhile, reinforcement learning follows a reward-driven learning paradigm, in which an agent interacts with its environment by performing actions and receiving feedback in the form of rewards or penalties. Through repeated interactions, the agent learns optimal strategies that maximise cumulative reward over time [106]. This learning framework has demonstrated remarkable success in complex decision-making tasks such as autonomous driving and strategy games like chess, where algorithms have achieved superhuman performance [107].

2.5 Deep Learning

Deep learning is a branch of machine learning that employs ANNs designed to mimic the information-processing capabilities of the human brain [108]. These networks are typically structured with an input layer, multiple hidden layers, and an output layer [109]. The term depth refers to the number of hidden layers within the network [110]. Each hidden layer processes data by applying weights to the inputs and transmitting the transformed outputs forward. During training, the network's predictions are compared with the target values, and the weights are iteratively adjusted to minimize the error [111, 112].

This optimisation is commonly performed using the backpropagation algorithm, which systematically propagates the error backward through the network to update the weights [102]. The

cycle of feeding data through the network, computing errors, and refining weights is repeated across multiple iterations, known as epochs [113]. Through sufficient training, deep neural networks learn to recognise complex, non-linear relationships within data, making them highly effective for tasks such as image classification, natural language processing, and human activity recognition [114].

2.5.1 Advantages of Deep Learning

DL offers several important advantages that make it highly effective for complex tasks. One key strength is its ability to automatically learn useful features directly from raw data, reducing the need for manual feature engineering [102]. DL models can capture highly complex, nonlinear patterns, allowing them to outperform traditional methods in areas such as image recognition, speech processing, and natural language understanding [101]. Furthermore, DL systems continue to improve as more data becomes available, enabling them to reach high levels of accuracy in large scale applications [115]. DL models are also highly flexible and can be adapted to different types of data, including text, audio, images, and sensor signals [116]. Because of these advantages, DL has become a leading approach across many fields, including healthcare, autonomous driving, robotics, and cybersecurity [117].

2.5.2 Application of Machine Learning and Deep Learning

In this work, ML and DL are used to enable accurate recognition of sign language gestures from both radar and video data. The radar signals are converted into micro-Doppler spectrograms, which capture detailed motion patterns over time and frequency. Deep learning models, particularly convolutional neural networks, are well suited for learning meaningful features from these spectrograms automatically, without the need for manual feature design.

Similarly, ML/DL methods are applied to video data to extract important spatial information related to hand and body movements. Their ability to model complex patterns and generalise across different participants makes them highly effective for achieving robust and reliable sign language recognition in this study.

2.5.3 Summary

This chapter reviewed various MM sensing technologies developed to assist individuals with hearing impairments by facilitating both verbal and non-verbal communication. Four main modalities were discussed; vision-based, wearable sensor-based, and RF-based systems. Each modality offers unique strengths and faces specific limitations. Vision-based systems achieve high gesture and facial recognition accuracy using DL models; however, they are sensitive to lighting, background clutter, and privacy concerns due to continuous video recording. Wearable

sensor-based systems, such as smart gloves, provide precise motion capture but can be uncomfortable, require frequent calibration, and are impractical for continuous or public use.

In contrast, RF-based sensing presents a promising contactless alternative that works reliably under various lighting and environmental conditions, maintaining user privacy while detecting subtle motion dynamics. Additionally, ML and DL techniques were highlighted as key enablers for processing and classifying MM data, enhancing recognition accuracy across all sensing types. Overall, while vision, and wearable systems have advanced MM communication, they remain limited by environmental and usability challenges. RF-based sensing offers a robust path forward for developing next-generation hearing assistive technologies that are more adaptive, non-intrusive, and user-friendly.

Chapter 3

Methodology

This chapter presents the methodologies used in the three research works that form this thesis. Although each work focuses on a different research problem, all studies follow a common approach based on contactless sensing, signal processing, and deep learning. The chapter summarises the main contributions of each paper and explains the experimental setup, data collection, signal processing, performance evaluation and DL classification models. In addition, this chapter highlights the similarities and differences between the proposed methodologies and explains how each study extends the overall research framework. This structured presentation helps to clarify the methodological progression across the thesis and supports the interpretation of the results presented in later chapters.

3.1 Performance Comparison of Video and Radar for Common Signs Recognition in BSL and ASL

3.1.1 Introduction

Sign language is an essential form of communication for individuals who are Deaf or hard of hearing, using hand gestures, facial expressions, and body movements. Automatic Sign Language Recognition (SLR) has therefore become an active area of research, with increasing interest in developing technologies that can translate non-verbal communication into spoken or written language. Existing approaches for capturing sign language are broadly categorised into vision-based, wearable, and contactless sensing methods [118].

As sign languages vary across regions, it is important to develop systems that can support multiple sign language types. In this work, radar sensing is explored as a contactless and privacy-preserving approach, alongside video-based sensing, to recognise 15 common signs shared between American Sign Language (ASL) and British Sign Language (BSL). The performance of radar-only and video-only systems is analysed to better understand the strengths of each sensing modality. Following are the key contributions of this work:

- Proposed a contactless recognition system to recognise and translate 15 common signs shared between ASL and BSL, making the system useful across multiple sign language communities.
- A robust and diverse dataset is collected that comprises of 3,600 samples (1,800 radar + 1,800 video) from 15 classes from four users (2 males and 2 females). The data is collected in different days under varying conditions (lighting, clothing, hairstyles, time of day).
- Performed unimodal analyses of radar-only and video-only models to assess their independent contributions, establishing a proof-of-concept for the integration of such systems into future hearing aid technologies.
- Implemented DL models (ResNet18) for sign classification to evaluate the performance of radar-based and video-based recognition approaches.
- Established a foundation for the development of MM, privacy-preserving, and scalable sign language recognition systems, with potential applications in assistive technologies for individuals with hearing impairments.

3.1.2 Experimental Setup and Data Acquisition

To investigate the effectiveness of two technologies, i.e., camera and radar for the classification of common signs in ASL and BSL, we have conducted some experiments with the UWB radar alongside the built-in camera of the laptop, as illustrated in Fig. 3.2. At the start of the data collection process, volunteers were shown target signs and asked to perform them in a 6-second window. The chair, where the participant was seated, was placed 141 cm from the sensors during data capture. Additionally, data were collected over multiple days at various times, under varying conditions (weather, lighting, clothing, hairstyles, time of the day) to capture diverse data. Data for 15 common signs in ASL and BSL (Fig.) were collected from four volunteers (two male and two female). Each sign was performed 30 times by each participant, resulting in a total of 3,600 data samples (1,800 from radar and 1,800 from video).

3.1.2.1 Data Collection using UWB radar

The experimental setup is depicted in Fig. 3.2, where the UWB radar module is precisely mounted above the laptop's integrated camera to enable synchronized data capture. Data acquisition was conducted in a standard indoor environment under typical ambient conditions. The setup was designed to account for environmental factors such as background noise, reflections from surrounding walls and objects, variations in lighting, and other potential sources of interference to ensure robustness of the collected dataset. During data collection, participants

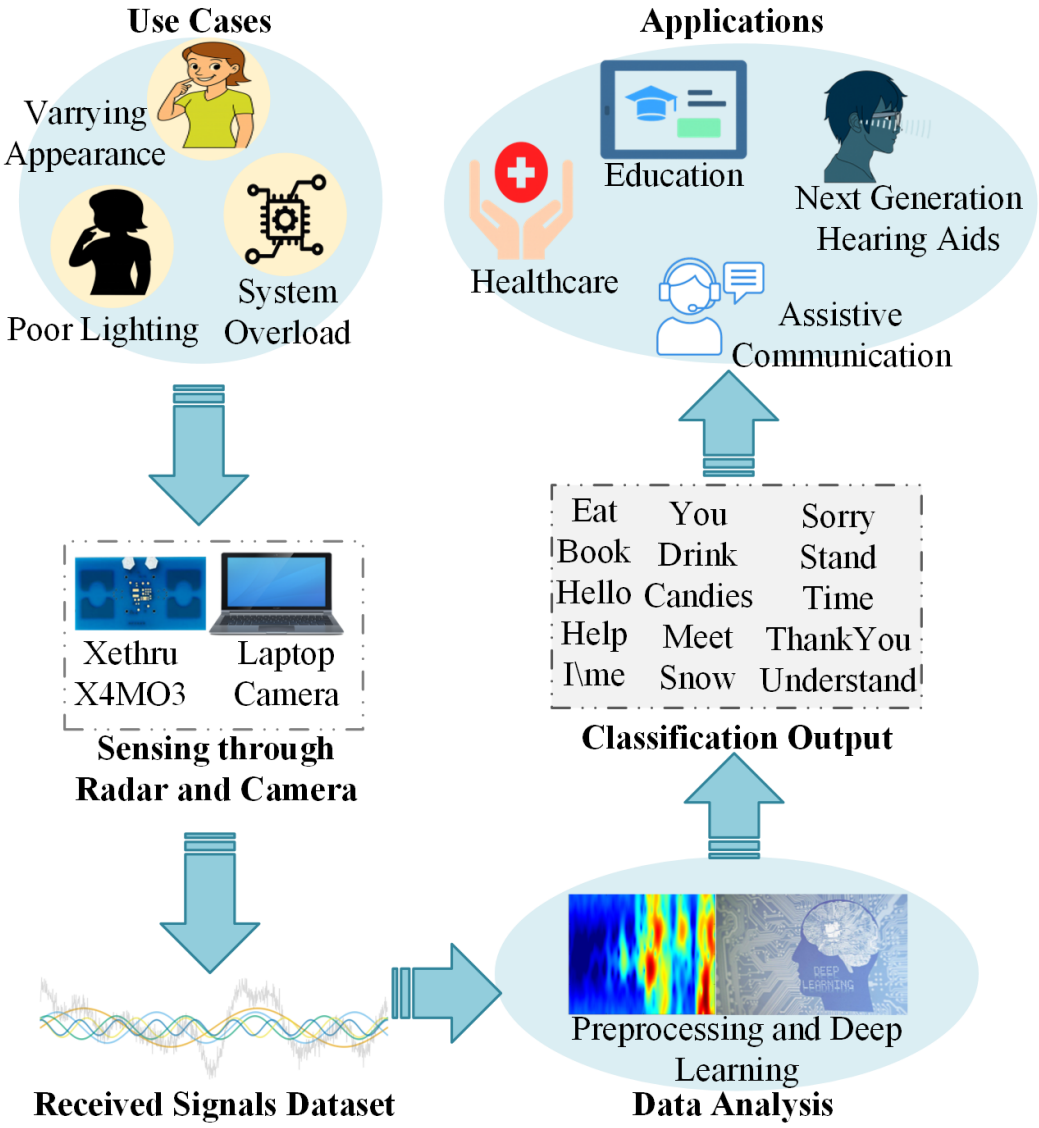


Figure 3.1: Workflow System Diagram of Proposed Work

were seated 141 cm in front of the sensors. The UWB radar, equipped with one transmitter and one receiver antenna, has a detection range of 9.6 meters. It operates with a transmit power of 6.3 dBm, and its Effective Radiated Power (ERP) complies with IEEE standards for UWB devices, ensuring minimal interference with other systems while maintaining operational safety. At the measurement distance of 141 cm, the radar exhibits a power density of approximately 1.68 W/m², which remains significantly below the IEEE C95.1-2005 recommended exposure limits of 50 W/m² for controlled environments and 10 W/m² for uncontrolled environments [119].

3.1.2.2 Data Collection using Camera

The built-in camera of laptop is used to capture the samples as full frame, RGB .avi files that give frame dimensions of $3 \times 1024 \times 576$ and a frame rate of 30 Frames Per Second (FPS). The

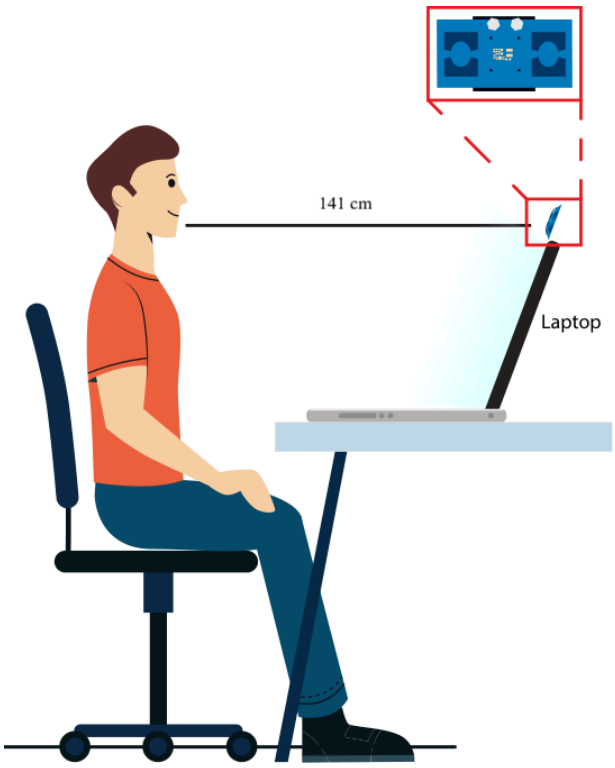


Figure 3.2: Setup for Data Collection

large size, high dimensionality, and rich color details in such files often include vast amounts of information that are irrelevant for sign language classification. Moreover, this excess data can place a heavy burden on computational resources and model capacity, potentially leading to inefficiencies or system overload. Full-frame videos are preprocessed to reduce their size and then converted to grayscale, resulting in dimensions of $1 \times 450 \times 165$ at 15 FPS. The processed video samples retain sufficient visual and temporal information to support effective model training and accurate prediction, as discussed in the Section 4.1.1.

3.1.3 Data Preprocessing and Evaluation Metrics

3.1.3.1 Radar Data Preprocessing

In this section, the preprocessing steps used for extracting spectrograms from radar bin files are described. Initially, the radar chip was configured via the XEP interface using the corresponding driver. Data were collected at a rate of 500 frames per second (FPS), with each sample encoded as a 32-bit floating-point value. The data file was then read iteratively and stored in a *DataStream* variable, which was subsequently converted into a complex Range-Time-Intensity (RTI) matrix. To suppress static clutter, a Moving Target Indication (MTI) filter was applied, ensuring that only moving targets contributed to the signal. This enhances the visibility of Doppler-induced frequency shifts. A filtering stage was then used to generate the Doppler-range map, followed by a secondary fourth-order MTI filter to refine the spectrogram.

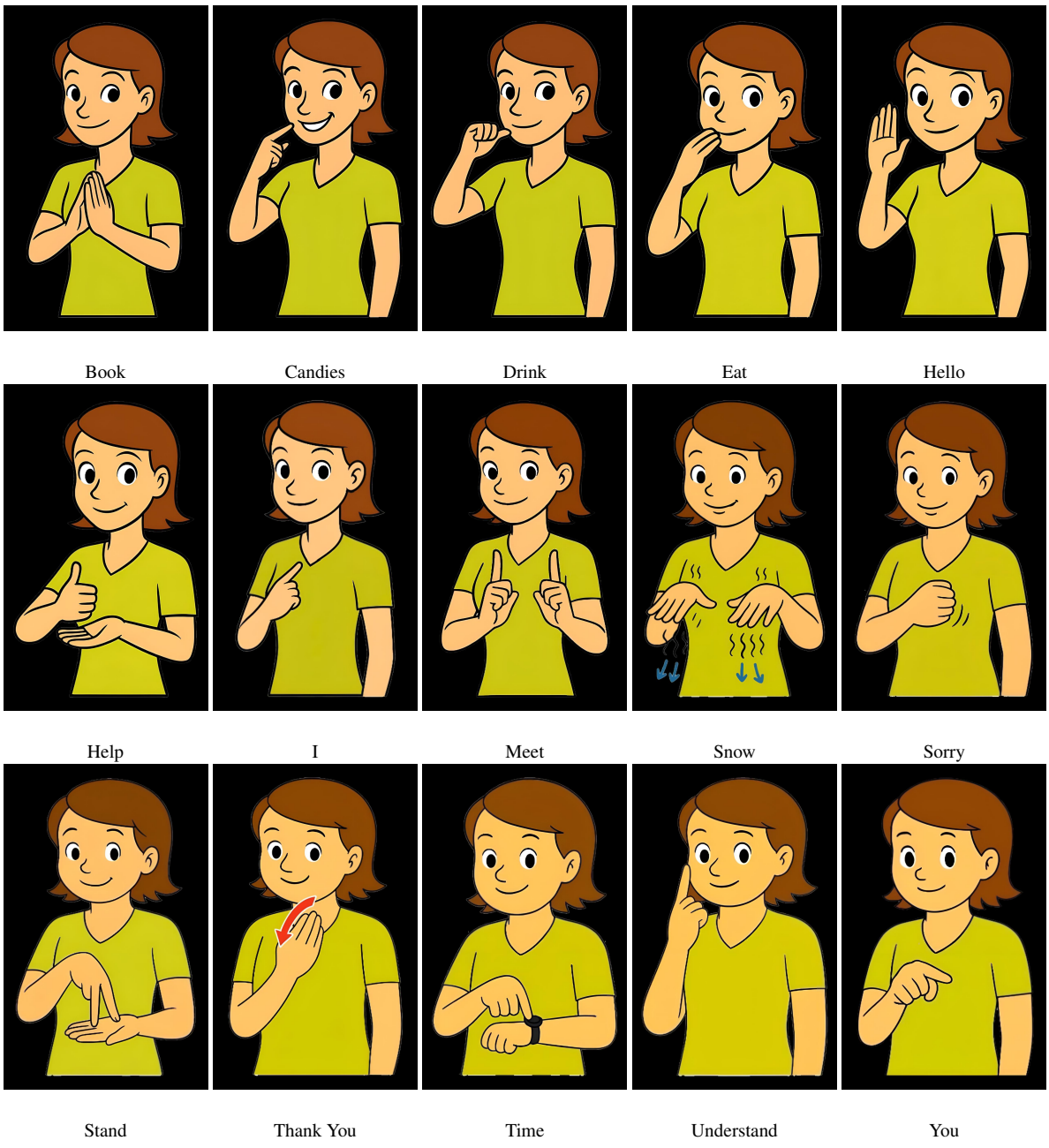


Figure 3.3: 15 Common Signs in ASL and BSL

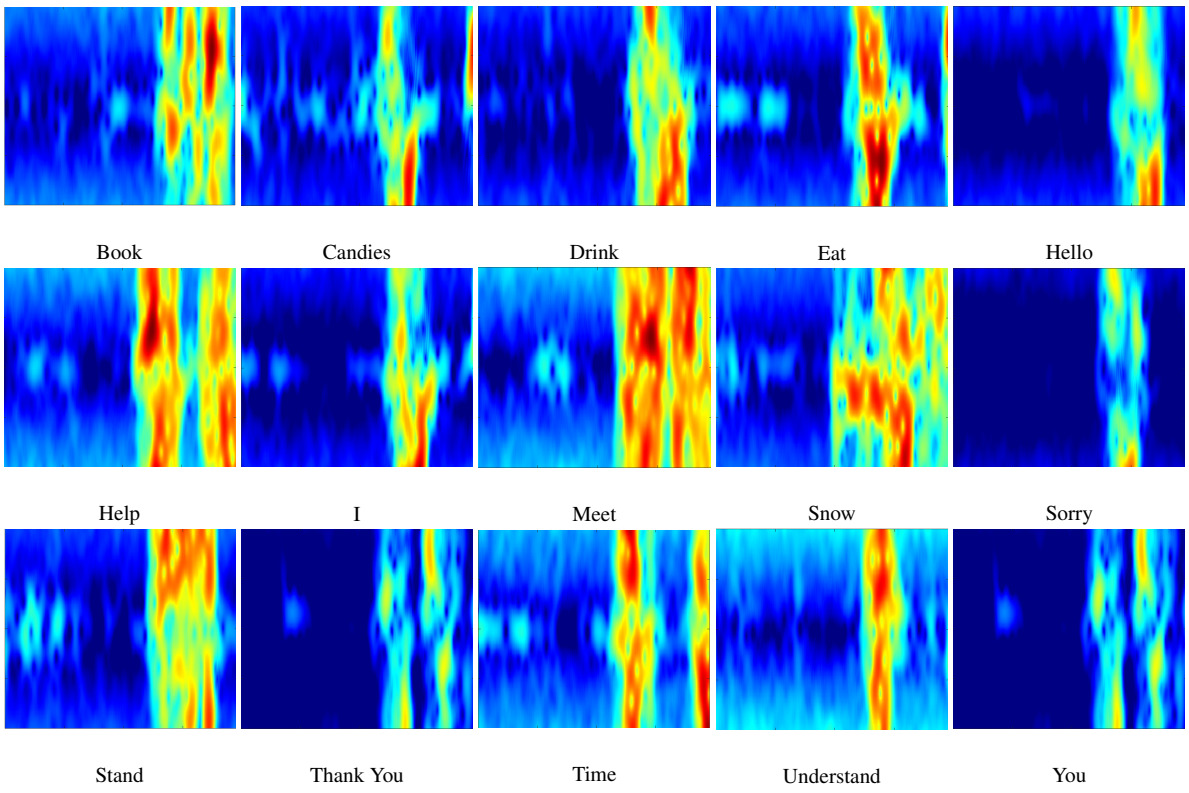


Figure 3.4: Spectrograms of 15 Common Signs in ASL and BSL

Several parameters were considered during processing, including a window length of 128 samples, overlap percentage, and a Fast Fourier Transform (FFT) zero-padding factor of 16. A range profile was obtained by applying an FFT to each chirp. Subsequently, a second FFT was performed across consecutive chirps for each range bin to extract Doppler information. Spectrograms were generated using the Short-Time Fourier Transform (STFT), which provides a joint time-frequency representation. Unlike the standard Fourier Transform (FT), which only provides frequency information, STFT enables analysis of temporal variations in frequency. The segmentation of signals allows FT to be applied locally, capturing both temporal and spectral characteristics. The choice of window length introduces a trade-off between time and frequency resolution.

The maximum unambiguous Doppler frequency is given by:

$$f_{d,max} = \frac{1}{2T_r} \tag{3.1}$$

where T_r denotes the chirp repetition interval.

The transmitted signal can be expressed as:

$$T_s(t) = A\cos(2\pi ft) \tag{3.2}$$

where A is the signal amplitude and f is the carrier frequency.

The received signal from a target at range $D(t)$ is:

$$R_s(t) = A_r \cos\left(2\pi f\left(t - \frac{2D(t)}{c}\right)\right) \tag{3.3}$$

where:

- A_r is the received signal amplitude,
- c is the speed of light,
- $D(t)$ is the time-varying distance between the radar and the target.

For a moving target with radial velocity $v(t)$, the received signal can be approximated as:

$$R_s(t) \approx A_r \cos\left(2\pi(f + f_d)t - \frac{4\pi D_0}{c}\right) \tag{3.4}$$

where D_0 is the initial range.

The Doppler shift is defined as:

$$f_d = \frac{2fv(t)}{c} \tag{3.5}$$

The received signal consists of contributions from multiple moving scatterers (e.g., head, hands, and torso). Thus, it can be expressed as:

$$R_s(t) = \sum_{k=1}^{N} A_k \cos\left(2\pi(f + f_{d,k})t - \frac{4\pi D_k}{c}\right) \tag{3.6}$$

where each component corresponds to an individual moving body part with distinct velocity and range. These micro-Doppler signatures collectively form unique patterns that enable activity classification. Spectrograms derived from these signals were used to construct a dataset, which was divided into training and testing subsets. Deep learning (DL) pre-trained models were then employed to analyse the spectrograms and classify human activities.

3.1.3.2 Video Data Preprocessing

The built-in laptop camera was used to capture video samples as full-frame RGB .avi files with dimensions of $3 \times 1024 \times 576$ at 30 FPS. Due to the high dimensionality and rich color details, these raw files contained substantial redundant information that was not directly relevant to sign language classification, while also imposing heavy computational demands on system resources. To address this, the videos were preprocessed by converting frames to grayscale and resizing them to $1 \times 450 \times 165$ at 15 FPS. This reduced both spatial and temporal resolution, thereby lowering computational overhead while preserving the essential visual and temporal features required for effective model training and accurate recognition.

Table 3.1: Parameters Setting of Radar

Parameter	Value
Activity duration	6 seconds
Radar Sensor	XeThru X4M03
Instrumental range	9.6 metres
Radar's distance from subject	141 cm
Radar frequency	7.29 GHz
Radar bandwidth	1.5 GHz
Tx power	6.3 dBm

3.1.3.3 Evaluation Metrics

The efficacy of DL models in gesture classification is assessed based on weighted average accuracy (degree of closeness between a measured or calculated value and its true or accepted value), precision, recall, and F1 Score. F1 Score, a widely adopted metric in classification literature, is computed using Equation 3.7. Precision and recall, integral components of F1 Score, are determined using 3.8 and 3.9, respectively.

$$F1 = 2\frac{Precision.Recall}{Precision + Recall} \tag{3.7}$$

$$Precision = \frac{\sum TruePositive}{\sum TruePositive + \sum FalsePositive} \tag{3.8}$$

$$Recall = \frac{\sum TruePositive}{\sum TruePositive + \sum FalseNegative} \tag{3.9}$$

The radar-based system's performance is influenced by specific hardware constraints. The Ultra-Wideband (UWB) radar used in this study has limited spatial resolution compared to vision-based systems, affecting its ability to detect fine facial movements accurately. The radar's Doppler sensitivity depends on its sampling rate and frame capture speed (500 FPS), which limits the accurate detection of rapid expressions or gestures. Additionally, the radar's performance is highly dependent on the angle at which it detects reflections; changes in subject orientation relative to the radar can weaken or alter Doppler signatures, unlike camera-based systems that capture full visual detail regardless of position.

Another factor is multipath reflections, where radar signals bounce off multiple surfaces before reaching the sensor. This can introduce noise into the Doppler signatures and impact classification accuracy. Despite these constraints, the radar-based approach remains advantageous for privacy-preserving, contactless recognition, with its resilience in low-light and through-surface scenarios making it suitable for diverse environments.

Table 3.2: Fine-tuned parameters for the selected models [120, 121]

DL Model	Parameters	Settings
GoogleNet	Learning_Rate	0.0001
	Batch_Size	16
	Learning_Algorithm	Adam
	Loss_Function	Cross entropy
	Total_Epochs	100
	Iteration_Per_Epoch	75
SqueezeNet	Learning_Rate	0.0001
	Batch_Size	16
	Learning_Algorithm	Adam
	Loss_Function	Cross entropy
	Total_Epochs	200
		75
VGG16	Learning_Rate	0.0001
	Batch_Size	16
	Learning_Algorithm	Adam
	Loss_Function	Cross entropy
	Total_Epochs	100
	Iteration_Per_Epoch	75
VGG19	Learning_Rate	0.0001
	Batch_Size	16
	Learning_Algorithm	Adam
	Loss_Function	Cross entropy
	Total_Epochs	100
	Iteration_Per_Epoch	75

3.1.4 ResNet for Sign Classification

Convolutional Neural Networks (CNNs) are widely regarded as the state-of-the-art approach for image-based classification tasks, as they are capable of extracting hierarchical spatial features from complex input data. A significant advancement in CNN architectures was the introduction of the Residual Network (ResNet) [122], designed to mitigate challenges associated with training very deep neural networks, particularly the issues of vanishing and exploding gradients.

3.1.4.1 Design Philosophy

The core concept of ResNet is the residual block, where the network learns residual mappings rather than direct input to output transformations. Formally, instead of approximating $H(x)$ directly, each residual block learns $F(x) = H(x) - x$, such that the output becomes $H(x) = F(x) + x$. This is achieved through identity (skip) connections, which allow gradients to flow more effectively during backpropagation, enabling stable training of deeper networks.

3.1.4.2 Residual Block Structure

A standard residual block comprises convolutional layers with batch normalization and ReLU activation, augmented with shortcut connections. For deeper variants such as ResNet50 or ResNet101, a bottleneck design is employed, incorporating a 1×1 convolution for dimensionality reduction, a 3×3 convolution for feature extraction, and a final 1×1 convolution to restore dimensionality. This structure allows the network to balance depth with computational efficiency.

3.1.4.3 Application to This Work

In this study, ResNet18 was employed as the backbone for classifying spectrograms (radar modality) and grayscale video frames (video modality). ResNet18 was selected due to its computational efficiency and strong capability for spatial feature extraction. For radar spectrograms, which exhibit relatively simple spatiotemporal structures, the shallow depth of ResNet18 provided sufficient representational capacity without risk of overfitting. For video frames, ResNet18 served as an effective feature extractor, though video's temporal dynamics remain a limitation when using static frame-based inputs.

3.1.4.4 Performance Considerations

ResNet18 demonstrated robust performance across both modalities, with radar-based classification achieving 96% test accuracy, benefiting from stable Doppler signatures, while video-based classification achieved 81% test accuracy, reflecting its greater sensitivity to environmental conditions such as lighting and background variation. The skip-connection architecture of ResNet

ensured stable convergence and mitigated gradient degradation, making it a suitable choice for this work.

3.2 Utilising Contactless Sensing Technology for the Identification of Hand and Head Movements in Conjunction with Facial Expressions

3.2.1 Introduction

Sign language recognition has emerged as a vital research challenge, with the objective of understanding communication through hand gestures, head movements, and facial expressions. However, most existing systems, such as vision-based, and wearable sensor-based technologies, face major limitations including privacy concerns, lighting dependence, noise issues, and maintenance difficulties. This work introduces a contactless radar-based approach that recognise different expressions performed with head, hand movements, and facial expressions by leveraging micro-Doppler signatures obtained from data collected using a radar sensor. In this work, sixteen different classes are being considered, i.e., Ashamed, Cheerful, Enormal, Furious, GoodIdea, Guilty, Lonely, Normal, Ok, Playful, Proud, Sad, Shocked, Surprised, Thinking, and Worried. An UWB Radar was employed to record the data. Four DL models are used to classify these sign. Spectrograms serve as representations of the recorded data, while spatio-temporal features are extracted using GoogleNet, SqueezeNet, VGG16, and VGG19 architectures.

Main contributions of this work are discussed below:

- We proposed a contactless head, hand movement, and facial expressions recognition system to recognise and translate expressions.
- A dataset is collected that contains 1440 samples from 16 different classes at a distance of 141 cm. The data is collected from two males and 1 female ages between 20 and 40.
- We applied four pre-trained DL models, i.e., GoogleNet, SqueezeNet, VGG16, and VGG19 on the dataset, offering a benchmark for future research in this field.
- VGG16 achieved the highest performance, with an accuracy of 94.2% across 16 expression classes.

3.2.2 Methodology

In this section, the proposed framework is explained in detail which is illustrated in Fig 3.5. It is divided into three phases, i.e., data collection, signal processing and utilisation of DL models. All phases are explained in detail in next sub-sections.

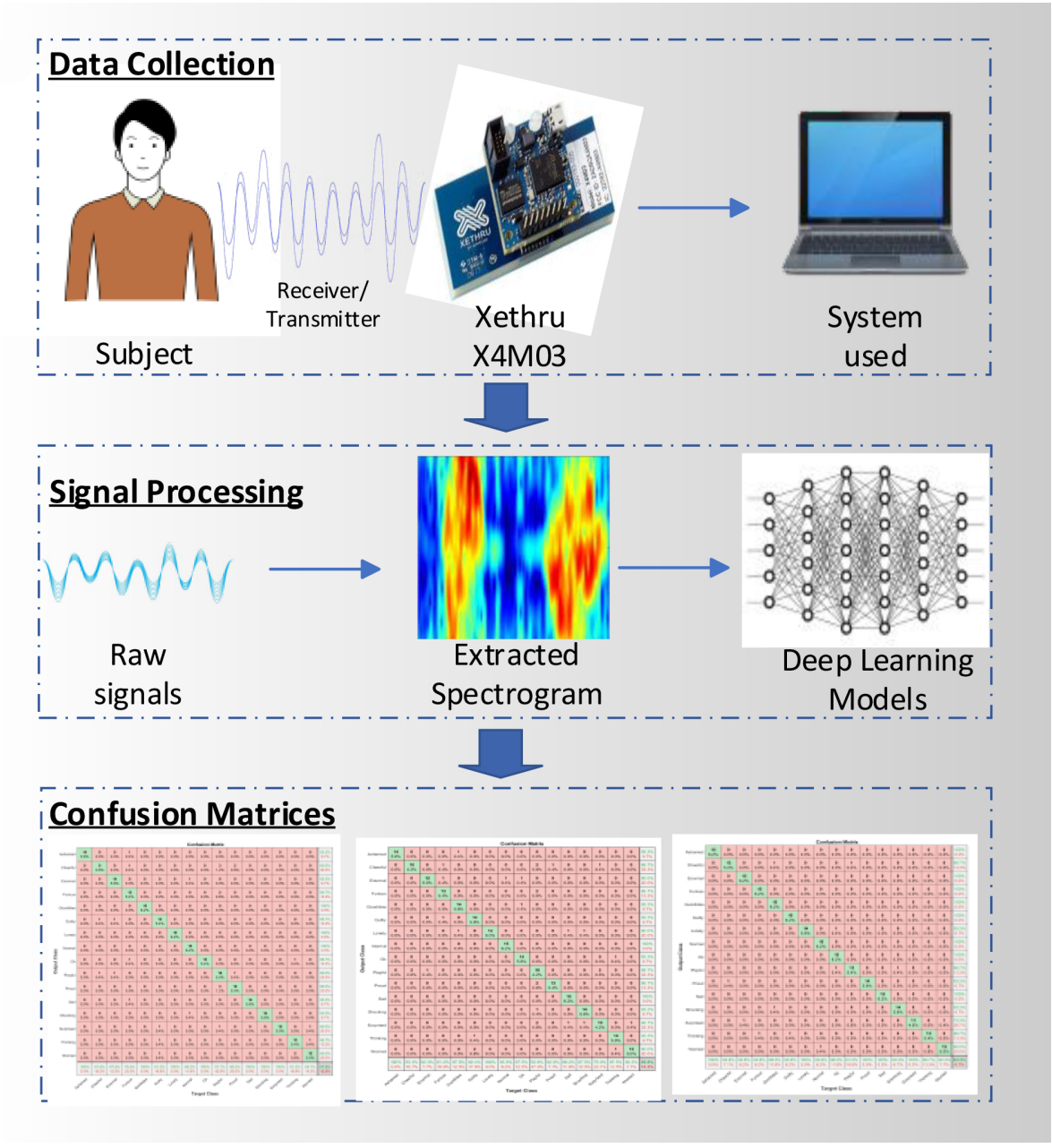


Figure 3.5: Proposed Workflow System Diagram

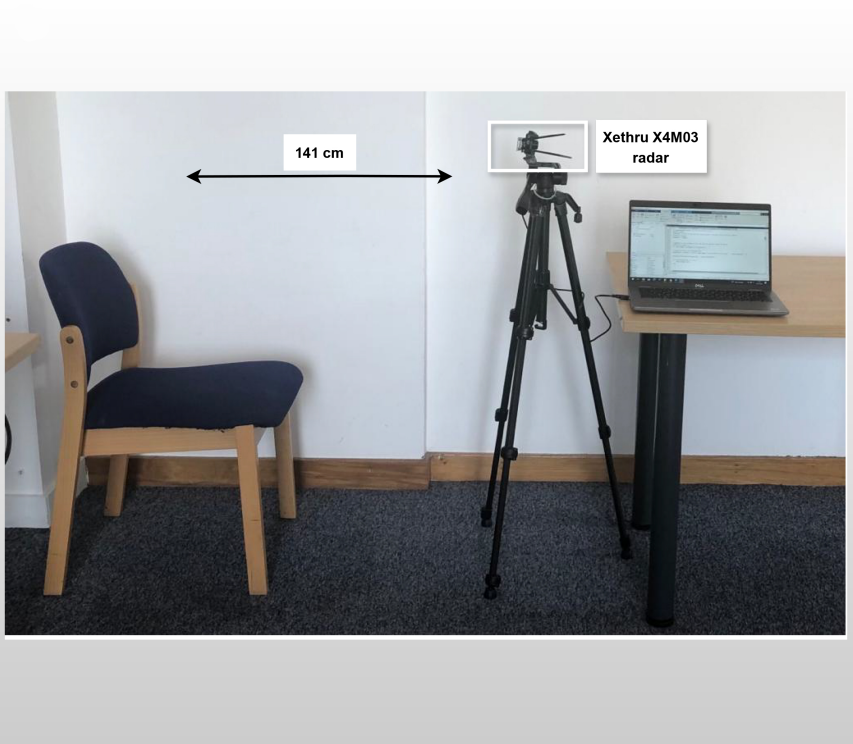


Figure 3.6: Experimental Setup

3.2.2.1 Experimental Setup and Data Collection

In the data collection phase, an UWB radar, namely Xethru X4MO3 is employed to recognise hand, head and facial expressions which is connected to a laptop where collected data is being stored. It has one Transmitter (Tx) and one Receiver (Rx) antenna, offering a maximum human motion detection range of 9.6 meters with all major and minor movements. Further, the Xethru app is installed on the laptop that operates the radar. The radar is mounted on a tripod stand (ensuring stability during the data collection process) which is placed in front of the chair where the user is sitting. The parameter setting of radar is given in Table 3.1. The distance between the radar and the seated participant was set to 141 cm, chosen to represent a typical one-to-one communication range, as non-verbal interactions generally occur at close distances. This distance enables effective gesture capture while ensuring reliable radar signal quality for optimal recognition and interaction which is shown in Fig 3.6. Our study was conducted exclusively in indoor environments due to the limitations of the radar equipment and experimental setup, which are not suited for outdoor conditions. Factors such as uncontrolled environmental variables, signal interference, and the need for stable radar positioning make outdoor experiments impractical. We collected data from a single user over multiple days, introducing natural variations in clothing, hat usage, and occasional changes in the chair. Despite these fluctuations, our system performed well, demonstrating its robustness in handling real-world changes. The data collection took place in an open indoor environment on the sixth level of the James Watt South Building at the University of Glasgow. This setting allowed people to move freely around the

radar setup, simulating real-world scenarios. Furthermore, data has been gathered from a total of three users, two males and one female, covering 16 distinct classes (activities) as shown in Fig 3.7. The age range of the users falls between 20 and 40 years. This study is conducted as a proof-of-concept to validate feasibility under controlled conditions. Every activity/gesture is performed 30 times, with each instance lasting for a fixed duration of 6 seconds. The decision to involve a larger number of participants was driven by the aim to enhance the dataset's realism and diversity. A dataset contains 1440 data samples that were collected during the experiment for 16 different classes, i.e., Ashamed, Enormal, Furious, GoodIdea, Guilty, Lonely, Normal, Ok, Playful, Proud, Sad, Shocked, Surprised, Cheerful, Thinking and Worried. In essence, each of these signs plays a vital role in daily interactions, helping individuals navigate emotions, communicate effectively, build relationships, and cultivate personal and collective wellbeing. Ethical approval for this study was granted by the University of Glasgow's Research Ethics Committee (Permission Numbers: 300200232, 300190109).

3.2.2.2 Signal Processing

Radar preprocessing 3.1.3 are adopted from the Section 3.1

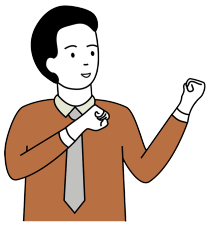
3.2.2.3 DL Models for Classification

In our classification task aimed at distinguishing activities, we utilise four specific pre-trained DL models: GoogLeNet, SqueezeNet, VGG16, and VGG19. These advanced convolutional neural network architectures, initially trained on ImageNet [121] , are adapted to analyse spectrogram images derived from radar data. During the fine-tuning process of these pre-trained models, adjustments are made to the top layers to enable the classification of collected data into sixteen predefined classes. Detailed descriptions of the CNN architectures used in this study are provided in subsequent subsections.

GoogLeNet: GoogLeNet, a prominent convolutional neural network (CNN) architecture renowned for image classification tasks, is composed of 22 layers, including convolutional, pooling, inception, and fully connected layers. The inception module, a key component, consists of six convolutional layers and a pooling layer. This module employs filters of different sizes (1×1, 3×3, and 5×5) to extract diverse patterns from input images. The feature maps produced by these filters are then concatenated at the module's output. Notably, 1×1 convolutions precede convolutions with larger filter sizes, contributing to parameter reduction and network efficiency. The specific parameter settings of GoogLeNet are detailed in Table 3.2.

SqueezeNet: The SqueezeNet architecture, another pre-trained model, comprises a total of 18 layers, as described by [121]. This architecture has gained popularity due to its comparable

(a) Ashamed



(b) Cheerful

(c) Enormal

(d) Furious

(e) Good Idea

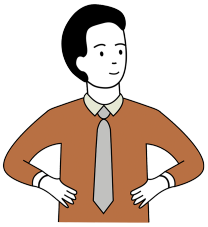
(f) Guilty

(g) Lonely

(h) Normal

(i) Ok

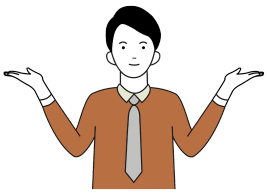
(j) Playful



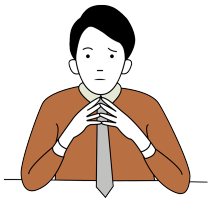
(k) Proud

(l) Sad

(m) Shocked



(n) Surprised



(o) Thinking

(p) Worried

Figure 3.7: A visual representation of 16 expressions

performance with significantly fewer parameters—roughly fifty times fewer—making it advantageous for resource-constrained applications. SqueezeNet employs three primary strategies. Firstly, it reduces the squeeze layer from 3x3 filters to 1x1 filters. Secondly, it utilises an expanded layer where 1x1 and 3x3 filters receive reduced input parameters. Lastly, it employs down-sampling later in the architecture with smaller stride values, resulting in larger activation maps in the final layer, which contributes to improved accuracy. Details regarding the parameter configurations for SqueezeNet are available in Table 3.2.

VGG16: VGG16 is comprised of 16 layers, as outlined by [121], boasting a total of 138 million parameters. The architecture primarily employs 3×3 filters with a stride of 1, while padding and max-pooling layers use 2×2 filters with a stride of 2. The layer arrangement follows a sequence of ReLU, convolutional, and max-pool layers, where ReLU layers contribute to computational efficiency and accelerated learning. Towards the end of the architecture, three fully connected layers and a softmax layer are incorporated for output generation. The specific parameter settings for VGG16 can be found in Table 3.2. With its well-structured layer design and parameter configuration, VGG16 serves as a robust framework for feature extraction and classification across diverse applications.

VGG19: The data underwent transformation through a distinctive layer featuring 3×3 filters across five stages of convolutional layers. Each stage was accompanied by pooling layers, followed by three fully connected layers, crucial for extracting essential image information. To bolster the extraction of image feature vectors, the convolutional kernel depth was progressively increased from 64 to 512, enhancing the capabilities of the VGG16 network. At each convolutional layer stage, pooling layers were applied, each with a size and stride of 2×2, as described by [120, 121].

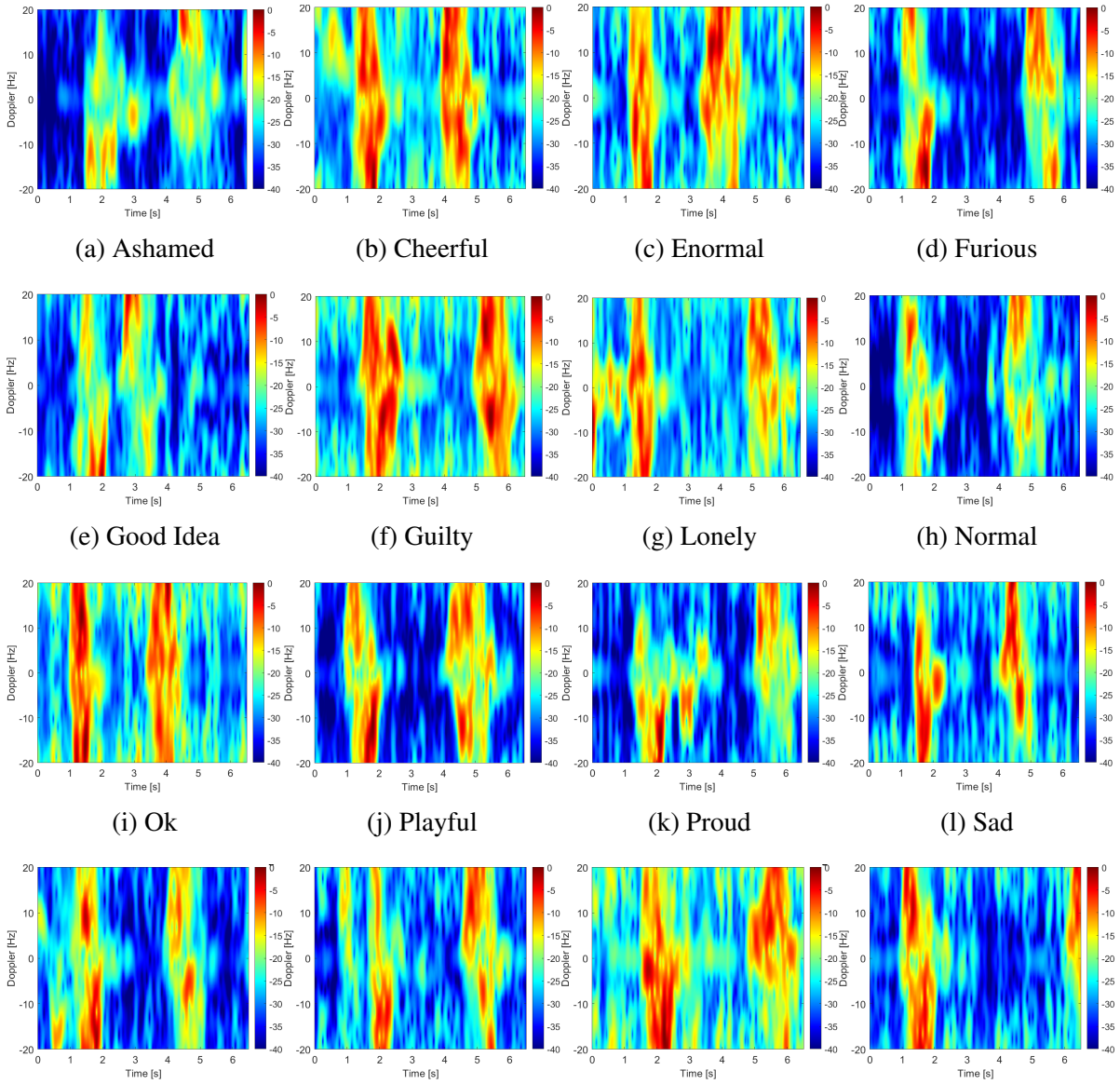
3.2.3 Research Experiments and Performance Evaluation

The description of dataset is discussed in this section along with the evaluation of system using pre-trained DL models.

3.2.3.1 Dataset

Following the data collection and signal processing steps, we obtained a set of spectrograms. The dataset has 1440 samples from 16 classes (ashamed, cheerful, enormal, furious, goodidea, guilty, lonely, normal, ok, playful, proud, sad, shocked, surprised, thinking and worried) that are collected from 3 users and each class contains 30 samples. Fig 3.8 shows the samples of spectrograms we obtained in the result of signal processing from the dataset which is then divided into two subsets, i.e., training and testing. The train set contains 1200 samples and the

test set has 240 samples. Further, these sets have equal representation of all the classes and subjects [121].



(a) Ashamed (b) Cheerful (c) Enormal (d) Furious

(e) Good Idea (f) Guilty (g) Lonely (h) Normal

(i) Ok (j) Playful (k) Proud (l) Sad

(m) Shocked (n) Surprised (o) Thinking (p) Worried

Figure 3.8: Obtained Spectrograms Sample of 16 Classes

3.2.3.2 Performance Evaluation

evaluation metrics 3.1.3.3 are adopted from the Section 3.1

3.3 Multimodal Sensing for Phrase Recognition in BSL

3.3.1 Introduction

This chapter presents a MM sign language recognition system that combines radar and video sensing to improve accuracy and reliability. It builds on the findings of the previous works, which showed how radar and video can each recognise sign language but also have individual limitations. In this study, both sensing methods are merged using DL models to take advantage of their strengths. The radar captures motion information from micro-Doppler signals, while the video provides visual features such as hand and body movements. These features are combined using a hybrid CNN–LSTM and ResNet-based model to handle challenges like lighting changes, occlusion, and differences in signing styles. In this work, we present a radar and vision-based framework capable of recognising 20 phrases, helping to bridge communication gaps for Deaf or hard of hearing. The key contributions of this study are as follows:

- Privacy-preserving, contactless sensing: We propose a non-intrusive radar-based recognition system capable of identifying and translating 20 phrases. The use of contactless sensing ensures privacy protection.
- Unimodal and MM evaluation: We conduct systematic analyses of radar-only, video-only, and MM fusion models to understand their individual and combined contributions. This establishes a proof-of-concept for integrating radar–vision systems into future assistive devices, including next-generation hearing aids.
- Deep learning–based classification: A ResNet-18 deep convolutional neural network is employed to evaluate three configurations: radar-only, video-only, and MM fusion, providing a systematic comparison of unimodal and MM performance.
- Foundation for scalable assistive technologies: We demonstrate the feasibility of a non-intrusive, privacy-preserving, and scalable sign language recognition framework with potential applications in accessibility technologies for people with hearing impairments.

3.3.2 Experimental Methodology and Signal Processing

The experimental setup is shown in Figure 3.9, where the UWB radar module is mounted above a laptop to allow synchronized data collection and processing. Data were recorded in a normal indoor environment under typical ambient conditions. The setup was carefully arranged to reduce background noise, reflections from nearby objects and walls, lighting variations, and other possible interferences. This ensured that the dataset collected was reliable and suitable for later analysis. During each session, participants sat approximately 141 cm in front of the radar module. The X4M03 includes one transmitter and one receiver antenna, and it operates in the 7.29

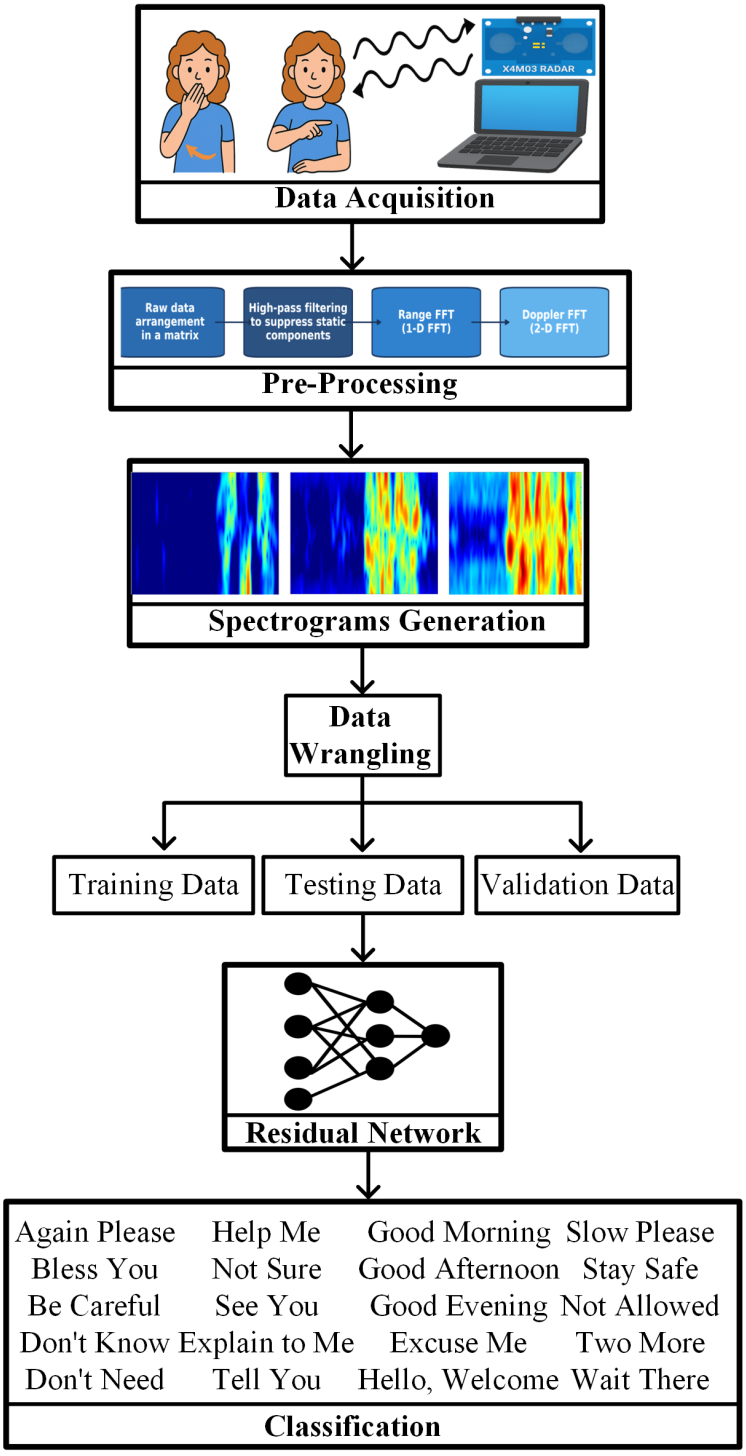


Figure 3.9: Framework of the Proposed Work

GHz–10.2 GHz frequency range (X-band). The system has a sampling rate of 23.328 GS/s and a maximum detection range of 9.6 m. The radar transmits signals with a power level of 6.3 dBm, and its ERP follows IEEE UWB standards, ensuring low interference with other systems and safe operation. At a sensing distance of 141 c cm, the radar's power density is approximately 1.68 W/m², which is well below the IEEE C95.1-2005 exposure limits of 50 W/m² for controlled environments and 10 W/m² for uncontrolled environments [119].

3.3.2.1 Radar System

Figure 3.10 illustrates the working mechanism of the XeThru X4M03 UWB radar module. The system begins with the UWB Radar (X4 Chip), which transmits and receives ultra-wideband pulses to detect motion within its sensing field. The reflected radar signals are processed by the Signal Controller/Buffer (ARM / FPGA / MCU), which manages radar timing, frame buffering, and data communication with the host system. A Power Regulator provides a stable 3.0 V output from a 4.5–5.5 V supply, while the 12 MHz Oscillator generates the precise reference clock required for reliable radar operation. The buffered radar data are transferred through the Data Interface (USB / UART / SPI) to an external Host Processing Unit for further analysis. Unlike systems with onboard signal processing, the X4M03 transmits raw radar frames to the host, allowing complete control over the data processing pipeline. On the host side, the received data undergo several sequential processing stages: preprocessing, feature extraction, and classification. In this work, the classification stage employs a ResNet-18 deep CNN to automatically learn spatial and temporal features from radar spectrograms and accurately identify sign language gestures. The Output of the system corresponds to the recognised gesture class.

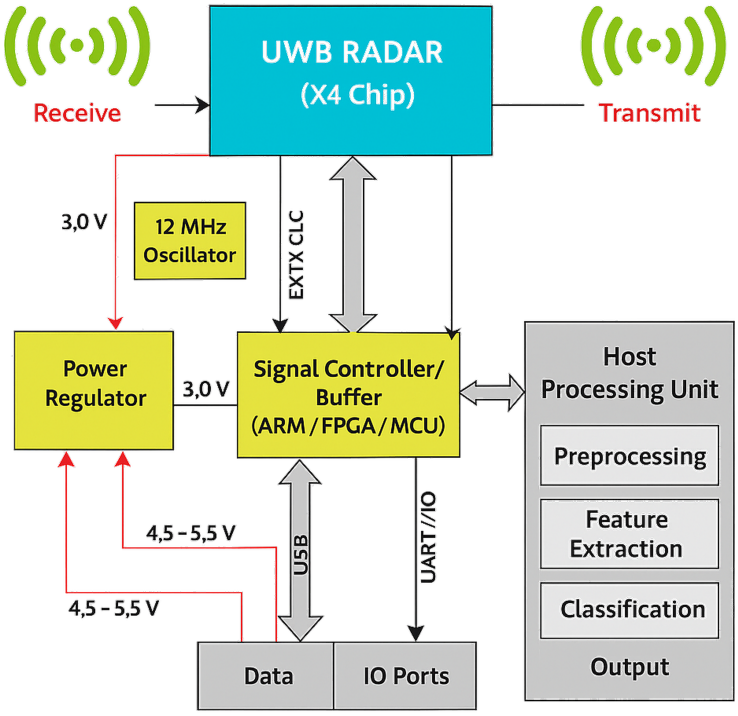


Figure 3.10: Block Diagram of UWB X4MO3 Radar System

A total of 3,600 radar samples were collected from six participants across 20 classes, with each class repeated 30 times. The use of six participants captures inter-subject variability, the 20 classes consist of commonly used daily phrases, and the total of 3600 samples (6×20×30) ensures a balanced dataset with sufficient repetitions for robust model training and evaluation.

The preprocessing steps applied to the radar data are the same as those described in Section 3.1.3. Each sign after signal processing exhibited a unique Doppler signature, as illustrated in Figure. 3.4.

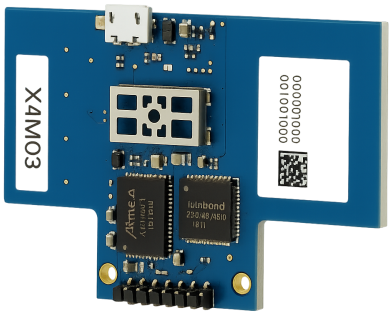


Figure 3.11: Xethru X4MO3 UWB Radar Sensor

3.3.2.2 Camera System

The signal processing steps for the data collected using the camera are adopted from Section 3.1.3.2.

Table 3.3: Parameter Settings of the Radar Sensor.

Parameter	Value
Activity duration	6 seconds
Radar Sensor	XeThru X4M03
Instrumental range	9.6 metres
Radar's distance from subject	141 cm
Radar frequency	7.29 GHz
Radar bandwidth	1.5 GHz
Tx power	6.3 dBm

3.3.3 ResNet for Sign Classification

CNNs are widely recognised as state-of-the-art for image-based classification due to their ability to extract hierarchical spatial features from complex input data. A major milestone in CNN development was the introduction of the ResNet [122], which addresses difficulties in training very deep neural networks, particularly vanishing and exploding gradients. The key idea behind ResNet is the use of residual blocks, where the network learns residual mappings instead of direct input–output transformations. Formally, rather than approximating $H(x)$ directly, each block learns $F(x) = H(x) - x$, producing the output $H(x) = F(x) + x$. These identity (skip) connections facilitate effective gradient flow during backpropagation, enabling the stable training of deeper architectures.

A typical residual block consists of convolutional layers followed by batch normalization and ReLU activation, enhanced by identity shortcuts. In deeper variants such as ResNet50 and ResNet101, a bottleneck structure is employed, using a 1×1 convolution for dimensionality reduction, a 3×3 convolution for feature extraction, and a final 1×1 convolution to restore dimensionality, balancing representational power and computational efficiency. In this work, ResNet18 was employed as the backbone model due to its computational efficiency and strong capability in extracting spatial features. It was used across three experimental settings: radar-only, video-only, and MM fusion. For radar spectrograms, which contain structured Doppler-based motion signatures, ResNet18 provided sufficient representational power while minimizing overfitting, achieving a classification accuracy of 96%. For video data, grayscale frames were used as inputs, and although ResNet18 effectively captured spatial information, the absence of temporal modeling introduced sensitivity to factors such as lighting and background variation, resulting in an 90% accuracy. Finally, the MM configuration, combining radar and video representations, leveraged complementary motion and visual cues, yielding improved overall

performance and demonstrating the advantage of sensor fusion for human-activity recognition. The residual learning mechanism contributed to stable convergence across all modalities, confirming ResNet18 as an effective and efficient architecture for both single-modality and MM classification tasks.

3.3.3.1 Evaluation Metrics

The evaluation metrics have already been described in Subsection 3.1.3.3.

3.4 Summary

This chapter has presented the methodology used in this research for recognising non-verbal communication using radar, video, and MM sensing. It described the system design, data collection process, preprocessing steps, feature representation, and the deep learning models used for classification. Different sensing modalities were considered to capture hand gestures, head movements, facial expressions, and sign language at both gesture and phrase levels. The next Chapter 4 presents the experimental results and performance evaluation, where the effectiveness of the proposed methods is analysed and compared in detail.

Chapter 4

Results and Discussion

4.1 Performance Comparison of Video and Radar for Common Signs Recognition in BSL and ASL

4.1.1 Results and Discussion

The performance metrics for training, validation and testing are summarized in Table 4.1. For the training and validation phases, loss and accuracy were the primary evaluation metrics. During testing, accuracy was used as the main performance indicator due to its simplicity and relevance in balanced, unbiased datasets. To provide a more comprehensive evaluation of the model's effectiveness, additional metrics such as precision, sensitivity, specificity, and F1 score were also included. The results across all modality configurations demonstrate strong model performance across all metrics.

Table 4.1: Performance Metrics for Different Modalities

Metric	**Radar Only**	**Video Only**
Training Loss	0.005	0.10
Training Accuracy	99.9%	86%
Validation Loss	0.02	0.18
Validation Accuracy	91%	78%
Testing Precision	96.54%	82.29%
Testing Sensitivity	96.13%	81.77%
Testing Specificity	96.52%	82.19%
Testing F1 Score	96.14%	80.18%
Testing Accuracy	96.13%	81.77%

To gain a comprehensive understanding of the MM model's performance and behavior, it is essential to first evaluate each modality independently. We proposed a proof of concept for MM sensing framework. In this work, we analyzed the modalities individually (unimodal analysis) to see their independent contributions. Two separate modeling approaches were carried out: radar

only and video only.

4.1.1.1 Radar Only

Training loss and accuracy for radar only over 150 epochs is graphically represented in Figure **??** (a). The training and validation accuracy begin to diverge around epoch 60, with final training accuracy reaching approximately 99.9% and validation accuracy stabilizing near 91%. Similarly, the training and validation losses reduce steadily, with training loss dropping to around 0.005 and validation loss settling close to 0.02. These trends indicate successful learning without overfitting. The consistently low loss values and high accuracies suggest that the radar-only model generalizes well despite the moderate complexity of the data. The slower convergence early in training could be attributed to the complexity of patterns in the Short Time Fourier Transform (STFT) images, which might benefit from deeper or more specialized architectures.

Figure 4.1 presents the corresponding radar-only confusion matrix (normalized percentages across 15 classes). The matrix reveals that the model predicted most classes with high confidence, with 7 classes achieving 100% true positive rates and several others above 90%. Some misclassifications are observed, notably in classes 6, 7, and 12, where samples were occasionally confused with neighboring classes. These errors likely stem from similarities in the radar-based motion signatures of certain signs, particularly those with overlapping or subtly distinct movements.

4.1.1.2 Video Only

Figure 4.2 shows that the video-only model required approximately 50 epochs to converge. Both training and validation metrics loss and accuracy progressed steadily throughout, with training loss decreasing to around 0.10 and validation loss to approximately 0.18. Training accuracy reached 86%, while validation accuracy peaked at around 78%, indicating that it successfully learned from the video data without significant overfitting, although some fluctuations in validation performance were observed. These variations may be attributed to noise or inconsistencies in the visual input, such as signer movement or lighting conditions.

Figure **??** (b) presents the video-only confusion matrix, highlighting mixed class-wise performance. While some classes achieved strong True Positive (TP) rates such as class 8 and class 9 with 100%, others like class 13 (67%) and class 14 (58%) showed considerable misclassifications. One class (class 2) recorded only 67% TPs, while several others fell between 70–92%. A pattern of confusion clustering is again visible, similar to what was observed in the radar matrix. Notably, class 4 in both modalities showed comparable lower TP rates, and certain misclassifications were concentrated around specific neighboring classes. These similarities may indicate shared difficulty across sensor types in distinguishing certain signs, potentially due to subtle variations in gesture execution or visually similar hand movements—suggesting that some recognition challenges are inherent across modalities.

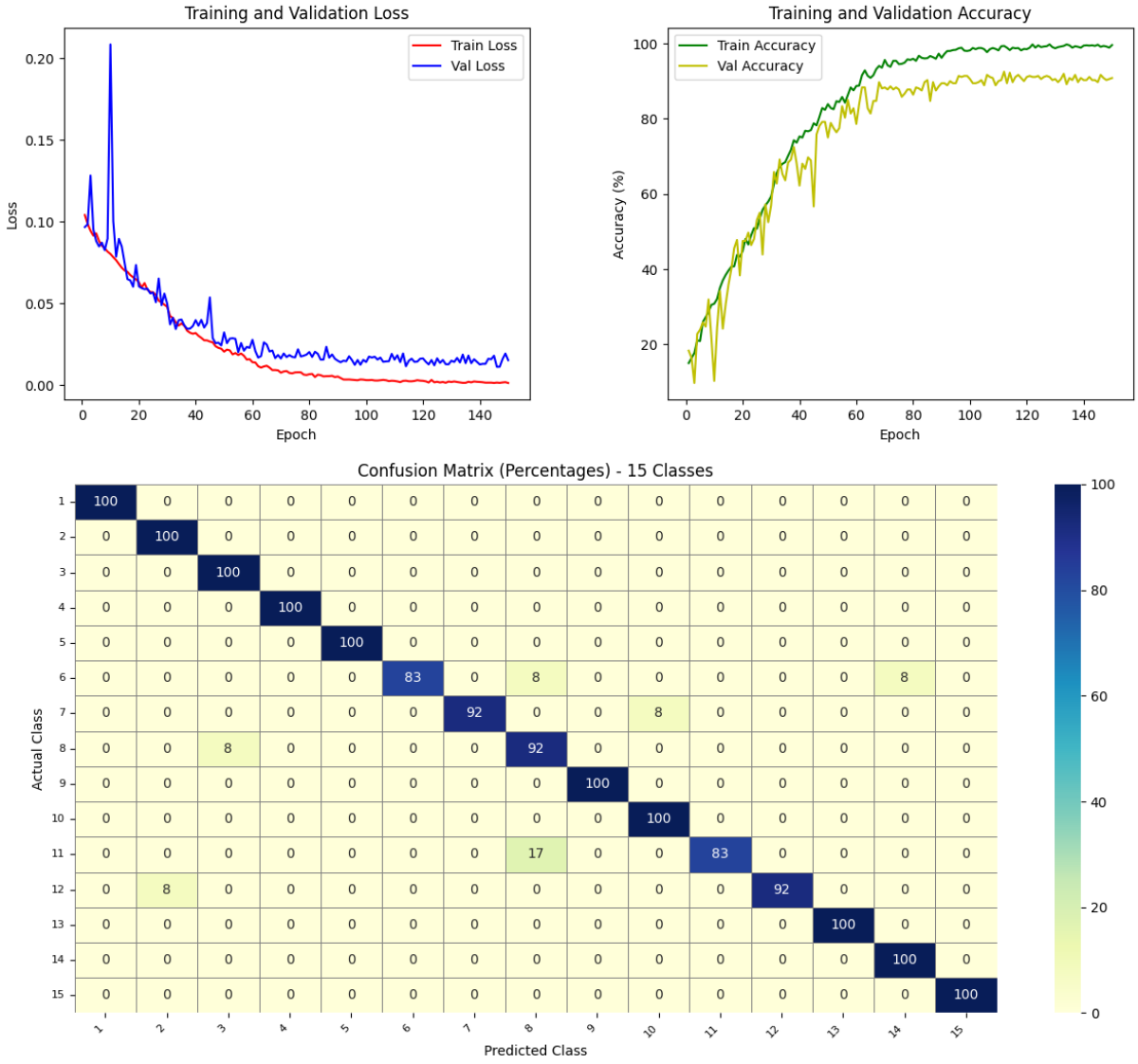


Figure 4.1: Radar Only Training and Accuracy Plots and Confusion Matrix

4.2 Utilising Contactless Sensing Technology for the Identification of Hand and Head Movements in Conjunction with Facial Expressions

4.2.1 Results and Discussion

Data is gathered using the UWB radar positioned 141 cm from the subject, then transformed into spectrograms through signal processing. These spectrograms are subsequently inputted into pre-trained DL models, and their performance is assessed on the acquired data. The parameter settings for 6 DL models are given in Table 3.2. In all experiments, train and test sets are fixed, i.e., 80/20 of the total data, respectively.

Table 4.2 provides a comprehensive view of the experimental outcomes from tests conducted

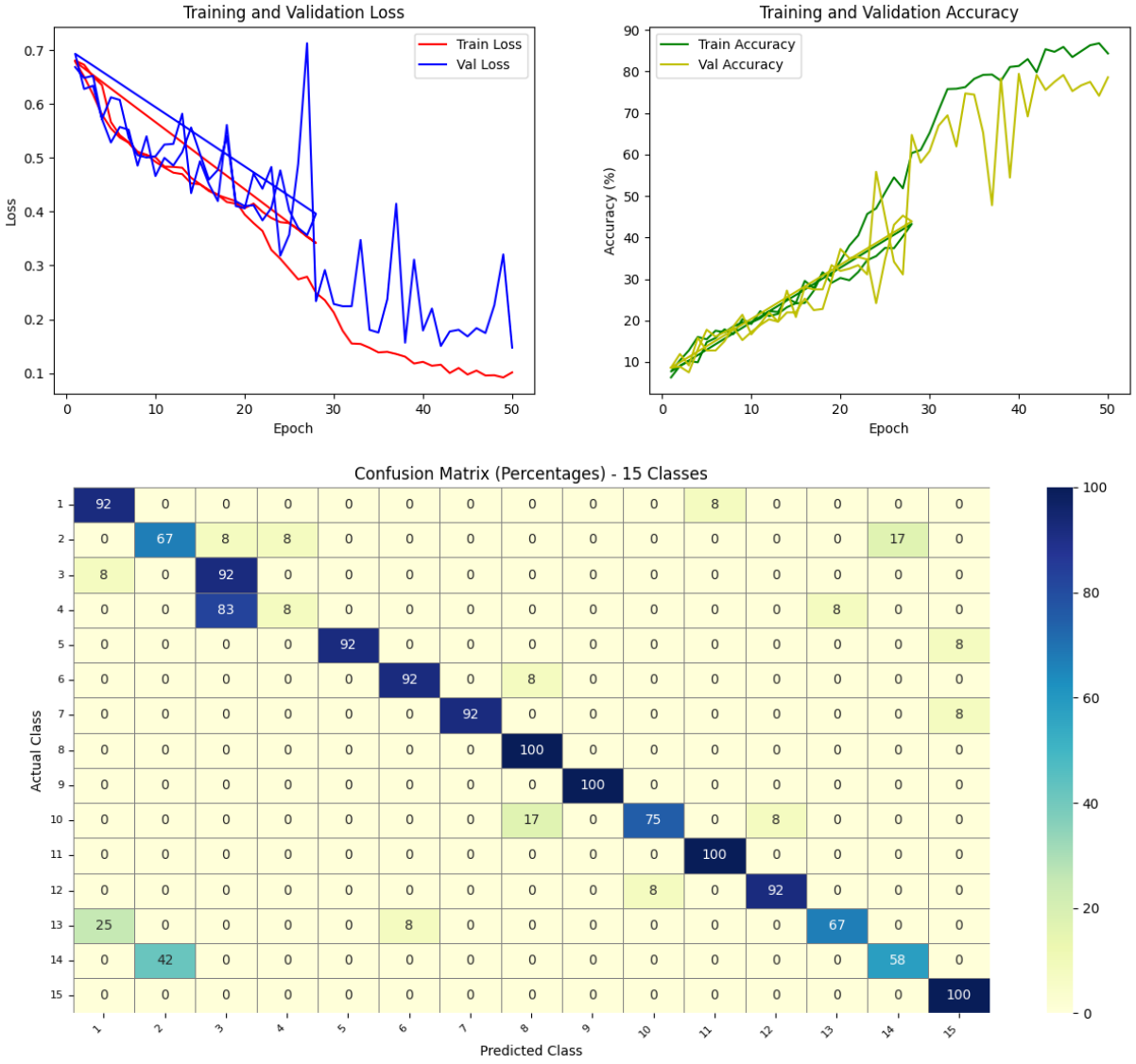


Figure 4.2: Video Only Training and Accuracy Plots and Confusion Matrix

at a distance of 141 cm, presenting detailed metrics, i.e., precision, recall, F1 score, and accuracy. All models performed well as microDoppler signature at 141 cm are more sensitive to hand and head movements [121]. Further, four pre-trained DL models, i.e., GoogleNet, SqueezeNet, VGG16, and VGG are applied for classification purposes and their confusion matrices are shown in Fig 4.3 and 4.4 which are explained below:

GoogleNet The confusion matrix in Fig 4.3a reveals that GoogleNet correctly classifies the majority of the expressions with near-perfect accuracy. However, the Cheerful class shows an accuracy drop to 60%. This misclassification is likely due to its resemblance to Furious, Playful, and Proud, as these expressions may share similar micro-Doppler patterns, leading to confusion in classification.

Table 4.2: Evaluation of pre-trained models

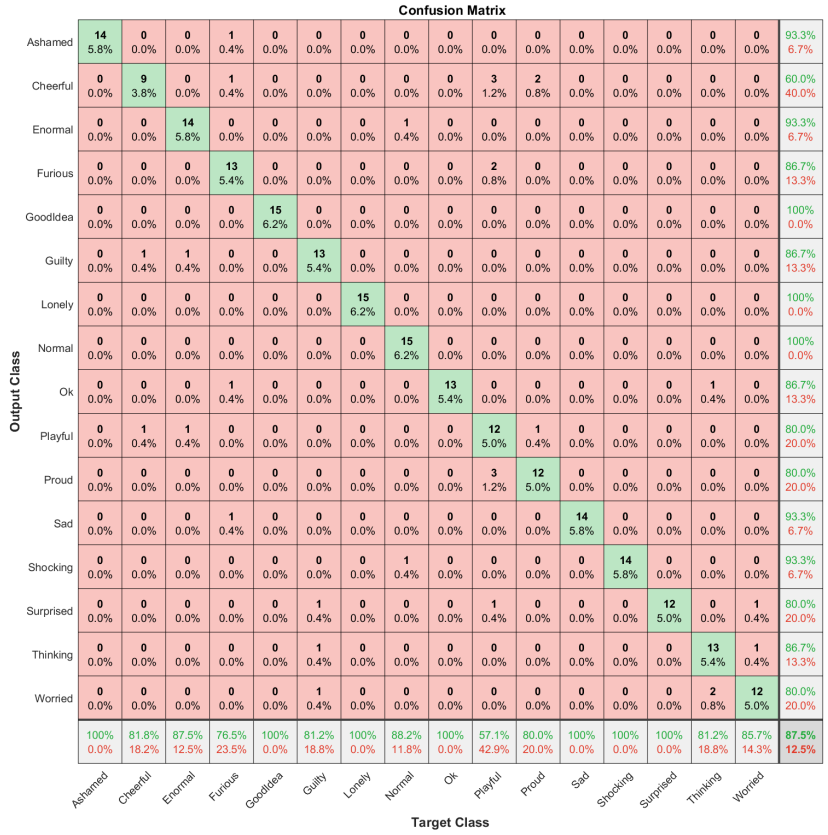
Model	Precision	Recall	F1 Score	Accuracy (%)	95% CI
GoogleNet	0.88	0.89	0.88	87.5	85.8–89.2
SqueezeNet	0.86	0.87	0.86	85.8	84.0–87.6
VGG16	0.90	0.94	0.92	94.2	92.5–96.0
VGG19	0.93	0.94	0.93	93.3	92.0–94.3

SqueezeNet Most of the activities are recognised with almost 100% accuracy except two classes, i.e., Cheerful and Surprised as shown in Fig 4.3b. Class Cheerful is 0.8% confused with class Playful 0.4% with Furious, Proud, and Surprised. Whereas the surprised class has a 0.8% resemblance with Playful and 0.4% with Sad, Shocking, and Worried. Overall, the lowest accuracy is 66.7% for classes Cheerful and Surprised.

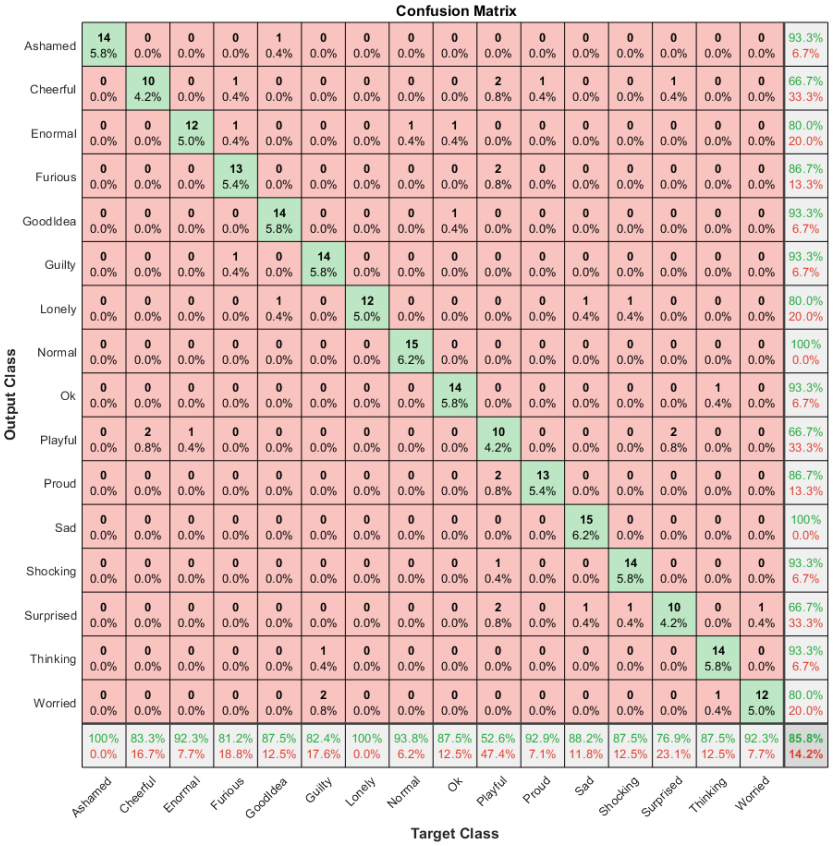
VGG16 The confusion matrix in Fig 4.4a demonstrates that VGG16 outperforms the other models, achieving an overall accuracy of 94.2%. However, five expression classes (Cheerful, Playful, Surprised, Thinking, and Worried) have relatively lower accuracy, with a minimum recorded accuracy of 86.7

- Cheerful is misclassified with GoodIdea and Playful (0.4% confusion).
- Playful shows a 0.4% resemblance with Ashamed and Ok.
- Surprised is confused 0.4% of the time with Guilty and Playful.
- Thinking has a 0.4% similarity with Furious and Sad.
- Worried shares a 0.8% resemblance with Thinking.

The relatively high performance of VGG16 suggests that its deeper layers effectively capture spatial and temporal features from the spectrograms. However, the misclassifications indicate that some expressions exhibit highly overlapping motion cues, making them challenging to distinguish.

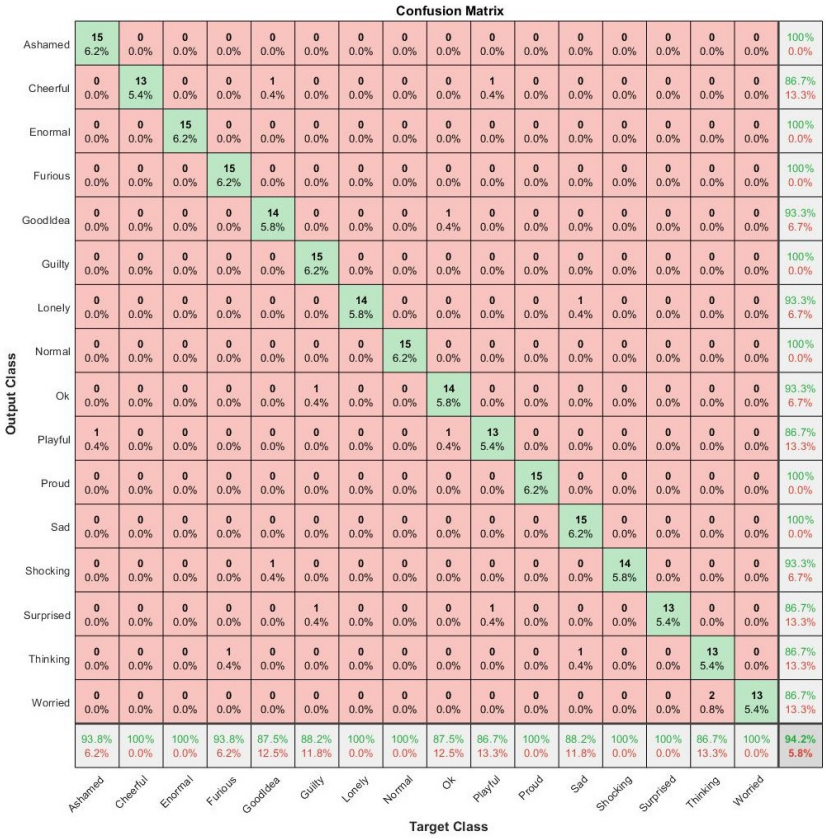


(a) Confusion matrix of GoogleNet

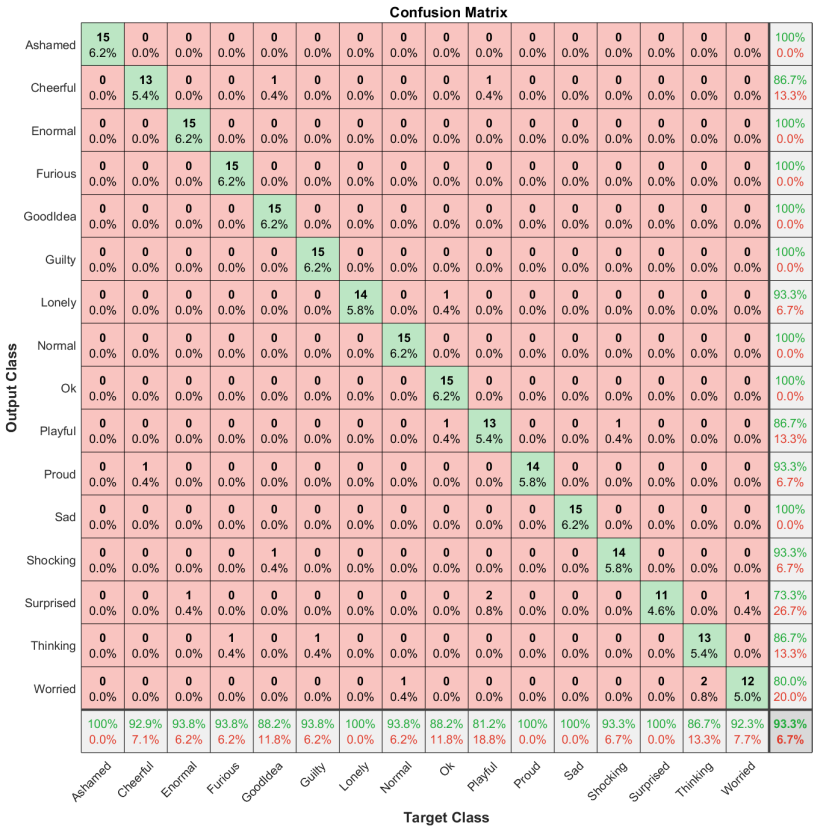


(b) Confusion matrix of SqueezeNet

Figure 4.3: Confusion Matrices of GoogleNet and SqueezeNet



(a) Confusion matrix of VGG16



(b) Confusion matrix of VGG19

Figure 4.4: Confusion Matrices of VGG16 and VGG19

VGG19 As depicted in the confusion matrix (Fig 4.4b), VGG19 also achieves near-perfect classification for most expression classes. However, Surprised is the only class with the lowest accuracy at 73.3%, as it shares 0.8% similarity with Playful and 0.4% with Enormal. The reduced accuracy for the Surprised class could be due to its motion signature being similar to other dynamic expressions such as Playful. The deeper architecture of VGG19 provides strong feature extraction capabilities, but the minor performance degradation suggests that subtle expression variations still pose a challenge.

4.3 Multimodal Sensing for Phrase Recognition in BSL

4.3.1 Results and Discussion

The performance results for the training, validation, and testing stages are shown in Table 4.3. During training and validation, loss and accuracy were used to monitor how well the model was learning over time. For testing, accuracy was chosen as the main performance measure because it is easy to interpret and appropriate for balanced datasets. To give a more complete picture of how the model performs, additional metrics were also included, such as precision, sensitivity (recall), specificity, and F1-score. These extra metrics help explain how well the model correctly identifies gestures, avoids false detections, and balances missed and incorrect predictions.

Table 4.3: Performance Metrics for Different Modalities.

Metric	**Radar Only**	**Video Only**	**MM**
Training Loss	0.10	0.09	0.50
Training Accuracy	99%	93.6%	96.8%
Validation Loss	0.15	0.18	0.80
Validation Accuracy	95.4%	89.7%	93.5
Testing Precision	95.8%	90.8%	93.9
Testing Sensitivity	95.4%	89.7%	93.5
Testing Specificity	95.8%	90.8%	93.9
Testing F1	95.4%	89.7%	93.6
Testing Accuracy	95.4%	89.7%	93.5

To clearly understand the effectiveness of the MM model, it is important to first examine the performance of each individual data source. For this reason, our framework begins with unimodal evaluation, where radar and video data are analysed separately to determine their independent strengths and limitations. This step helps identify how much each modality contributes to the recognition task before combining them. After completing the unimodal assessments, we then evaluate the MM approach, which integrates both radar and video information to improve overall performance. In total, three modelling configurations were explored in this study: radar-only, video-only, and MM fusion.

4.3.1.1 Radar Only

Training loss and accuracy for the radar-only model over 100 epochs are illustrated in Figure 4.5. The training and validation accuracy curves rise steadily throughout training, with a noticeable divergence beginning around epoch 60. By the final epoch, the model achieves approximately 97% training accuracy, while the validation accuracy stabilises near 90%. Similarly, both training and validation loss values decrease consistently, with training loss reaching approximately 0.15 and validation loss converging around 0.35. These learning curves indicate effective optimisation and good generalisation, with no major signs of overfitting. The smooth and gradual reduction in loss, paired with consistently increasing accuracy, suggests that the radar-only model successfully learns discriminative motion features associated with sign-language phrases.

Figure 4.5 shows the corresponding radar-only confusion matrix, normalized across 20 sign-language classes. The matrix demonstrates strong classification performance, with most classes achieving high true-positive rates. Several classes reach 100% accuracy, while many others exceed 89–94%, confirming reliable recognition across diverse gestures. A few classes, such as Class 7 and Class 15, achieve approximately 83–89% accuracy, indicating occasional misclassification. These errors are likely caused by similarities in the radar-based motion signatures of certain sign-language phrases, particularly those with overlapping hand trajectories or subtle variations in movement speed and direction. Despite these minor confusions, the dominant diagonal pattern in the matrix confirms that the radar-only model performs robustly and is capable of accurately distinguishing complex phrase-level sign-language gestures using Doppler-based motion information alone.

4.3.1.2 Video Only

Figure 4.6 shows that the video-only model converged within approximately 25 epochs, with both training and validation curves demonstrating stable learning behavior. Training loss decreased to around 0.10, while validation loss reached approximately 0.18. Training accuracy climbed to about 86%, and validation accuracy peaked near 90%, indicating that the model successfully learned meaningful visual features without overfitting. Although small fluctuations were observed in validation accuracy, these are expected in video-based sign recognition due to factors such as lighting sensitivity, changes in background, camera angle differences, motion blur during fast gestures, and natural variations in signer movement and hand orientation.

Figure 4.6 also presents the confusion matrix, revealing strong class-wise performance for most signs, with several achieving 100% TP recognition. However, certain classes showed lower accuracy, such as Class 18 with 61% and Class 20 with 72%, demonstrating noticeable confusion with visually similar gestures. These misclassifications often occurred between signs with overlapping hand shapes or subtle finger movements, which can be difficult for an RGB camera to capture precisely, especially under poor lighting, rapid hand motion that introduces blur, partial occlusion, or varying signer distance from the camera. Additionally, some gestures

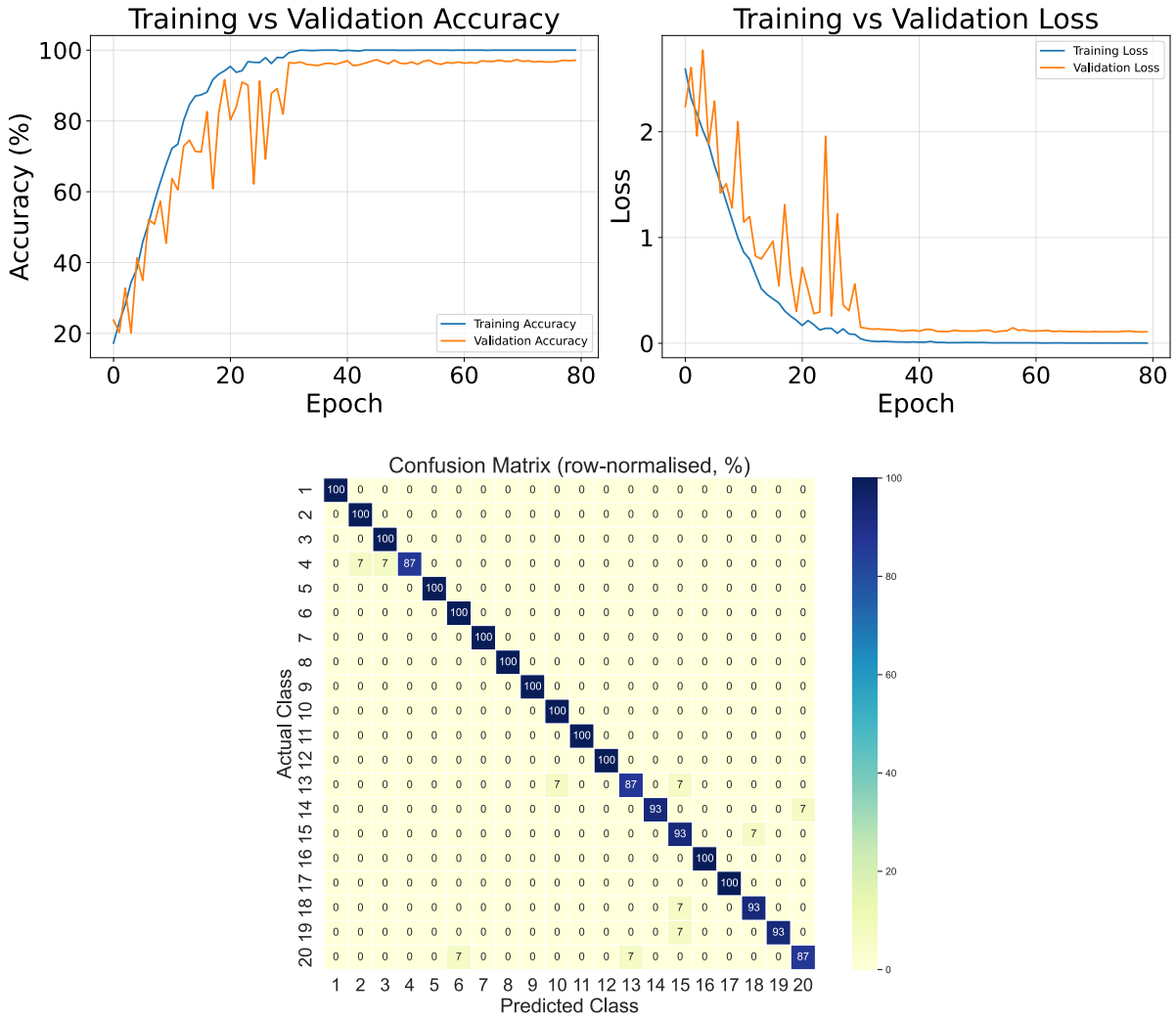


Figure 4.5: (a) Training and validation loss and accuracy and the confusion matrix for the radar test datasets

involve slight depth or angle variations that are challenging to distinguish using 2D video alone.

4.3.1.3 Multi-Modal

Figure 4.7 presents the training and validation accuracy and loss of the MM model over the training epochs. Both training and validation accuracy increase steadily, showing that the model learns effectively as training progresses. By the final epochs, training accuracy reaches approximately 97%, while validation accuracy reaches about 92–94%. The small difference between these values indicates good generalization and no significant overfitting. At the same time, both training and validation loss decrease smoothly from around 3.0 to approximately 0.5 and 0.8, respectively. Minor fluctuations in the validation curves are expected and are mainly caused by natural variations in MM data. Overall, these results confirm that the model converges stably and learns meaningful features from the combined modalities.

Figure 4.7 shows the confusion matrix, which provides a detailed evaluation of class wise

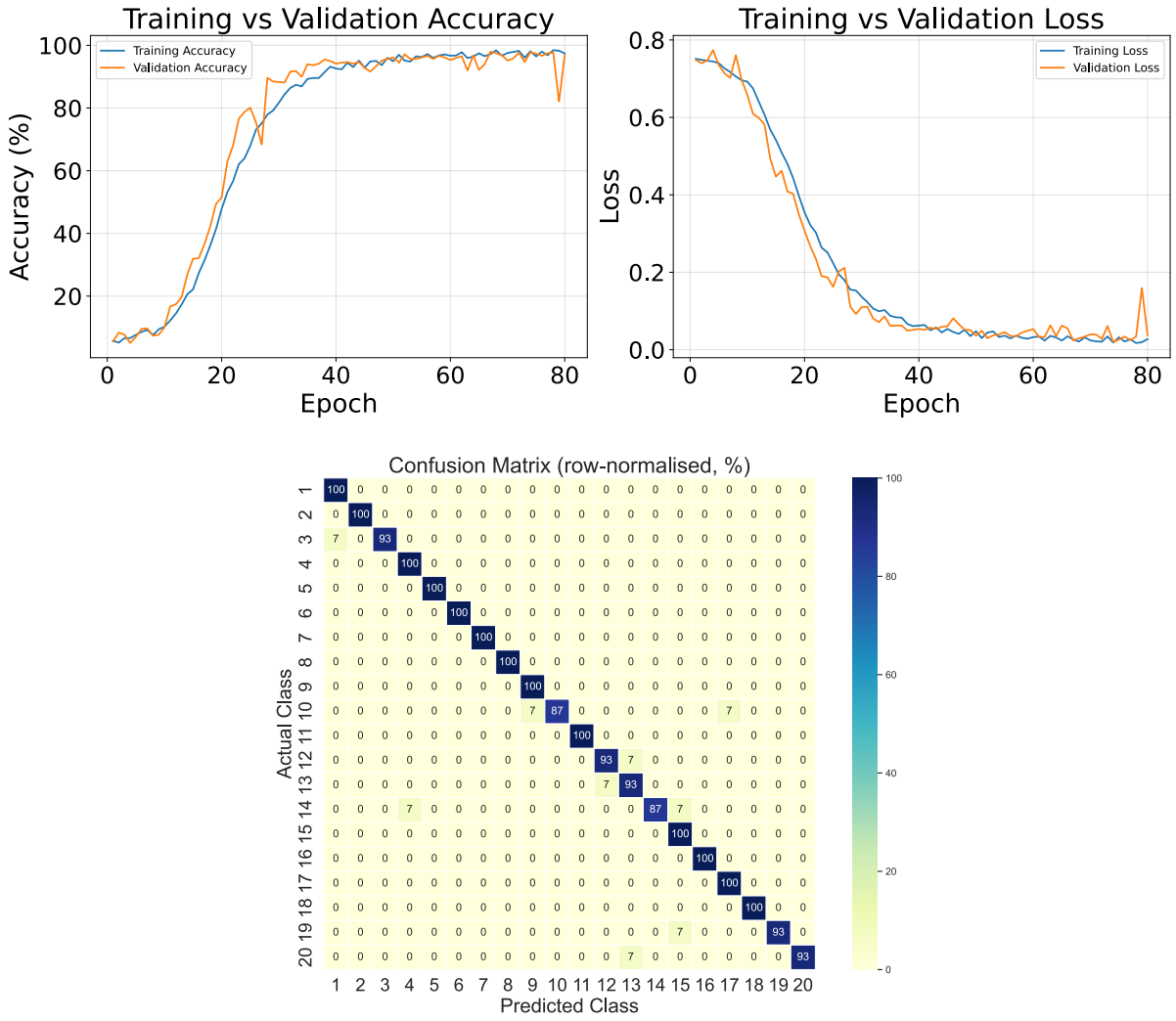


Figure 4.6: Training and validation loss and accuracy and the confusion matrix for the video test datasets

performance. Most classes achieve high recognition rates, typically between 93% and 100%, indicating strong and consistent classification performance. The strong diagonal pattern confirms that the model correctly predicts the majority of samples in each class. A small number of classes, such as class 10 and class 14, show slightly lower accuracy (around 87%), with errors mainly occurring between visually similar classes. These misclassifications are limited and expected due to similarities in gesture shape or movement. Overall, the confusion matrix demonstrates that the MM model performs reliably across all classes and achieves balanced and robust recognition performance. The MM model achieves better performance by combining radar and video data. Radar provides reliable motion and depth information that is robust to lighting and background changes, while video captures detailed visual features such as hand shape and orientation. Together, these complementary modalities allow the model to learn more discriminative features. As a result, the MM model shows more stable training, higher accuracy, lower loss, and clearer class separation.

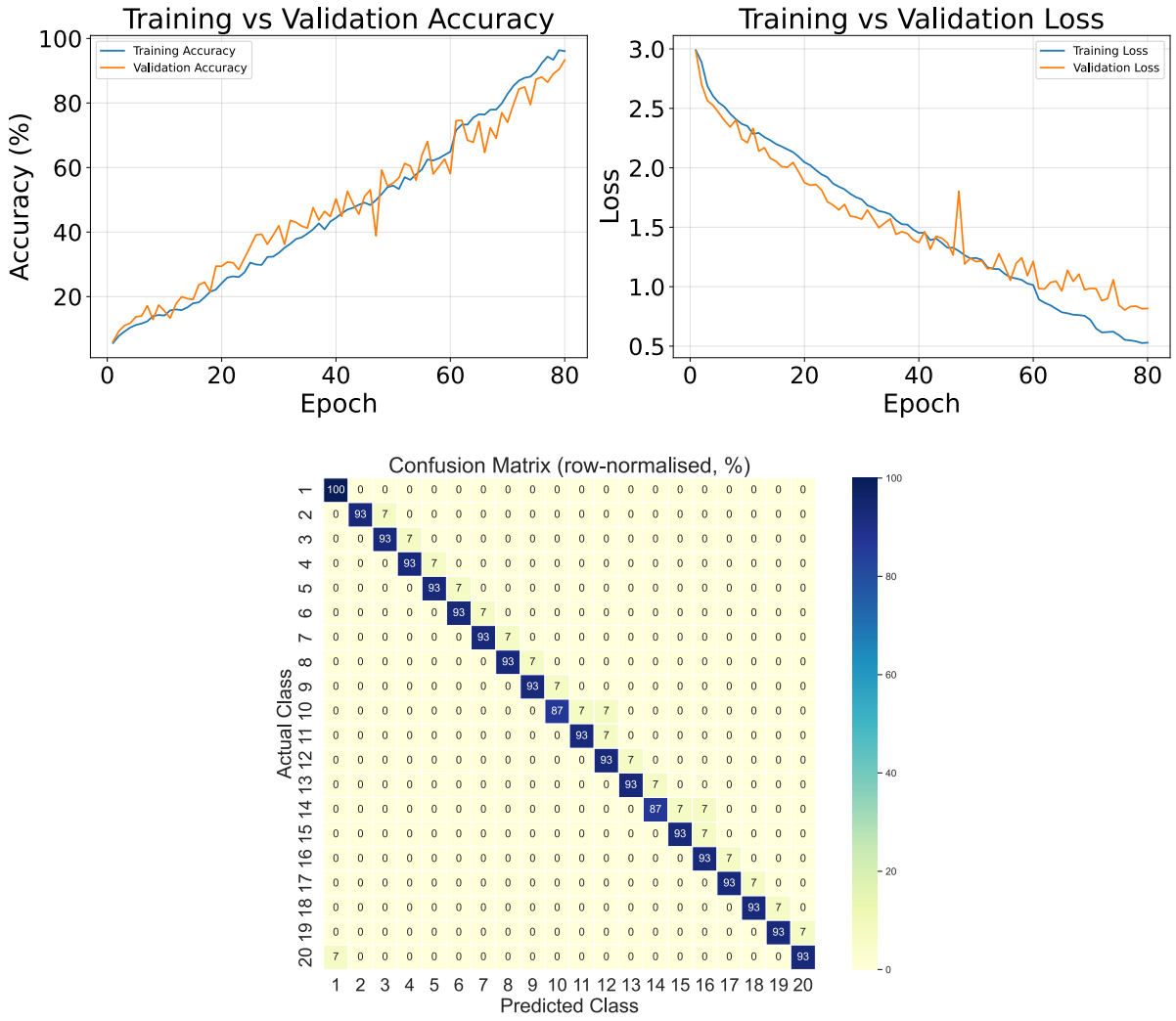


Figure 4.7: Training and validation loss and accuracy and the confusion matrix for the MM dataset

4.4 Summary

This chapter presents the experimental results and performance analysis of the proposed radar-only, video-only, and MM recognition systems. The results are evaluated using accuracy, loss curves, and confusion matrices to assess classification performance across facial expressions, hand and head movements, common ASL/BSL signs, and phrase-level sign language recognition. A detailed comparison of the different approaches is provided to highlight their strengths and limitations. The findings discussed in this chapter form the basis for the conclusions and future research directions presented in the next chapter.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The main purpose of this thesis is to propose the concept of future MM hearing aid devices based on RF sensing capable of bridging communication between deaf communities and hearing individuals through the application of DL techniques. This research contributes to the development of contactless and privacy-preserving assistive systems designed to recognise non-verbal communication, i.e., hand gestures, facial expressions, head movements, and sign language phrases, by combining radar and camera-based sensing into a unified MM framework.

The thesis also provides an overview of existing technologies used to assist individuals with hearing impairments, focusing on both contact-based and contactless approaches. Contact-based methods, including wearable sensors, have shown high recognition accuracy but also present limitations such as discomfort, limited battery life, and the possibility of users forgetting to wear them consistently. These issues can reduce the long-term practicality and usability of such systems, particularly for individuals with cognitive challenges, who may struggle to maintain consistent use of wearable devices. Therefore, there is a growing demand for contactless methods that offer greater comfort, convenience, and reliability.

Contactless approaches; such as camera, radar technologies represent the next-generation of assistive systems. Camera-based systems are effective for recognising lip movements, sign language, head gestures, and facial expressions. However, their use raises privacy concerns, as continuous video recording may be perceived as intrusive or legally sensitive in certain regions.

RF sensing has effectively resolved several of the challenges faced by previous methods. In radar-based frameworks, Doppler shift spectrograms are used to represent subtle micro-movements, which are processed and classified through DL architectures. The extracted features are then interpreted using ML models to recognise detailed gestures and facial expressions.

To support the Deaf and hard-of-hearing community, this thesis presents a proof-of-concept MM monitoring system using RF signals to detect and interpret micro-movements associated with verbal and non-verbal communication. The research demonstrates how ML and DL models

can identify RF signal patterns that correspond to specific gestures and expressions. Multiple ML and DL architectures were tested and analysed for accuracy, efficiency, and adaptability, ensuring that the proposed system meets the practical requirements for real-world hearing aid applications.

The experiments were carried out in three main stages. The first stage demonstrated that radar sensing could successfully recognise facial expressions and movements of the head and hands. DL models such as VGG16 and GoogLeNet achieved high accuracy, confirming that radar can detect detailed human motions without visual input. The second stage compared radar-only and video-only systems for recognising common ASL and BSL gestures, showing that radar performs better in low-light and noisy conditions, while video provides strong visual information. The final stage combined both modalities for phrase-level recognition in BSL. The fusion of radar and video data significantly improved recognition accuracy, robustness, and performance across diverse environments.

This progression proves that RF signals can effectively detect and classify human micro-movements such as lip, hand, head, and facial gestures. The integration of ML and DL techniques ensures precise classification, establishing a solid foundation for the future development of MM hearing aid devices. The findings confirm that RF-based multimodal sensing can serve as a powerful communication bridge between Deaf and hearing individuals, offering a contactless, intelligent, and privacy-preserving solution.

In summary, this thesis highlights the potential of advanced MM assistive technology in improving both verbal and non-verbal communication within the Deaf community. By introducing an innovative RF-based sensing framework supported by ML and DL models, this work sets the foundation for next-generation hearing aids that are intelligent, adaptable, and user-friendly. The research demonstrates how RF-based MM hearing aids can outperform traditional wearable and vision-based technologies, offering new possibilities for inclusive communication. It marks an important step towards the development of assistive devices that enhance accessibility, independence, and quality of life for Deaf and hard-of-hearing individuals.

5.2 Limitation

There are some limitations in RF sensing-based technology, which are described in the following points.

- **Limited Amount of Data:** The accuracy of RF sensing systems depends on the amount and variety of data collected. When the number of participants or data samples is small, it can reduce how well the ML or DL models perform in new situations. Collecting a larger and more diverse dataset with different users, gestures, and environments will help improve the reliability of the system.

- **Regulatory Constraints:** The use of RF sensors is controlled by national and international regulations that limit the power of transmission and the frequency bands that can be used. These restrictions can sometimes limit how the system is designed or where it can be deployed, especially in countries with strict wireless communication rules. Following these regulations is important to ensure safe and legal operation.

- **Susceptibility to Interference:** RF sensing devices can be affected by interference from other wireless systems that operate on similar frequencies. This interference can reduce the quality of the data collected and may affect the accuracy of the system. To minimize these effects, it is important to apply good signal filtering, frequency control, and environmental adjustments during operation.

5.3 Future Work

In future, several areas will be further developed to improve the performance and real-world use of the proposed system.

- **Diverse Environmental Data Collection:** In future, the development of MM hearing aid devices can be further enhanced through broader environmental data collection. Expanding the dataset to include a variety of positions, orientations, and user interactions will strengthen the adaptability of the system across diverse conditions. Data gathered under varying lighting, background, and acoustic settings will enable future MM hearing aids to perform effectively in both indoor and outdoor environments, supporting reliable recognition of gestures and sign language in real-world scenarios.

- **Advanced Signal Processing and Noise Reduction:** Future research will focus on refining signal processing techniques to enhance the clarity and precision of RF data. Applying advanced filtering and adaptive noise reduction algorithms will improve feature extraction and classification accuracy. Enhanced signal processing will ensure that RF-based sensing remains consistent and efficient across different spatial arrangements and environmental conditions.

- **Integration of Practical RF Technologies:** The next phase of research will explore the use of real Wi-Fi–based RF sensing to support cost-effective and scalable system development. Utilising Wi-Fi CSI can provide a practical alternative for continuous and contactless activity detection. The implementation of low-cost embedded platforms, such as Raspberry Pi and other microcontroller-based systems, will enable efficient data capture and real-time processing, facilitating smooth integration into future hearing aid prototypes.

- **Optimisation of Deep Learning Models:** Further studies will focus on optimizing DL models to enhance recognition accuracy and real-time performance. Advanced architectures such as transformer and attention-based fusion models will be explored to effectively combine radar and video modalities. Additionally, model compression and quantization techniques will be employed to improve computational efficiency and enable deployment on lightweight embedded devices suitable for wearable applications.

- **Towards Real-Time Multimodal Hearing Aid Systems:** In future, the integration of radar, Wi-Fi, and visual modalities will lead to the creation of intelligent, real-time MM hearing aid systems. These advanced devices will combine high-speed data processing with robust DL models to provide seamless recognition of gestures and sign language. The successful implementation of this vision will contribute significantly to next-generation assistive technologies that enhance communication, inclusivity, and independence for the Deaf and hard-of-hearing community.

Bibliography

[1] National Health Service (NHS). *Hearing Loss*. https://www.nhs.uk/conditions/hearing-loss/. Accessed: January 14, 2026. 2025.

[2] WHO. *Deafness and hearing loss*. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. Accessed: 25 April 2024.

[3] UK Health Security Agency. *Health Matters: Hearing Loss Across the Life Course*. https://ukhsa.blog.gov.uk/2019/06/05/health-matters-hearing-loss-across-the-life-course/. Accessed: 25 April 2024. 2019.

[4] Ankita Wadhawan and Parteek Kumar. "Deep learning-based sign language recognition system for static signs". In: *Neural Computing and Applications* 32.12 (2020), pp. 7957–7968.

[5] Yu Liu et al. "Sign language recognition from digital videos using feature pyramid network with detection transformer". In: *Multimedia Tools and Applications* 82.14 (2023), pp. 21673–21685.

[6] Xuping Wu et al. "Ultra-robust and sensitive flexible strain sensor for real-time and wearable sign language translation". In: *Advanced Functional Materials* 33.36 (2023), p. 2303504.

[7] Sarah Qahtan et al. "A comparative study of evaluating and benchmarking sign language recognition system-based wearable sensory devices using a single fuzzy set". In: *Knowledge-Based Systems* 269 (2023), p. 110519.

[8] James McCleary et al. "Sign language recognition using micro-Doppler and explainable deep learning". In: *2021 IEEE Radar Conference (RadarConf21)*. 2021. DOI: 10.1109/RadarConf2147009.2021.9455257.

[9] M. Mahbubur Rahman et al. "Word-Level Sign Language Recognition Using Linguistic Adaptation of 77 GHz FMCW Radar Data". In: *Proceedings of the 2021 IEEE Radar Conference (RadarConf 21)*. Atlanta, GA, USA, 2021, pp. 1–6. DOI: 10.1109/RadarConf2147009.2021.9455190.

[10] Yukang Yan et al. “PrivateTalk: Activating Voice Input with Hand-On-Mouth Gesture Detected by Bluetooth Earphones”. In: *Proceedings of the 32nd ACM User Interface Software and Technology Symposium (UIST ’19)*. ACM, Oct. 2019. DOI: 10.1145/3332165.3347950.

[11] Xuyu Wang et al. “Vital signs monitoring with RFID: Opportunities and challenges”. In: *IEEE Network* 33.4 (2019), pp. 126–132. DOI: 10.1109/MNET.2019.1800014.

[12] Gerasimos Potamianos et al. “Audio-visual automatic speech recognition: An overview”. In: *Issues in visual and audio-visual speech processing* 22 (2004), p. 23.

[13] Yijia Lu et al. “Decoding lip language using triboelectric sensors with deep learning”. In: *Nature communications* 13.1 (2022), p. 1401.

[14] Hira Hameed et al. “Pushing the limits of remote RF sensing by reading lips under the face mask”. In: *Nature communications* 13.1 (2022), p. 5168.

[15] Yao Ge et al. “A comprehensive multimodal dataset for contactless lip reading and acoustic analysis”. In: *Scientific Data* 10.1 (2023), p. 895.

[16] Shigeng Zhang et al. “Hearme: Accurate and real-time lip reading based on commercial rfid devices”. In: *IEEE Transactions on Mobile Computing* (2022).

[17] Tao Zhang et al. “WiGrus: A Wifi-Based Gesture Recognition System Using Software-Defined Radio”. In: *IEEE Access* 7 (2019), pp. 131102–131113.

[18] Ahsen Tahir et al. “WiFreeze: Multiresolution Scalograms for Freezing of Gait Detection in Parkinsons Leveraging 5G Spectrum with Deep Learning”. In: *Electronics* 8.12 (2019), p. 1433.

[19] Hira Hameed et al. “Recognizing British Sign Language Using Deep Learning: A Contactless and Privacy-Preserving Approach”. In: *IEEE Transactions on Computational Social Systems* (2022).

[20] Elaine Rashbrook and Clare Perkins. *UK Health Security Agency, Health Matters: Hearing loss across the life course*. https://ukhsa.blog.gov.uk/2019/06/05/health-matters-hearing-loss-across-the-life-course. Accessed: 10 Jan 2024.

[21] Maree Johnson et al. “A systematic review of speech recognition technology in health care”. In: *BMC medical informatics and decision making* 14.1 (2014), pp. 1–14.

[22] Mostafa Magdy Balaha et al. “A vision-based deep learning approach for independent-users Arabic sign language interpretation”. In: *Multimedia Tools and Applications* 82.5 (2023), pp. 6807–6826.

[23] Deep R. Kothadiya et al. “Signformer: Deepvision transformer for sign language recognition”. In: *IEEE Access* 11 (2023), pp. 4730–4739.

[24] Ronglai Zuo, Fangyun Wei, and Brian Mak. “Natural Language-Assisted Sign Language Recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 14890–14900.

[25] Oscar Koller et al. “Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.9 (2019), pp. 2306–2320.

[26] Anshul Mittal et al. “A modified LSTM model for continuous sign language recognition using leap motion”. In: *IEEE Sensors Journal* 19.16 (2019), pp. 7056–7063.

[27] Chao Sun et al. “Discriminative exemplar coding for sign language recognition with Kinect”. In: *IEEE Transactions on Cybernetics* 43.5 (2013), pp. 1418–1428.

[28] Lamia Trabelsi et al. “Advancements and Challenges in Vision-Based Sign Language Recognition: A Comprehensive Review”. In: *Information Fusion* 126 (2025). Comprehensive review of vision-based sign language recognition using classical ML and deep learning (CNNs, LSTMs, Transformers). Highlights datasets, benchmark methods, and open challenges for scalable, multilingual, real-time SLR systems., p. 103626. DOI: 10.1016/j.inffus.2025.103626.

[29] Alexander Brettmann et al. “Breaking the Barriers: Video Vision Transformers for Word-Level Sign Language Recognition”. In: *arXiv preprint* (2025). Introduces a Video Vision Transformer (ViViT) for dynamic word-level ASL recognition using WLASL100. Achieved 75.6% Top-1 accuracy, outperforming I3D CNN baselines. Demonstrates superior temporal modeling but remains computationally heavy. arXiv: 2504.07792 [cs.CV].

[30] Abdulelah Ali Alkhoraif et al. “Ensemble Transformer-Based Word-Level Sign Language Recognition with Multi-Modal Input Fusion”. In: *Journal of Engineering Research* (2025). Proposes a two-stream Swin Transformer ensemble integrating appearance (RGB) and pose modalities via landmark-to-image transformation. Achieved 93.5% accuracy on WLASL and 92.6% on MS-ASL; outperforms prior multimodal SLR methods. DOI: 10.1016/j.jer.2025.07.006.

[31] Juan Song et al. “Hand-Aware Graph Convolution Network for Skeleton-Based Sign Language Recognition”. In: *Journal of Information and Intelligence* 3 (2025). Introduces HA-GCN emphasizing hand topology in skeleton-based SLR with adaptive DropGraph regularization. Achieved 96.8% and 99.6% accuracy on AUTSL and INCLUDE datasets. Reduces overfitting and enhances fine-grained hand motion modeling., pp. 36–50. DOI: 10.1016/j.jiixd.2024.08.001.

[32] Hamzah Luqman. “SignVLM: A Pre-Trained Large Video Model for Sign Language Recognition”. In: *PeerJ Computer Science* 11 (2025). Presents SignVLM, a CLIP-based pre-trained large video model for sign language recognition using Transformer decoders. Achieved state-of-the-art accuracy across KArSL, WLASL, LSA64, and AUTSL datasets with strong zero/few-shot generalization., e3112. DOI: 10.7717/peerj-cs.3112.

[33] Yan Ren et al. “Multi-Modal Isolated Sign Language Recognition Based on Deep Learning”. In: *Pattern Recognition Letters* 180 (2025). Proposes a CNN–LSTM framework fusing RGB and skeleton modalities for isolated sign recognition. Achieves higher accuracy than single-modality baselines and demonstrates robustness across lighting conditions. Limitations include high computational cost and lack of continuous SLR support., pp. 1–12. DOI: 10.1016/j.patrec.2025.02.005.

[34] Qian Zhou et al. “Fusion of Multimodal Spatio-Temporal Features and 3D Deformable Convolution Based on Sign Language Recognition in Sensor Networks”. In: *Sensors* 25.14 (2025), p. 4378.

[35] Mohamed Aly and Islam S. Fathi. “Recognizing American Sign Language Gestures Efficiently and Accurately Using a Hybrid Transformer Model”. In: *Scientific Reports* 15 (2025). Develops a hybrid CNN–Transformer model for static ASL alphabet recognition. Attains 99.97% accuracy and 110 FPS on ASL datasets with low computational complexity. Excels in efficiency but limited to static gestures without temporal modeling., p. 24291. DOI: 10.1038/s41598-025-09483-0.

[36] Esraa Hassan et al. “A Novel Model for Expanding Horizons in Sign Language Recognition”. In: *Scientific Reports* 15 (2025). Presents the Sign Nevestro DenseNet Attention (SNDA) model integrating DenseNet, attention mechanisms, and Nadam optimization for ASL recognition. Achieves 99.76% accuracy on ASL datasets; robust to lighting variation but limited to static signs., p. 24358. DOI: 10.1038/s41598-025-09643-2.

[37] Shi-Wei Gan et al. “Vision-Based Sign Language Translation via a Skeleton-Aware Neural Network”. In: *Journal of Computer Science and Technology* 40.2 (2025). Proposes SANet, a skeleton-aware neural network for vision-based sign language translation. Combines RGB and skeleton modalities via GCN-based feature weighting and achieves superior accuracy on SLT datasets. Includes smartphone deployment; limited by pose-estimation accuracy and computational demand., pp. 378–396. DOI: 10.1007/s11390-024-2978-y.

[38] Abdulrahman Baihan et al. “Sign Language Recognition Using Modified Deep Learning Network and Hybrid Optimization: A Hybrid Optimizer (HO) Based Optimized CNNSa-LSTM Approach”. In: *Scientific Reports* 14 (2024). Introduces a CNNSa-LSTM model with a hybrid optimizer for vision-based sign language recognition. Achieves im-

proved accuracy and convergence over CNN and LSTM baselines. Demonstrates strong generalization but limited interpretability and high computational complexity., p. 16273. DOI: 10.1038/s41598-024-59764-9.

[39] Saud Alyami and Hamzah Luqman. "A Comparative Study of Continuous Sign Language Recognition Techniques". In: *arXiv preprint* (2025). Compares Transformer-, CNN-LSTM-, and graph-based continuous sign language recognition (CSLR) models on PHOENIX-Weather, CSL, and AUTSL datasets. Finds Transformers yield best contextual performance, while CNN-RNN hybrids are better for low-resource setups. Dataset dependence and multilingual limitations noted. arXiv: 2406.12369 [cs.CV].

[40] Shahzeen Ijaz Ahmad et al. "Sign Assist: Real-Time Isolated Sign Language Recognition and Translator". In: *3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement (AVSEC)*. Proposes Sign Assist, a real-time ISLR system integrating MediaPipe holistic keypoint extraction with LSTM-based sequential modeling and GPT-driven translation. Introduces PSL20 dataset (20 dynamic gestures). Achieves strong real-time accuracy but limited vocabulary and signer independence. Kos, Greece: ISCA, 2024, pp. 82–90. DOI: 10.21437/AVSEC.2024-18.

[41] Naif A. Baghdadi et al. "Toward Robust Arabic Sign Language Recognition via Vision Transformers and LIME Integration". In: *ScienceOpen Preprints* (2024). Presents an interpretable Vision Transformer framework with LIME-based explainability for Arabic Sign Language recognition. Achieves 99.46% and 99.88% accuracy on ArSL21L and RGB Arabic Alphabets datasets. Offers transparency and robustness but limited to static signs and high compute cost. DOI: 10.14293/S2199-1006.1.SOR-.PPV1HPZ.v1.

[42] Eduardo Carneiro et al. "Sign Language Recognition Based on Deep Learning and Low-Cost Handcrafted Descriptors". In: *Multimedia Tools and Applications* 83.12 (2024). Combines deep learning and handcrafted descriptors for efficient vision-based SLR. Employs CenterNet for hand/face detection and CNN–RNN modeling on AUTSL dataset, improving accuracy by 7.96% with minimal computational cost. Limited by occlusion sensitivity and dependency on accurate detections., pp. 34617–34641. DOI: 10.1007/s11042-024-17982-5.

[43] Wei Zhao et al. "Continuous Sign Language Recognition With Correlation Network". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Introduces CorrNet, a correlation-based deep learning model for continuous SLR. Learns inter-frame body trajectories via correlation and identification modules. Achieves state-of-the-art accuracy on PHOENIX14, CSL-Daily, and CSL datasets. High training complexity remains a limitation. IEEE, 2024, pp. 18322–18331. DOI: 10.1109/CVPR52733.2024.01832.

[44] Yuqi Liu et al. “SCOPE: Sign Language Contextual Processing with Embedding from LLMs”. In: *arXiv preprint* (2024). Proposes SCOPE, a context-aware vision-language framework integrating LLMs with vision-based SLR and SLT. Aligns motion features with LLM embeddings and fine-tunes via Q-LoRA. Introduces a 72-hour Chinese Sign Language dialogue dataset; achieves state-of-the-art results. Computationally intensive and language-specific. arXiv: 2409.01073 [cs.CV].

[45] Zhen Li et al. “Multi-View Spatio-Temporal Graph Transformer for Continuous Sign Language Translation”. In: *IEEE Transactions on Multimedia* 26 (2024). Develops a Multi-View Spatio-Temporal Graph Transformer (MV-STGT) for continuous sign language translation. Uses pose graphs and RGB views with cross-attention fusion, achieving superior BLEU-4 and ROUGE-L on PHOENIX14T and CSL-Daily. High compute and pose dependency are main constraints., pp. 4218–4231. DOI: 10.1109/TMM.2024.3390211.

[46] Hui Zhang et al. “Vision-Based Sign Language Recognition Using Depth-Aware Spatio-Temporal Transformer”. In: *Pattern Recognition* 154 (2025). Introduces a Depth-Aware Spatio-Temporal Transformer (DASTT) combining RGB and depth modalities for fine-grained sign recognition. Enhances robustness to lighting and background changes, improving Top-1 accuracy by 5.8% on AUTSL and MS-ASL. Limited by need for depth sensors and higher latency., p. 110667. DOI: 10.1016/j.patcog.2024.110667.

[47] Chenghong Lu, Shingo Amino, and Lei Jing. “Data Glove with Bending Sensor and Inertial Sensor Based on Weighted DTW Fusion for Sign Language Recognition”. In: *Electronics* 12.3 (2023), p. 613.

[48] Yun Li et al. “A sign-component-based framework for Chinese sign language recognition using accelerometer and sEMG data”. In: *IEEE Transactions on Biomedical Engineering* 59.10 (2012), pp. 2695–2704. DOI: 10.1109/TBME.2012.2190734.

[49] Vasiliki E. Kosmidou and Leontios J. Hadjileontiadis. “Sign language recognition using intrinsic-mode sample entropy on sEMG and accelerometer data”. In: *IEEE Transactions on Biomedical Engineering* 56.12 (2009), pp. 2879–2890. DOI: 10.1109/TBME.2009.2013200.

[50] Karly Kudrinko et al. “Wearable Sensor-Based Sign Language Recognition: A Comprehensive Review”. In: *IEEE Reviews in Biomedical Engineering* 14 (2021), pp. 305–320. DOI: 10.1109/RBME.2021.3071052.

[51] M. A. Ahmed and A. R. Elshafee. “A Review on Systems-Based Sensory Gloves for Sign Language Recognition”. In: *Sensors* 18.6 (2018), p. 1884. DOI: 10.3390/s18061884.

[52] A. Ben Haj Amor and A. Ben Hamida. “Sign Language Recognition Using the Electromyographic Signal: A Systematic Literature Review”. In: *Sensors* 23.5 (2023), p. 2341. DOI: 10.3390/s23052341.

[53] M. Taneja, R. Kumar, and S. Singh. “A Comprehensive Review of Sensor-Based Sign Language Recognition Models”. In: *Journal of Xi’an University of Architecture & Technology* 15.4 (2023), pp. 122–134.

[54] Qian Zhang, Yue Gu, et al. “WearSign: Pushing the Limit of Sign Language Translation Using Inertial and EMG Wearables”. In: *Proceedings of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 6.3 (2022), pp. 1–27. DOI: 10.1145/3546739.

[55] A. Qaroush, M. Al-Qudah, and M. Al-Zyoud. “Smart, Comfortable Wearable System for Recognizing Arabic Sign Language in Real-Time Using IMUs and Feature-Based Fusion”. In: *Expert Systems with Applications* 179 (2021), p. 115029. DOI: 10.1016/j.eswa.2021.115029.

[56] İ. Umut and Y. Kumdereli. “Novel Wearable System to Recognize Sign Language in Real Time”. In: *Sensors* 24.2 (2024), p. 451. DOI: 10.3390/s24020451.

[57] Y. Liu, X. Wang, and H. Zhang. “A Wearable System for Sign Language Recognition Enabled by Attachable Sensors”. In: *Advanced Functional Materials* 33.14 (2023), p. 2301123. DOI: 10.1002/adfm.202301123.

[58] J. Park, S. Kim, and D. Lee. “A Multi-Sensor Fusion Framework for Robust Real-Time Korean Sign Language Recognition Using EMG and IMU Wearables”. In: *IEEE Access* 10 (2022), pp. 80152–80163. DOI: 10.1109/ACCESS.2022.3198156.

[59] S. Lee, J. Kim, and T. Kang. “A Flexible Textile-Based Wearable Sensor Network for Continuous Sign Language Recognition”. In: *Sensors* 20.12 (2020), p. 3402. DOI: 10.3390/s20123402.

[60] P. Ghosh and R. Banerjee. “Hand Motion and Flexion Sensor Fusion for Real-Time Indian Sign Language Recognition”. In: *IEEE Sensors Journal* 23.19 (2023), pp. 20845–20856. DOI: 10.1109/JSEN.2023.3310502.

[61] L. Chen, J. Li, and Z. Wu. “Hybrid EMG–IMU Wearable System for Continuous Chinese Sign Language Recognition”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022), pp. 1452–1464. DOI: 10.1109/TNSRE.2022.3178445.

[62] M. Ibrahim, H. Khaled, and S. Al-Saad. “Low-Power Embedded Glove System for Arabic Sign Language Recognition Using Edge Learning”. In: *Proceedings of the International Conference on Internet of Things and Embedded Systems*. 2024, pp. 134–140. DOI: 10.1109/ICITES.2024.01234.

[63] H. El Khoukhi, A. Jbari, and L. Boukhriss. “Moroccan Sign Language Recognition with a Sensory Glove Using Artificial Neural Networks”. In: *Sensors* 25.4 (2025), p. 1761. DOI: 10.3390/s25041761.

[64] A. A. Saleh and P. Joshi. “Sensor-Equipped Gloves for Recognizing Arabic Sign Language Using Machine Learning”. In: *International Journal of Advanced Manufacturing (IJAM)* 12.3 (2025), pp. 78–88.

[65] H. Abdelrahman, N. Eldin, and R. Shahid. “Smart Glove System for Real-Time American Sign Language Recognition Using IMU and Flex Sensors”. In: *IEEE Access* 11 (2023), pp. 103245–103257. DOI: 10.1109/ACCESS.2023.3298912.

[66] P. Yadav and M. Singh. “Multimodal Wearable Sensor Fusion for Continuous Indian Sign Language Recognition”. In: *IEEE Sensors Letters* 8.2 (2024), pp. 150–158. DOI: 10.1109/LSENS.2024.3341821.

[67] S. Rahman, T. Ali, and A. Rehman. “Low-Cost Embedded Sign Language Translator Using Arduino-Based Glove and Machine Learning”. In: *Microprocessors and Microsystems* 100 (2023), p. 105785. DOI: 10.1016/j.micpro.2023.105785.

[68] Hyeon-Jun Kim and Soo-Whang Baek. “Application of Wearable Gloves for Assisted Learning of Sign Language Using Artificial Neural Networks”. In: *Processes* 11.4 (2023), p. 1065. DOI: 10.3390/pr11041065.

[69] A. Ben Haj Amor and A. Ben Hamida. “Sign Language Recognition Using the Electromyographic Signal: A Systematic Literature Review”. In: *Sensors* 23.5 (2023), p. 2341. DOI: 10.3390/s23052341.

[70] Sevgi Z. Gurbuz et al. “American sign language recognition using RF sensing”. In: *IEEE Sensors Journal* 21.3 (2020), pp. 3763–3775. DOI: 10.1109/JSEN.2020.3022376.

[71] M. Mahbubur Rahman et al. “Word-level ASL recognition and trigger sign detection with RF sensors”. In: *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, Ontario, Canada, 2021, pp. 8233–8237. DOI: 10.1109/ICASSP39728.2021.9414063.

[72] Victor C. Chen. *The micro-Doppler effect in radar*. 2nd. Norwood, MA: Artech House, 2019, p. 450. ISBN: 9781630815462.

[73] Sevgi Zubeyde Gurbuz et al. “Micro-Doppler-based in-home aided and unaided walking recognition with multiple radar and sonar systems”. In: *IET Radar, Sonar & Navigation* 11.1 (2017), pp. 107–115. DOI: 10.1049/iet-rsn.2016.0055.

[74] Zhu Wang et al. “Gesture-Radar: A Dual Doppler Radar Based System for Robust Recognition and Quantitative Profiling of Human Gestures”. In: *IEEE Transactions on Human-Machine Systems* 51.1 (2020), pp. 32–43. DOI: 10.1109/THMS.2020.3036637.

[75] Zhiwen Yu et al. “SoDar: Multitarget Gesture Recognition Based on SIMO Doppler Radar”. In: *IEEE Transactions on Human-Machine Systems* 52.2 (2022), pp. 276–289. DOI: 10.1109/THMS.2021.3088234.

[76] J. Shang, C. Wu, and J. Yang. “A Robust Sign Language Recognition System with Multiple Wi-Fi Devices (WiSign)”. In: *Proceedings of the ACM Workshop on Wireless of the Students, by the Students, for the Students (S3)*. ACM. 2017. DOI: 10.1145/3084450.3084458.

[77] Y. Ma et al. “SignFi: Sign Language Recognition Using WiFi”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2.1 (2018), 23:1–23:21. DOI: 10.1145/3191755.

[78] T. Ahmed, M. Nafees, and M. Shihab. “Wi-Fi CSI Based Human Sign Language Recognition Using LSTM Network”. In: *International Conference on Electrical, Computer and Communication Engineering (ECCE)*. 2021. DOI: 10.1109/ICECCE52056.2021.9514223.

[79] T. Ahmed et al. “DF-WiSLR: Device-Free Wi-Fi-based Sign Language Recognition”. In: *Pervasive and Mobile Computing* 68 (2020), p. 101266. DOI: 10.1016/j.pmcj.2020.101266.

[80] Z. Zhang et al. “WiSign: Ubiquitous American Sign Language Recognition Using Commercial Wi-Fi Devices”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 11.6 (2020), pp. 1–22. DOI: 10.1145/3406109.

[81] R. Santhalingam and F. Adib. “mmASL: Environment-Independent ASL Gesture Recognition Using 60 GHz Millimeter-Wave Signals”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 4.4 (2020), pp. 1–26. DOI: 10.1145/3432224.

[82] S. Z. Gurbuz et al. “American Sign Language Recognition Using RF Sensing”. In: *IEEE Sensors Journal* 20.22 (2020), pp. 13736–13747. DOI: 10.1109/JSEN.2020.3009820.

[83] E. Kurtoglu, M. Rahman, and S. Z. Gurbuz. “ASL Trigger Recognition in Mixed Activity/Signing Sequences for RF Sensor-Based User Interfaces”. In: *IEEE Transactions on Human-Machine Systems* 52.6 (2022), pp. 1000–1012. DOI: 10.1109/THMS.2022.3148935.

[84] E. Kurtoglu et al. “Interactive Learning of Natural Sign Language with Radar”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2024. DOI: 10.1109/ICASSP48485.2024.9745013.

[85] M. M. Rahman et al. “Word-Level ASL Recognition and Trigger Sign Detection with RF Sensors”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021. DOI: 10.1109/ICASSP39728.2021.9414063.

[86] M. Kulhandjian, A. Awan, and R. Kachroo. “Sign Language Gesture Recognition Using Doppler Radar and Deep Learning”. In: *IEEE Sensors Letters* 3.12 (2019), pp. 1–4. DOI: 10.1109/LSENS.2019.2955382.

[87] R. Gao, L. Zhang, and H. Liu. “A Multitask Sign Language Recognition System Using Commodity Wi-Fi”. In: *Wireless Communications and Mobile Computing* (2023). DOI: 10.1155/2023/9374105.

[88] J. Qin et al. “WiASL: American Sign Language Writing Recognition System Using Commercial WiFi Devices”. In: *Measurement* 215 (2023), p. 112123. DOI: 10.1016/j.measurement.2023.112123.

[89] H. Zhang et al. “RF-Sign: Position-Independent Sign Language Recognition Using Passive RFID Tags”. In: *IEEE Internet of Things Journal* 10.20 (2023), pp. 18123–18136. DOI: 10.1109/JIOT.2023.3322228.

[90] R. Mahbub et al. “mm-SLR: Millimeter-Wave Radar-Assisted Sign Language Understanding in Dynamic Environments”. In: *IEEE Radar Conference (RadarConf)*. 2024. DOI: 10.1109/RadarConf.2024.9763211.

[91] M. Nafee et al. “WiBaSL: A Wi-Fi-Based Word-Level Deep Bangladeshi Sign Language Dataset with CSI”. In: *Sensors* 25.4 (2025), p. 1234. DOI: 10.3390/s25041234.

[92] M. Tariq et al. “Deep Learning Sign Language Recognition System Based on Wi-Fi CSI”. In: *IEEE Access* 8 (2020), pp. 228145–228158. DOI: 10.1109/ACCESS.2020.3046835.

[93] M. Thariq, M. Nafees, and K. Rahman. “Sign Language Gesture Recognition with Bispectrum Features Using SVM”. In: *IEEE International Conference on Signal Processing and Integrated Networks (SPIN)*. 2019. DOI: 10.1109/SPIN.2019.8711712.

[94] H. Zheng et al. “CSI-Cro: A Cross-Domain CSI Sign Language Recognition System Based on Dual-Attention Feature Fusion”. In: *IEEE Transactions on Mobile Computing* (2025). DOI: 10.1109/TMC.2025.3334512.

[95] Joe G Greener et al. “A guide to machine learning for biologists”. In: *Nature Reviews Molecular Cell Biology* 23.1 (2022), pp. 40–55.

[96] Ramon Mayor Martins and Christiane Gresse Von Wangenheim. “Findings on Teaching Machine Learning in High School: A Ten-Year Systematic Literature Review”. In: *Informatics in Education* (2022).

[97] Jafar Alzubi, Anand Nayyar, and Akshi Kumar. "Machine learning from theory to algorithms: an overview". In: *Journal of physics: conference series*. Vol. 1142. 1. IOP Publishing. 2018, p. 012012.

[98] Rakesh Singh Khanzode and Sachin Sarode. "Advantages and applications of machine learning". In: *International Journal of Advanced Research in Computer Science* 11.3 (2020), pp. 1–5.

[99] Thomas Wuest et al. "Machine learning in manufacturing: advantages, challenges, and applications". In: *Production & Manufacturing Research* 4.1 (2016), pp. 23–45.

[100] Michael I Jordan and Tom M Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245 (2015), pp. 255–260.

[101] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[102] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[103] Andre Esteva et al. "A guide to deep learning in healthcare". In: *Nature Medicine* 25 (2019), pp. 24–29.

[104] Batta Mahesh. "Machine learning algorithms-a review". In: *International Journal of Science and Research (IJSR).[Internet]* 9 (2020), pp. 381–386.

[105] Mohamed Alloghani et al. "A systematic review on supervised and unsupervised machine learning algorithms for data science". In: *Supervised and unsupervised learning for data science* (2020), pp. 3–21.

[106] Sergey Levine et al. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems". In: *arXiv preprint arXiv:2005.01643* (2020).

[107] Ian Osband et al. "Behaviour suite for reinforcement learning". In: *arXiv preprint arXiv:1908.035* (2019).

[108] Adam C Mater and Michelle L Coote. "Deep learning in chemistry". In: *Journal of chemical information and modeling* 59.6 (2019), pp. 2545–2559.

[109] Timothy P Lillicrap et al. "Random synaptic feedback weights support error backpropagation for deep learning". In: *Nature communications* 7.1 (2016), pp. 1–10.

[110] Abul Bashar et al. "Survey on evolving deep learning neural network architectures". In: *Journal of Artificial Intelligence* 1.02 (2019), pp. 73–82.

[111] Ochin Sharma. "Deep challenges associated with deep learning". In: *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*. IEEE. 2019, pp. 72–75.

[112] Ajay Shrestha and Ausif Mahmood. "Review of deep learning algorithms and architectures". In: *IEEE access* 7 (2019), pp. 53040–53065.

[113] Tobias Gruber et al. “On deep learning-based channel decoding”. In: *2017 51st Annual Conference on Information Sciences and Systems (CISS)*. IEEE. 2017, pp. 1–6.

[114] Musab Coşkun et al. “An overview of popular deep learning methods”. In: *European Journal of Technique (EJT)* 7.2 (2017), pp. 165–176.

[115] Jürgen Schmidhuber. “Deep learning in neural networks: An overview”. In: *Neural Networks* 61 (2015), pp. 85–117.

[116] Li Deng and Dong Yu. “A Tutorial Survey of Deep Learning”. In: *Foundations and Trends in Signal Processing* 7.3–4 (2014), pp. 197–387.

[117] Geert Litjens et al. “A survey on deep learning in medical image analysis”. In: *Medical Image Analysis* 42 (2017), pp. 60–88.

[118] Aisha Fatima et al. “Utilizing Contactless Sensing Technology for the Identification of Hand and Head Movements in Conjunction with Facial Expressions”. In: *IEEE Sensors Journal* (2025). DOI: 10.1109/JSEN.2025.xxxxx.

[119] IEEE. *IEEE Standard for Safety Levels With Respect to Human Exposure to Radio Frequency Electromagnetic Fields, 3 kHz to 300 GHz*. IEEE Standard C95.1-2005. Publisher: IEEE. Apr. 2006.

[120] Hira Hameed et al. “Wi-Fi and Radar Fusion for Head Movement Sensing Through Walls Leveraging Deep Learning”. In: *IEEE Sensors Journal* (2023).

[121] Hira Hameed et al. “Recognizing British Sign Language Using Deep Learning: A Contactless and Privacy-Preserving Approach”. In: *IEEE Transactions on Computational Social Systems* 10.4 (Aug. 2023), pp. 2090–2098. ISSN: 2329-924X, 2373-7476. DOI: 10.1109/TCSS.2022.3210288. URL: https://ieeexplore.ieee.org/document/9918647/ (visited on 01/05/2024).

[122] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.