Contactless AI-Enabled Hybrid Sensing For Cognitive Impairment

Hira Hameed

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

School of Engineering
College of Science and Engineering
University of Glasgow

January 2024

Abstract

This thesis covers various aspects of Multi Model (MM) hearing impairments including human speech, sign language, behavior analysis, and facial expressions, which facilitate the deaf communities. Recent research using wearable, audio, and visual technologies for monitoring cognitive impairments in deaf individuals has its benefits but also presents certain limitations. For instance, while wearable devices provide body-mounted monitoring, their constant use can be uncomfortable, and there is a risk that deaf individuals might forget to wear them. Moreover, these devices need regular removal for recharging. In audio noise, even individuals with regular hearing may struggle to clearly hear someone's voice. Camera-based visual information raises privacy concerns, and legal implications might restrict its broad usage in public and private areas due to issues like filming without consent, which is illegal in many countries. This thesis explores the use of Radio Frequency (RF) signals to sense human speech, sign recognition, behavior identification, and facial expressions using Wi-Fi, radar, and Radio Frequency Identification (RFID) signals. RF sensing provides an exciting opportunity for next-generation MM hearing aid devices. The RF-based hearing aid just requires Tx and Rx on a single chip. Additionally, RF signal in the form of Wi-Fi is currently present in many homes. People move around Wi-Fi signals, signal propagation is affected. Channel State Information (CSI) in Wi-Fi describes how a signal propagates from the transmitter to the receiver. In this thesis, the data collected in the form of CSI, micro-doppler, and Received Signal Strength Indicator (RSSI) signals are fed into Machine Learning (ML) and Deep Learning (DL) techniques for classification purposes. The proposed techniques successfully differentiate various activities, such as speech recognition, sign language recognition, and behavior analysis by using head movements, and facial expressions to understand the expressions of individuals when communicating with deaf people. These techniques utilise RF signals to individually differentiate each activity and achieve over 90% test accuracy. This thesis serves as a proof of concept for contactless MM hearing aid systems that can assist deaf people with different perspectives to live independently without the need to wear monitoring devices, audio, and visual devices.

Contents

List of Tables

List of Figures

List of Abbreviations

2DCRNN	2D Convolutional Recurring Neural Network
3DCNN	3D Convolutional Neural Network
ADC	Analog-to-Digital Converter
AI	Artificial Intelligence
ASL	American Sign Language
AHGR	Amphibious Hierarchical Gesture Recognition
ANN	Artificial Neural Network
AV	Audio-Visual
CI	Confidence Interval
CL	Cognitive Load
CSI	Channel State Information / Communication, Sensing, and Imaging
CSV	Comma Separated Value
DL	Deep Learning
DSN	Domain Separation Network
EMG	Electromyogram
FC	Fully Connected
FMCW	Frequency Modulated Continuous Wave
FPS	Frame Per Second
FFT	Fast Fourier Transform
FPR	False Positive Rate

FL	Federated Learning
GAN	Generative Adversarial Network
GNU	GNU’s Not Unix
HAs	Hearing Aids
HMM	Hidden Markov Models
HOG	Histogram of Oriented Gradients
IC	Integrated Circuit
ICA	Independent Component Analysis
CIR	Impulse Response
ISL	Indian Sign Language
IMUs	Inertial Measurement Units
IOT	Internet Of Things
KNN	K-Nearest Neighbour
LOS	Line Of Sight
LSTM	Long Short-Term Memory Networks
ML	Machine Learning
MSPS	Mega Samples Per Second
MM	Multi Modal
MTI	Moving Target Indicator
MCNN	Multimodal Convolutional Neural Network
MV-DNN	Multi-View Deep Neural Network for Chinese Sign Language
NaN	Not a Number
NN	Neural Network
NNP	Neural network pattern
NLOS	Non Line Of Sight

PCA	Principal Component Analysis
RBF	Radial Basis Function
RFID	Radio Frequency Identification
Rx	Reciever
ReLU	Rectified Linear Unit
RFaceID	RF-based face recognition system
SDRs	Software Defined Radios
SE	Speech Enhancement
SLR	Sign Language Recognition
STFT	Short Time Fourier Transform
SL	Sign Language
SVM	Support Vector Machine
SUS	System Usability Scale
TCN	Temporal Neural Networks
TX	Transmitter
TPR	True Positive Rate
UHF	Ultra High Frequency
UK	United Kingdom
UWB	Ultra Wideband
USRP	Universal Software Radio Peripheral
VGG	Visual Geometry Group
VSR	Visual Speech Recognition

List of Publications

During my PhD. tenure, 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. Here is a comprehensive list of all my publications from this period.

A. *Articles*

(1) **Hameed, H**., Usman, M., Tahir, A., Hussain, A., Abbas, H., Cui, T.J., Imran, M.A. and Abbasi, Q.H., 2022. Pushing the limits of remote RF sensing by reading lips under the face mask". Nature Communications, 13(1), p.5168.

(2) **Hameed, H**., Usman, M., Tahir, A., Ahmad, K., Hussain, A., Imran, M.A. and Abbasi, Q.H., 2022. Recognizing British Sign Language Using Deep Learning: A Contactless and Privacy-Preserving Approach. IEEE Transactions on Computational Social Systems.

(3) **Hameed, H**., Wi-Fi and Radar Fusion for Head Movement Sensing Through Walls Leveraging Deep Learning, in IEEE Sensors Journal, doi: 10.1109/JSEN.2023.3337515.

(4) **Hameed, H**., Artificial Intelligence-Enabled Smart Mask For Speech Recognition For Future Hearing Devices, (Transactions on Audio, Speech, and Language Processing, IEEE/ACM) under Revision.

(5) **Hameed, H**., RF Sensing Enabled Tracking Of Human Facial Expressions Using Machine Learning Algorithms, (IEEE Transactions on Biometrics, Behavior, and Identity Science) Under Review.

(6) Abbasi, Q. H., Tang, C., Ghadban, N., **Hameed, H**., Usman, M., Hussain, A. and Imran, M. A. (2023) Towards AI-assisted RF hearing aids. ENT and Audiology News, 32(2).

(7) Lubna, L., **Hameed, H**., Ansari, S., Zahid, A., Sharif, A., Abbas, H.T., Alqahtani, F., Mufti, N., Ullah, S., Imran, M.A. and Abbasi, Q.H., 2022. Radio frequency sensing and its innovative applications in diverse sectors: A comprehensive study. Frontiers in Communications and Networks, 3, p.1010228.

(8) Saeed, U., Shah, S. A., Ghadi, Y. Y., Khan, M. Z., Ahmad, J., Shah, S. I., **Hameed, H**. and Abbasi, Q. (2023), Extracting visual micro-doppler signatures from human lips motion using UoG radar sensing data for hearing aid applications. IEEE Sensors Journal, 23(19), pp. 22111-22118. (doi: 10.1109/JSEN.2023.3308972).

(9) Saeed, U., Shah, S. A., Ghadi, Y. Y., **Hameed, H.**, Shah, S. I., Ahmad, J., and Abbasi, Q. (2023) British Sign Language Detection Using Ultra-Wideband Radar Sensing and Residual Neural Network. IEEE Sensors Journal.Under Review

B. *Conference Proceedings*

(1) **Hameed, H**., Usman, M., Khan, M. Z., Hussain, A., Abbas, H. , Imran, M. A. and Abbasi, Q. H. (2022) Privacy-Preserving British Sign Language Recognition Using Deep Learning. In: 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'22), Glasgow, Scotland, United Kingdom, 11-15 July 2022, pp. 4316-4319. ISBN 9781728127828 (doi: 10.1109/EMBC48229.2022.9871491).

(2) **Hameed, H**., Elsayed, M., Farooq, M., Kaur, J., Usman, M., Hussain, A., Abd El-Latif, A. A., Imran, M. and Abbasi, Q. H. (2023) Non-Invasive Hand Gestures Recognition With Machine Learning Algorithms. In: International Conference on Cybersecurity, Cybercrimes, and Smart Emerging Technologies (CCSET2023), Riyadh, Saudi Arabia, 5-7 December 2023, (Accepted for Publication).

(3) **Hameed, H**., Identification of Hearing-Impaired People in Crowded Environments Using Wi-Fi Signals, 2023 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting (USNC-URSI), Portland, OR, USA, 2023, pp. 279-280, doi: 10.1109/USNC-URSI52151.2023.10237903.

(4) **Hameed, H**., Contactless Privacy-Preserving Head Movement Recognition Using Deep Learning for Driver Fatigue Detection, 2023 International Symposium on Networks, Computers and Communications (ISNCC), Doha, Qatar, 2023, pp. 1-6, doi: 10.1109/ISNCC58260.2023.10323825.

(5) Ghadban N, Usman M, Tang C, Ghanam H, **Hameed H**, Vinciarelli A, Abbasi QH, Imran MA. (2023) Detecting Phonetic Characters using Radar Data. In: IEEE International Radar Conference 2023, Sydney, Australia, 6-10 Nov 2023, (Accepted for Publication).

(6) **Hameed, H**., Privacy-Preserving Visual Cues Communication for Hearing-Impaired People Using Deep Learning (Under Review).

(7) **Hameed, H**., Lubna, Ghadban, N., Usman, M., Arshad, K., Assaleh, K., Alkhayyat, A., Imran, M. A. and Abbasi, Q. (2023) TAQWA: Teaching Adolescents Quality Wadhu/Ablution Contactlessly Using Deep Learning. In: 6th International Conference of Signal Processing and Intelligent Systems (ICSPIS'23), Dubai, United Arab Emirates, 8-9 Nov 2023, (Accepted for Publication).

(8) **Hameed, H**., Azam, N., Usman, M., Abbas, H. , Imran, M. A. and Abbasi, Q. H. (2022) RF Sensing For Smoking Detection At Oil Fields. In: 2022 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting, Denver, CO, USA, 10-15 Jul 2022, pp. 944-945. ISBN 9781665496582 (doi: 10.1109/AP-S/USNC-URSI47032.2022.9887288).

(9) Farooq, M., **Hameed, H.**, Taha, A., Imran, M., Abbasi, Q. H. and Abbas, H. T. (2023) Contactless Respiration Variability Detection and Accuracy Test Using UWB Radar. In: 18th European Conference on Antennas and Propagation (EuCAP 2024), Glasgow, Scotland, 17-22 March 2024, (Accepted for Publication)

C. *Published Datasets*

(1) **Hameed, H**., Usman, M., Tahir, A., Imran, M. and Abbasi, Q. (2022) Recognizing British Sign Language using Deep Learning: A Contact-less and Privacy-Preserving Approach. [Data Collection].

(2) **Hameed, H**., Usman, M., Tahir, A., Abbas, H. , Imran, M. and Abbasi, Q. (2022) Pushing the Limits of Remote RF Sensing: Reading Lips Under Face Mask. [Data Collection].

(3) **Hameed, H**., Tahir, A., Usman, M., Jiang, Z., Lubna, , Abbas, H. , Naeem, R., Tie Jun, C., Imran, M. A. and Abbasi, Q. (2024) Wi-Fi and Radar Fusion for Head Movement Sensing Through Walls Leveraging Deep Learning. [Data Collection].

Acknowledgements

Throughout the course of my doctoral research, which spanned two and a half years, I was honored to benefit from the invaluable support and mentorship of numerous esteemed individuals, to whom I express my deepest gratitude. I would like to extend my earnest appreciation to my academic colleagues, whose pivotal contributions played a crucial role in the fruition of my project.

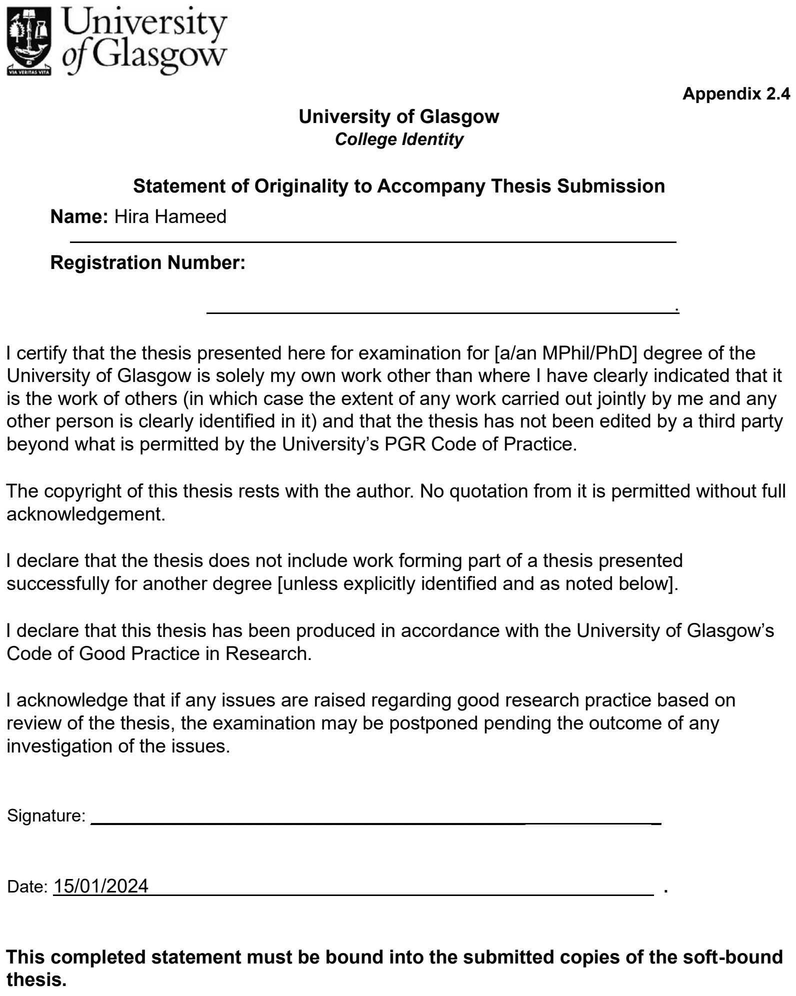
I would like to seize this moment to extend my heartfelt gratitude to my esteemed academic supervisors, Prof. Qammer H. Abbasi, for their unwavering support and invaluable guidance throughout my PhD journey. His profound insights and thoughtful mentorship have not only shaped my academic pursuits but have also been a source of great inspiration and encouragement in my personal life. His dedication and commitment to excellence have profoundly impacted my growth and development, both as a scholar and an individual.

I would like to thank Prof. Muhammad Imran for his continued support and invaluable input throughout my PhD. I am also indebted to my colleagues at the Communication, Sensing, and Imaging (CSI) Lab, especially Dr. Muhammad Usman and Dr. Ahsen Tahir, for their insightful discussions, advice, and participation in my experiments.

I am also extremely thankful to the University of Glasgow for supporting me in my hardship time.

Last but certainly not least, my heartfelt thanks go to my parents and my husband for their unwavering support and encouragement throughout my PhD journey. Their belief in my endeavors has been a constant source of motivation.

Declaration



Appendix 2.4

University of Glasgow
College Identity

Statement of Originality to Accompany Thesis Submission

Name: Hira Hameed

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: 15/01/2024 ______________________ .

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

Normal hearing is defined as the ability to hear a sound of 20dB level and above. Inability to hear this threshold can be recognised as hearing loss[1]. Hearing loss can be mild or severe and the subjects are referred to as ‘hard of hearing’. Hearing loss and deafness is a major impediment to normal communication and learning. Overall, 5% of the World’s population, around 430 million people suffer from hearing impairments. The number is expected to increase to 700 million people by 2050 [1]. In the United Kingdom (UK) alone, around 11 million individuals live with hearing impairments and age-related hearing loss has become a serious concern [2]. Next-generation hearing aids by 2050 require transformative Multi Modal (MM) processing, uninhibited by limitations of speech or sound enhancement. Lip-reading, sign language, head movements, and face reading have gained notable research attention in recent years due to its significance in many applications. These existing approaches are vision-based [3, 4, 5, 6, 7], sensor-based [8, 9], and motion-based [10, 11].

Vision-based systems face fundamental issues like the necessity to record targets, raising privacy concerns, and limiting real-world applications. Poor lighting affects image/video quality, and the presence of face masks during COVID-19 has made camera-based lip-reading nearly impossible. These systems also fail in complete darkness when visual observation of lip-movements is not feasible. Sensor-based systems have their own flaws, requiring targets to wear or carry devices, which can disrupt daily routines. In motion-based systems recently, RF based micro-movement detection research has shown encouraging results. These approaches use fluctuations in widely available RF signals to recognise movement. RF-based techniques, including Ultra-Wideband (UWB) [12], Wi-Fi [13], Bluetooth [14], and Radio Frequency Identification (RFID) [15, 16] are used to achieve precise detection of micro-movements. Wi-Fi-based, RFID-based, and radar-based micro-movement monitoring is a promising solution when compared to other RF-based approaches since it is a cost-effective due to the reuse of existing infrastructure. Now, there is a need for the development of accurate and reliable RF-based MM sensing systems for hearing impairment in the deaf community. This is because RF sensing offers three main advantages, which are outlined as follows:

1. **Non-Invasive and Unobtrusive Monitoring:** RF-based sensing, in contrast to wearable or camera-based systems, doesn't necessitate direct contact or a Line-of-Sight (LOS) connection with the subject. This characteristic allows for uninterrupted and unobtrusive monitoring, which is especially advantageous in sectors like healthcare, elderly care, and security. Additionally, ambient RF sensing, as opposed to wearable sensing technologies, lowers the risk of contact transmission infections by enabling contactless measurement of vital signs and macro-health indicators in Non-Line-of-Sight (NLOS) environments [17].

2. **Operational in Various Environmental Conditions:** RF-based systems are less sensitive to environmental variables like lighting conditions or weather in contrast to camera-based systems, which may encounter difficulties in poor lighting or when views are obstructed [18].

3. **Preservation of Anonymity:** RF-based sensing can track and analyse micro-movements and vital signs without capturing facial or other identifiable features, inherently preserving more anonymity than visual monitoring systems [19].

1.2 Problem Statement

The number of people with hearing impairments is expected to increase to 700 million by 2050 [1]. With this anticipated rise in the need for Hearing Aids (HAs), it is noteworthy that there are currently only three existing technologies for HAs, each with its own disadvantages, which are listed below.

1. **Wearable Sensors:** Wearable devices are electronic gadgets designed to be worn on the body, typically as accessories. They can track and monitor various personal metrics such as health and fitness data and provide a range of smart functionalities. However, these sensors, when attached to the body for continuous monitoring, may not always be comfortable for users to wear all the time, potentially disrupting their daily routines [9]. Additionally, wearable hearing devices often require frequent charging or battery replacements and regular maintenance, which can be inconvenient for users.

2. **Audio-Based:** Audio refers to the electronic representation, processing, or reproduction of sound, usually within the range of frequencies audible to the human ear. Audio-based devices designed for hearing impairments encounter challenges like ineffective noise filtration in loud environments and sound distortion. Unfortunately, in noisy environments, these devices often struggle, making it difficult to recognise individual voices [20].

3. **Vision-Based:** A camera is a device that captures and records images, either as still photographs or moving videos by allowing light to fall onto a light-sensitive surface, such as film or an electronic image sensor. Camera-based techniques face significant challenges, including privacy concerns from recording subjects and poor lighting that affects image quality. The widespread use of face masks during the COVID-19 era further reduces their effectiveness in lip-reading, and these systems become ineffective in complete darkness [21]. These unresolved issues continue to pose challenges for researchers.

1.3 Aims and Objectives

The purpose of this thesis is to create a non-invasive, contactless, MM sensing approach for hearing impairments. Every study carried out for this thesis received ethical approval from the College of Science and Engineering at the University of Glasgow. The purpose of MM aids is not limited to Speech Recognition (SR); they also support the interpretation of visual cues such as hand gestures, head movements, and facial expressions. The goals of this research include:

1. Review and evaluate the current systems for HAs and investigate the systems build for lip-reading, hand movements, head movements, and facial Expression.

2. Analyse the visual information collected from a range of sensors, including USRP, radar, and RFID.

3. Investigate Machine Learning (ML) and Deep Learning (DL) strategies for creating more generalized, low-latency, energy-efficient, and privacy-preserving hearing aid systems for deaf individuals.

4. Explore and integrate (with Ob3 models) ambitious wireless-based privacy-preserving MM Lip-Reading (LR) and end-user cognitive load (CL) use prediction models to deliver a personalised Audio-Video Speech Enhancement (AVSE) framework.

1.4 Contributions

This thesis proposes the effectiveness of MM sensing in applications related to hearing impairments. It covers various technologies, including communication-based Wi-Fi, UWB radar-based, Frequency Modulated Continuous Wave (FMCW) radar signals, and RFID-based signals. Ethical approval for conducting these experiments was obtained from the University of Glasgow's Research Ethics Committee (approval nos.: 300200232, 300190109). The significant contributions of this work are outlined as follows:

1. Investigate the accuracy and computational cost of a range of ML and DL algorithms in the context of MM sensing for hearing impairment.

2. Lip reading recognition using RF-sensing-based technology is designed to accurately detect lip movements whether a single subject is wearing a mask or not. This serves as a proof of concept for lip reading detection.

3. Hand gesture recognition system using RF sensing-based technology is proposed that is capable of accurately detect hand movements when performed by a single subject. This serves as a proof of concept for hand gesture detection.

4. Head gesture recognition using RF sensing based technology is proposed that is designed to accurately detect head movements whether a single subject performs them with or without a wall obstruction. This serves as a proof of concept for head gesture detection.

5. Facial recognition using RF sensing is proposed that is capable of accurately detecting facial expressions when performed by a single subject. This serves as a proof of concept for facial detection.

1.5 Thesis Organisation

This thesis is organised into the following chapters:

Chapter 2 list the current technologies and literature related to the field of lip reading, hand gesture recognition, head movements recognition, and facial expression.

Chapter 3 review examines how contactless sensing technologies can be utilised for lip-reading detection under a face mask, particularly in the COVID-19 era. It compares technologies such as Wi-Fi, Radar, and RFID sensing. The chapter concludes with findings that, while technologies like audio, video, and wearables are highly accurate, they do not function effectively under face masks and lack privacy preservation. In contrast, technologies like Radar, Wi-Fi, and RFID sensing can be implemented with privacy considerations, work effectively under face masks, and yield high-quality results.

Chapter 4 details the process of data collection using radar sensors from deaf individuals, and how DL techniques are applied for image classification of radar micro-doppler signatures. The chapter employs DL algorithms to analyse radar images, which depict the movements of British Sign Language (BSL). The findings indicate that, through a range of data processing techniques, DL can accurately distinguish between various movements captured in the radar micro-doppler signatures.

Chapter 5 discusses a novel approach that combines Wi-Fi and radar technologies, enhanced by DL techniques, to detect head movements through walls. The potential of integrating these two sensing modalities to overcome the limitations of each technology when used separately. By leveraging the strengths of both Wi-Fi and radar, along with the sophisticated pattern recognition capabilities of DL algorithms, this chapter demonstrates improved accuracy and reliability in sensing head movements, even through physical barriers like walls. This advancement could have significant implications for various applications, including security, health monitoring, and smart home systems.

Chapter 6 develops an RF sensing-based system for facial expression recognition. The system utilises frequency FMCW radar combined with ML techniques to classify facial expressions. The study specifically focuses on five common facial expressions: Happy, Sad, Fear, Surprise, and Neutral. The data, recorded as micro-doppler signals, are processed using state-of-the-art ML models such as Super Learner (SL), Linear Discriminant Analysis (LDA), Random Forest (RF), K Nearest Neighbor (KNN), Long Short-Term Memory (LSTM), and Logistic Regression (LR) to extract relevant features. These features, derived from the radar data, are then input into ML models for classification. The results demonstrate highly promising accuracy in facial expression classification.

Chapter 7 concludes the thesis and details future work to be considered for expanding the work discussed throughout the thesis.

Chapter 2

Literature Review

This chapter explores all contact and contactless techniques for hearing impairments that support lip reading, hand recognition, head movements, and facial recognition. The World Health Organisation (WHO) reports that in the UK alone, about 11 million people live with hearing impairments [1]. Therefore, the need for MM sensing techniques in this area is a critical aspect of healthcare research [22, 23, 24, 25]. MM hearing impairment technology is essential for supporting the deaf community in communicating with others across various environments. Future healthcare systems are looking to integrate such technologies to surpass the limitations of current methods like wearables, cameras, and audio. This chapter also discusses how MM hearing impairment devices enable communication through both verbal and non-verbal cues. It introduces SR technologies, including contact and contactless methods, in Section 2.1. A review of BSL in state-of-the-art technologies is presented in Section 2.2. Section 2.3 covers the application of hearing aid devices in identifying human behavior. Section 2.4 discusses current technologies that assist in the identification of facial expressions. Moreover, this chapter includes an overview of ML and DL techniques for classifying data collected by these devices, detailed in Section 2.5. Finally, the findings of this chapter are summarised in Section 2.6.

2.1 Speech Recognition

This section investigates how lip-reading can be detected through both contact and non-contact methods with the lips, as shown in Table 2.1.

Table 2.1: Summary of Speech Recognition Technologies: Sensor-, Audio-, Camera-, Radar-, Wi-Fi-, and RFID

References	Technology Used	Application	Key Outcomes	Limitation
Sensor-Based Speech Recognition				

Continued on next page

Table 2.1 – continued from previous page

References	Technology Used	Application	Key Outcomes	Limitation
Xu et al. [26]	Zwitterionic hydrogel sensor	Silent-SR via throat-worn device	5x sensitivity of nonionic hydrogels, 38ms response, mimics human skin	-
Kim et al. [27]	Ultrathin silicon strain gauges with DL	Silent speech interfaces	Detects minor physical changes, facilitates non-verbal communication	Sensitivity to external factors
Dong et al. [28]	EMG technology with less intrusive dry electrodes	Lip-reading for SR	Captures electrical signals from lip movements, interprets signals for lip-reading	Limited by dry electrode reliability
Lu et al. [29]	Flexible triboelectric sensors	Decoding lip movements	Positioned inside a pseudo mask for clear visibility, identifies lip movements	Mask design may not fit all users
Audio-Based Speech Recognition				
Tsouvalas et al. [30]	FedSTAR in Federated Learning (FL)	Audio recognition on smartphones	Enhances audio recognition with minimal labeled data; notable gains in federated settings	Dependency on data distribution
Adeel et al. [31]	Acoustic modeling	Speech quality and intelligibility enhancement	Tested in real-world scenarios for improved speech quality and intelligibility	Requires clear visual input

Continued on next page

Table 2.1 – continued from previous page

References	Technology Used	Application	Key Outcomes	Limitation
Wu et al. [32]	Two-layer LSTM with pruning for SR	Process automation via human voice	Converts spoken words to text, with applications in process automation	Pruning may remove relevant features
Johnson et al. [33]	Voice recognition systems	Healthcare voice data conversion	Converts sound to digital signals and data to text for healthcare applications	Susceptibility to background noise
Camera-Based Speech Recognition				
Ma et al. [34]	Video Speech Recognition	Improving VSR accuracy	Enhanced prediction and optimisation outperform larger datasets; effective across multiple languages	-
Various [35, 36, 37]	Video Speech Recognition	Communication for the deaf, biometric authentication, AVSE	Gained notable research attention for various applications	Challenges in complex scenarios
Kamil et al. [38]	Vision system for lip-reading	Aiding pronunciation for the hearing-impaired	Tracks lip movements to distinguish phonemes without an instructor	Limited by visual clarity and user's ability to mimic
Kastaniotis et al. [39]	CNN-TCN model for lip-reading	Predicting Greek phrases using mobile phones	Surpasses CNN-LSTM methods, enhancing stability and efficiency	Requires high-quality video input

Continued on next page

Table 2.1 – continued from previous page

References	Technology Used	Application	Key Outcomes	Limitation
Nemani et al. [40]	Custom 3-D CNN for VSR	Analysing spoken words from video	Strong performance across conditions; 80.2% and 77.9% accuracies	Dataset specificity and generalisation challenges
Radar-Based Speech Recognition				
Yue Ma et al. [41]	Auditory radar and webcam	Speech reconstruction	Accurate speech reconstruction for the character "A" with portable radar and webcam	Limited to single character analysis
Hameed et al. [22]	Radar and Wi-Fi sensing, ML/DL algorithms	Vowel classification	High accuracy (91.67%) in vowel classification without mask	Specific to vowels, may not generalise to all speech
Ge et al. [42]	Multimodal dataset with radars, camera, and sensors	SR research	Introduction of RVTALL dataset for diverse SR applications	Dataset complexity and integration challenges
Wi-Fi-Based Speech Recognition				
Hameed et al. [22]	Radar and Wi-Fi RF sensing with ML/DL	Vowel classification	High classification accuracy of 95% with Wi-Fi data for vowel recognition	Specific to vowels, may not generalise to all speech
Wang et al. [43]	CSI-based recognition system	Pronunciation classification	91% accuracy for single user, 74% for three users with WiHear system	Decreased accuracy with multiple speakers
RFID-Based Speech Recognition				

Continued on next page

Table 2.1 – continued from previous page

References	Technology Used	Application	Key Outcomes	Limitation
Zhang et al. [44]	Commercial RFID technology with multiple tags embedded on a transparent sheet	Words detection	95% detection accuracy for user speech; vocabulary recognition of 20 words with an average classification accuracy of 88%	Limited vocabulary size; potential challenges in expanding the system to recognise a broader vocabulary
Wang et al. [45]	RFID tattoos attached around the user's face	Speech recognition	86% accuracy rate among 10 users	Discomfort from wearing tattoos; need for multiple tags for a single word increases costs; re-calibration difficulty; limited communication range of 2.5 meters; concerns about cost efficiency and user convenience

2.2 Sign Language Recognition

This section examines how Sign Language can be recognised through methods that involve both contact with and non-contact with the hand, as shown in Table 2.2.

Table 2.2: Summary of Sign Language Recognition Technologies: Sensor, Camera-, Radar-, Wi-Fi-, and RFID

Reference	Technology Used	Application	Key Outcomes	Limitation
Sensor-Based Sign Language Recognition				
Gu et al. [46]	Smart wearable system with IMUs on fingertips and back of the hand	Interpreting ASL	Average accuracy of 99.81% in identifying ASL gestures	-
Fan et al. [47]	Smart data glove with sensors and an IMU	Recognising various hand gestures in land and underwater environments	High accuracy rates (over 98%), adaptive gesture recognition model, 94% accuracy for new users/devices	-
Harish et al. [48]	Wearable glove with flex sensors and accelerometers	Human Machine Interface for ISL	Improved accuracy from 74.12% to 97.2% with accelerometers	Initial lower accuracy rate without accelerometers
Preetham et al. [49]	Data glove with 10 flex sensors	Hand gesture recognition	-	Limited to single-hand gestures
Faisal et al. [50]	Dataglove with sensors, microcontroller	ASL recognition	Over 82% accuracy for static and 97% for dynamic gestures	-
Pan et al. [51]	Wearable system with bimodal capacitive sensors, 5G technology	Detecting finger movements and hand location	Over 99% accuracy in static and 91% in dynamic gesture recognition	-

Continued on next page

Table 2.2 – continued from previous page

Reference	Technology Used	Application	Key Outcomes	Limitation
Wang et al. [52]	Miniature inertial sensors with ML algorithms	Hand gesture recognition	High accuracy with random forest for static gestures and enhanced LSTM for dynamic gestures (up to 98.3%)	-
Camera-Based Sign Language Recognition				
Jadooki et al. [53]	Kinect sensor systems	Recognition of static signs of SL	Fusion of depth data and color information	-
Bauer et al. [54]	Colored gloves, HMM	Automatic recognition of SL	Data classified using the KNN algorithm	-
Mohandes et al. [55]	2D video cameras	SL recognition	Cameras widely used for SLR	-
Pigou et al. [56]	Deep end-to-end neural network	Gesture identification in video	Enhances frame-wise gesture recognition	-
Neverova et al. [57]	CNN architecture	Integrates data from multiple channels	Learns from grayscale video, depth, and skeletal joints	-
D et al. [58]	HOG, ANN	BSL classification	Utilised HOG for image examination	-
Aditya et al. [59]	Attentive multi-feature network	CSLR in video streams	Introduces extra keypoint features and attention layers	Struggles with limited information during training
Alyami et al. [60]	2DCRNN, 3DCNN	Classifying Arabic sign language	High accuracy levels (92% with 2DCRNN and 99% with 3DCNN)	-

Continued on next page

Table 2.2 – continued from previous page

Reference	**Technology Used**	**Application**	**Key Outcomes**	**Limitation**
Radar-Based Sign Language Recognition				
McCleary et al. [61]	CW radar and DL	BSL gesture recognition	92.81% accuracy in recognising 36 finger movements	-
Rahman et al. [62]	FMCW Radar, GAN	Word-level ASL recognition	-	-
Gurbuz et al. [63]	RF sensor and CNN	ASL recognition	Real-time sign recognition at 77 GHz	-
Wang et al.(2023) [64]	MV-DNN, millimeter-wave radar	CSL recognition	96% accuracy for eight CSLs	-
Li et al. [65]	UWB radar	SL and hand gesture recognition	Shows improved accuracy	-
Gavin et al. [66]	Millimeter-wave radar, CNN	ASL recognition	Effective in scenarios with overlapping movements	-
Wi-Fi Based Sign Language Recognition				
Shang et al. [67]	Wi-Fi CSI, SVM	SL classification	Higher classification accuarcy of SL	-
Ji et al. [68]	Wi-Fi devices, Neural Network	Constructing hand skeletons	Generates 2D and 3D hand models, enables finger tracking and SLR	-
Lin et al. [69]	Smartphone, Wi-Fi router, ML models	Human activity recognition	97.25% accuracy in classifying 20 types of human activities	-
Gao et al. [70]	Wi-Fi CSI, CNN, KNN	SLR	High accuracies on the SignFi dataset	-

Continued on next page

Table 2.2 – continued from previous page

Reference	Technology Used	Application	Key Outcomes	Limitation
Zhang et al. [71]	Wi-Fi signals, PCA, Residual-MultiHead model	SLR for phrases	95.03% recognition accuracy for English phrases translation	-
RFID-Based Sign Language Recognition				
Xu et al. [72]	Passive RFID tag	CSL recognition	Average F1-scores of 96.67% for new users and 97.50% for new environments	-
Zou et al. [73]	COTS RFID readers	Gesture recognition	High accuracy of 96.5% in fixed-position and 92.8% in diverse-position scenarios	-
Ma et al. [74]	RFID technology and Siamese network	Gesture recognition	Achieves an accuracy of 0.93 with a single sample per gesture	-
Zhao et al. [75]	RFID technology	Human action recognition	Developed a system for spatiotemporal data analysis	-
Dian et al. [76]	RFID with MCNN	Gesture recognition	Significantly outperforms existing RFID solutions	-

2.3 Head Movements Recognition

This section discusses techniques for recognising head gestures that both require and do not require contact with the head, as shown in Table 2.3.

Table 2.3: Summary of Head Movements Recognition Technologies: Sensor, Camera-, Radar-, Wi-Fi-, and RFID

Reference	Technology Used	Application	Key Outcomes	Limitation
Sensor-Based Head Movements Recognition				
Liu et al. [77]	Smart pillow with strain-sensing threads	Tracking head movements	Successful real-time tracking for flexible bioelectronics	-
Gonzalez et al. [78]	Wheelchair operated by head movements	Mobility impairments assistance	"Very good" SUS rating; response times detailed	Head motion errors in variable speed
Lee et al. [79]	Infrared Sensors	Minimising car accidents	78% success in detecting drowsy driving	-
Jiang et al. [80]	Human-machine interface with flexible strain-sensing threads	Accurate head motion tracking	92% accuracy in predicting head orientations	-
Camera-Based Head Movements Recognition				
Al et al. [81]	Eye-gaze and head movement tracking	Human-computer interaction	Comprehensive methods overview	-
Horprasert et al. [82]	Facial symmetry and anthropometric measurements	Determining head orientation	Empirical validation with photographs	-
Neto et al. [83]	Real-time head movement estimation via video camera	Communication interface	Tuned computer-vision algorithms	Environment specificity
Arcoverde et al. [84]	API for mobile phones with computer vision	Real-time head pose estimation	Robust across diverse conditions	-
Merrouche et al. [85]	Vision-based fall detection with a depth camera	Human shape analysis, head tracking	93.25% accuracy on SDUFall dataset	-

Continued on next page

Table 2.3 – continued from previous page

Reference	Technology Used	Application	Key Outcomes	Limitation
Wang et al. [86]	Head pose tracking with gaze estimation	Driver's gaze zone estimation	Effective gaze direction prediction	-
Radar-Based Head Movements Recognition				
Raja et al. [87]	Millimeter-wave doppler radar	3D head tracking in vehicles	92% movement-prediction accuracy	-
Ding et al. [88]	FMCW radar	Detecting inattentive driving behaviors	Average accuracy of about 95%	-
Bresnahan et al. [89]	Millimeter-wave FMCW radar	Classifying driver head movements	High classification accuracy in both stationary and moving cars	-
Nguyen et al. [90]	Millimeter-wave radar with One-shot learning	Monitoring head movements	High accuracy of 100%	-
Bu et al. [91]	LFMCW radar with multidomain fusion network	Human head movement recognition	Enhanced accuracy using a multidomain approach	-
Sun et al. [92]	Automotive-Radars in driving-assistance systems	Advantages and challenges discussion	Higher angular resolutions with fewer antennas	Hardware and size constraints
Wi-Fi Based Head Movements Recognition				
Shang et al. [67]	Kernel-based SVM	Classification of SL using CSI patterns	Utilised for SL classification	-
Ji et al. [68]	Neural network with Wi-Fi devices	Constructing hand skeletons for SLR	Generates 2D and 3D hand models; surpasses previous methods	-
Lin et al. [69]	ML models with Wi-Fi CSI	Human activity recognition	97.25% accuracy in classifying human activities	-

Continued on next page

Table 2.3 – continued from previous page

Reference	**Technology Used**	**Application**	**Key Outcomes**	**Limitation**
Gao et al. [70]	CNN and KNN with Wi-Fi CSI	Wi-Fi-based SLR	High accuracies on the SignFi dataset	-
Zhang et al. [71]	PCA and Residual-MultiHead model with Wi-Fi signals	SLR for phrases	95.03% recognition accuracy for English phrase translation	-
RFID-Based Head Movements Recognition				
Chen et al. [93]	RFID for passive sensing	Head gesture recognition	91% accuracy in recognising head gestures	Interference from other body movements
He et al. [94]	RFID with cross circular polarisation	Non-contact human activity detection	Improved SNR and sensing range, 230% increase in detection area	-
Figueiredo et al. [95]	RFID, antennas, IMU	Tracking head orientation	Accurate prediction of Euler angles, minimal error	Challenges with noise and sampling
Yang et al. [96]	RFID tag phase responses for Nod-Track	Driving fatigue detection	High accuracy in detecting nodding motion	-

2.4 Facial Recognition

This section discusses techniques for recognising facial expressions that include both contact and non-contact with the face, as shown in Table 2.4.

Table 2.4: Summary of Facial Recognition Technologies: Sensor, Camera-, Radar-, Wi-Fi-, and RFID

Reference	**Technology Used**	**Application**	**Key Outcomes**	**Limitation**
Sensor-Based Facial Expression Recognition				

Continued on next page

Table 2.4 Continued from previous page

Reference	Technology Used	Application	Key Outcomes	Limitation
Yan et al. [97]	EmoGlass: emotion-detecting glasses	Enhancing emotional awareness	Recognizes seven facial expressions; valuable insights for emotional health tech	-
Matthies et al. [98]	Wearable sensing methods survey	Understanding emotions for HCI	Reviews technologies for emotion recognition	-
Verma et al. [99]	Inertial sensors	Facial expression recognition	89.9% average accuracy in detecting expressions	-
Masai et al. [100]	Photo Reflective Sensors	Recognizing facial expressions during daily activities	Identifies eight expressions; resembles regular glasses	-
Matthies et al. [101]	EarFieldSensing (EarFS)	Detecting facial expressions via electric field changes	High accuracy in facial gesture recognition	-
Camera-Based Facial Expression Recognition				
O et al. [102]	Cameras with computer vision and DL	Facial expression recognition	Effective face expression classification	-
Wu [103]	Edge computing with mobile network design	Real-time facial expression recognition on mobile devices	Efficient and real-time recognition	-

Continued on next page

Table 2.4 Continued from previous page

Reference	Technology Used	Application	Key Outcomes	Limitation
Komagal [104]	PTZ cameras with Faster R-CNN	Facial expression analysis for educational technology	Enhances student engagement recognition	-
Siddiqi et al. [105]	Depth camera with PCA and ICA	Facial Expression Recognition (FER) system	98.0% accuracy rate	-
Suk et al. [106]	Smartphone app with SVMs	Real-time facial expression recognition	86% accuracy on standard dataset; 72% in real-world testing	Limited frame rate
Terissi et al. [107]	Single camera with 3D face model	Tracking head pose and facial expressions	Effective tracking without specific training phase	Challenges with occlusions and varying distances
Radar-Based Facial Expression Recognition				
Mostafa et al. [108]	IF waveform generation and acquisition circuits	Radar system applications	Suitable for radar system applications	-
Shah et al. [109]	RF sensing technologies (Doppler radar, RFID, Wi-Fi)	Monitoring health parameters	Explores various RF-based sensing techniques	-
Dang et al. [110]	Millimeter-wave radar with DL model	Emotion recognition	High accuracy in emotion recognition	-
Zhang et al. [111]	CW radar and camera technology	Non-contact emotion recognition	High accuracy in emotion recognition	Interference reduction required
Gouveia et al. [112]	Non-contact radar for emotion recognition	Mental health care applications	Accuracy rates between 60% and 70%	-

Continued on next page

Table 2.4 Continued from previous page

Reference	Technology Used	Application	Key Outcomes	Limitation
Dang et al. [113]	Millimeter-wave radar with DL	Classifying emotions	High recognition accuracy	-
Wi-Fi Based Facial Expression Recognition				
Chen et al. [114]	Wi-Fi signal analysis (WiFace)	Facial expression recognition	94.80% accuracy for six expressions	-
Gu et al. [115]	Wi-Fi and vision-based analysis	Hybrid emotion recognition	79.05% accuracy for seven emotions	-
Gu et al. [116]	Wi-Fi signal processing (EmoSense)	Emotion sensing	Unobtrusive detection	-
Khan et al. [117]	Wireless signals with DL	Emotion state recognition	71.67% classification accuracy	-
Jia et al. [118]	RF system with Wi-Fi (BeAware)	Behaviour recognition	Effective user behavior recognition	-
RFID-Based Facial Expression Recognition				
Ramli et al. [119]	Stretchable strain sensor	Facial expression detection	High sensitivity; can stretch up to 20% strain	-
Xu et al. [120]	RFID for facial authentication	Enhancing privacy and anti-spoofing	Over 95.7% success in authentication; EER of 4.4%	-
Battaglia et al. [121]	RFID with face recognition and pattern matching	Authentication solution	Addresses computational complexity and data storage issues	-
Luo et al. [122]	RFaceID: RFID-based face recognition	Facial recognition	93.1% recognition accuracy	-

2.5 Machine and Deep Learning

ML involves algorithms that enable computers to identify patterns and make decisions or predictions based on data [123]. ML and DL are widely used in various applications, including self-driving cars, SR, and etc [124, 125]. ML encompasses diverse algorithms designed to address different problems [126]. These algorithms create models using training data, which provide sample patterns for the algorithms to recognise in future, unseen data. By learning from these patterns, ML algorithms can make predictions or decisions based on past data examples. For instance, when a model is fed sensor information about specific movements, ML algorithms can detect patterns of these movements in new, unseen samples. DL algorithms, a subset of ML, utilise neural networks that mimic the functioning of the human brain.

2.5.1 Advantages of Machine Learning

ML has the advantage of requiring no specific programming to recognise patterns. Instead, ML algorithms learn from sample data, known as training data, allowing them to find patterns on their own. This method saves time and uncovers patterns in data that people may find difficult to discern[127, 128].

2.5.2 Machine Learning Approaches

ML algorithms are broadly categorised into three groups: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves algorithms learning from labeled training data to recognise patterns and subsequently predict or classify data based on these learned patterns [129]. Unsupervised learning, on the other hand, works with unlabeled data, with computers clustering data into clusters based on identified patterns [130]. Reinforcement learning is a reward-based system in which the algorithm learns to attain a specific reward through trial and error. This sort of learning is used in a variety of applications, including as self-driving automobiles and strategic games such as chess [131].

2.5.3 Advantages of Deep Learning

DL provides flexibility across applications such as image identification and natural language processing. It specialises in processing huge and complicated datasets and automatically extracting crucial features. It is an essential tool because of its robustness, adaptability, and ability to deal with unstructured data and make predictions in real-time. Enhancing its efficiency and usefulness is its automated decision-making and transfer learning capabilities [132, 133, 134].

2.5.4 Deep Learning Approaches

DL is the subset of ML, that utilises artificial neural networks similar to those in the human brain [135]. These networks consist of interconnected neurons [136]. DL architectures include an input layer for training the data, several hidden layers, and an output layer. The deep in DL refers to the number of hidden layers, which can be quite extensive [137]. These layers process input data by multiplying it with assigned weights and producing outputs [138]. The final output is compared to the input layer at the output layer, resulting in weight recalibration for better alignment with the input data. As the input is passed through the layers the final output is represented in the output layer. The output layer is then compared to the input layer and weights are recalculated to be more in line with the input data [139]. The weight values are then readjusted which results in small changes in the network's perception of patterns in the data to increase accuracy [140]. The input data is then fed through the network with the updated weights. Backpropagation algorithms are used to adjust the values of the weights based on the results of the previous run [141]. This process is repeated many times until optimal results are obtained. The number of times the data is fed through the neural network is known as epochs [142].

2.6 Summary

In this chapter, we have discussed various technologies that support future hearing impairment applications designed to assist the deaf in communicating with hearing individuals through both verbal and non-verbal modes. The concept of future MM hearing aids encompasses speech recognition (SR), sign language recognition (SLR), head movement recognition, and facial recognition to identify others' behaviors. However, existing technologies such as wearables, audio systems, and cameras face certain limitations in supporting future hearing aid devices. Table 2.5 outlines the advantages and limitations of current and prospective technologies that support hearing aid devices. The work of this thesis seeks to fill this gap in the literature. The next chapter describes the important aspects of future MM hearing aid devices that use RF-based SR.

The work of this thesis seeks to fill this gap in the literature. The next chapter describes the important aspects of future MM hearing aid devices that use RF-based SR.

Table 2.5: Technologies supporting future hearing aid device

Technology Type	**Advantages**	**Limitations**
Wearable-Based Devices	Collect and monitor various personal metrics, such as health and fitness data, and provide smart features.	Continuous attachment can cause discomfort, limited battery life necessitating frequent charging.
Audio-Based Devices	Allows communication in verbal modes.	Insufficient noise filtering, sound distortion in noisy environments, difficulty distinguishing voices.
Camera-based Devices	Support non-verbal communication modes like facial recognition and lip-reading.	Privacy concerns, degraded image quality in poor lighting, ineffective with face masks and in total darkness.
RF Sensing (Radar, Wi-Fi, RFID)	Can penetrate masks easily, leverage existing communication signals and infrastructure, not dependent on light, offers solutions to assisted living and contactless monitoring.	Gaps in literature on RF sensing for MM hearing aids, the need for a Tx and Rx chip addition to the hearing aid.

Chapter 3

Lip Reading Under Face Mask Using Contactless Sensing

Lip-reading has emerged as a vital research challenge with the objective of recognising speech through lip movements. However, most existing lip-reading technologies, such as Sensor-based, Audio-based, and Vision-based systems, encounter substantial limitations. Sensor-based devices often cause discomfort in daily usage, Audio-based methods face challenges in noisy environments, and Vision-based techniques have become almost impractical during the COVID-19 era due to mask usage. This chapter introduces three innovative approaches for MM hearing impairment based on speech recognition, emphasising non-invasive methods to detect speech in the presence of face masks such as radar, Wi-Fi, and RFID. It also highlights potential future research directions essential for developing advanced hearing impairment technologies adapted to the unique challenges posed by the COVID-19 pandemic.

3.1 Introduction

Hearing abilities are often measured by a person's capacity to understand noises at decibel levels (dB) of 20 and above. When people fail to distinguish sounds at this level or higher, it is a sign of a problem with hearing, from small to serious [1]. Those with such disorders frequently encounter considerable difficulty in ordinary communication and learning processes and are commonly termed as 'hard of hearing' or 'deaf'. Globally, approximately 5% of the population, or around 430 million individuals, currently experience hearing impairments. According to predictions, this amount will rise to 700 million by 2050. Around 11 million people in the United Kingdom have hearing impairments, with age-related hearing loss emerging as a significant issue [2]. Future developments in hearing aids are expected to significantly evolve, emphasizing MM processing that extends beyond the traditional scope of speech and sound enhancement.

Existing hearing impairment technologies have several drawbacks. Sensor-based systems require the user to wear or carry devices, which can be uncomfortable for prolonged periods and

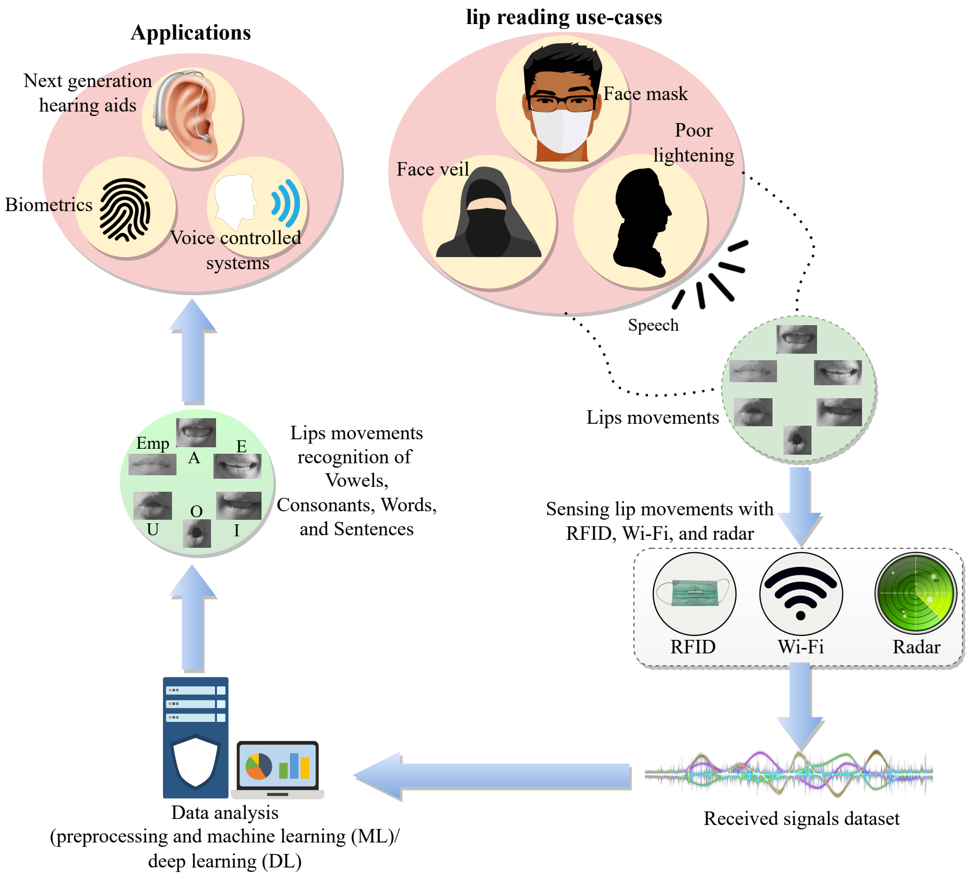

Applications
Next generation hearing aids
Biometrics
Voice controlled systems
lip reading use-cases
Face mask
Poor lightening
Face veil
Speech
Lips movements
Emp
A
E
U
O
I
Lips movements recognition of Vowels, Consonants, Words, and Sentences
Sensing lip movements with RFID, Wi-Fi, and radar
RFID
Wi-Fi
Radar
Received signals dataset
Data analysis (preprocessing and machine learning (ML)/ deep learning (DL)

Figure 3.1: Conceptual illustration of the proposed lip-reading framework.

disrupt daily activities. These devices often need frequent charging or battery replacement, and require regular maintenance, adding to the inconvenience. Audio-based devices, while designed for hearing impairments, struggle with noise filtration in loud environments and can distort sounds. This makes it challenging to recognise individual voices in noisy environments. Vision-based systems, which use visual information like lip-reading for speech recognition, face privacy concerns. The use of cameras in hearing aids could be seen as recording without consent, which is legally problematic in many regions. Moreover, the widespread use of face masks during the COVID-19 pandemic has further limited the effectiveness of vision-based hearing aids.

The limitations of existing hearing aid devices indicate that RF sensing is the only viable solution for MM hearing aid devices. The need for advancement of next-generation MM hearing aids could be propelled by RF sensing technology. This approach, adept at detecting lip and mouth movements, provides precise cues for hearing aids by identifying spoken sounds and speech patterns through ML and DL techniques. A notable benefit of RF sensing is its efficacy even with face masks on, as RF signals can penetrate masks to capture essential visual cues, overcoming a major limitation of existing hearing aid devices. This chapter describes the design, development, and demonstration of a functioning RF sensing-based system for the detection of spoken sounds via face masks. The suggested RF sensing device can function both on its own and in combination with the hearing aids, helping to identify lip and mouth movements that are hidden by face masks, which often obstruct visual cues in vision-based hearing aid systems.

A conceptual illustration of the proposed lip-reading framework is presented in Figure. 3.1. This framework detects variations in wireless CSI amplitudes, resulting from lip and mouth movements, through the use of ML and DL algorithms. These variations are then categorised into distinct forms of speech, such as words, phonemes, or letters. In particular, the radar-based system within this framework uses Doppler shift spectrograms, which a DL model identifies to classify various lip movements. Additionally, the framework incorporates a passive RFID tag, commonly found in Ultra high frequency (UHF) Textile Laundry products, which is integrated into a standard mask for data collection. This design ensures comfort and eliminates discomfort for the wearer. The gathered data, expressed in RSSI values, are then analysed with different ML models. The applications of this RF-based lip-reading framework are diverse, extending to fields such as hearing aid enhancement, biometric security, and voice-activated controls in smart home and vehicle infotainment systems.

3.2 Radar and Wi-Fi Speech Recognition

3.2.1 Methodology

The methodology used in this chapter has five main steps. Firstly, the experimental setup of radar and Wi-Fi is described, followed by data collection as the second step. The third step details the pre-processing phases. In the fourth step, the parameter settings of the considered algorithm are described. Finally, the evaluation metrics of the classification model are outlined. The subsequent subsections provide a detailed discussion of each stage in the proposed methodology.

Radar Based Experimental Setup

The hardware setup of radar-based lip-reading system is shown in Figure. 3.2, where Figure. 3.2a shows the front view and Figure. 3.2b represents the top view. Correspondingly, the front and top views of Wi-Fi-based setup are shown in Figure. 3.2c and Figure. 3.2d, respectively. For radar-based setup, Xethru X4M03 an UWB was used in this experiment, which was placed on top of the screen of the laptop. The Xethru X4M03 is a UWB radar sensor with built-in transmitter (Tx) and receiver (Rx) antennas, providing a maximum detection range of 9.6 metres. Key parameter settings of the radar are indicated in Table. 3.1. The subject was sitting 0.45 metres away from the radar while pronouncing vowels as illustrated in Figure. 3.2b. The body was in normal position and the only movements were the lip movements along with slight head movements, which are common while talking. The duration of each activity was set to 6 seconds, where an activity represents the data collection of a single vowel from a single subject. The RF signal was transmitted and received from the radar within this duration. The UWB radar-based system setup for lip-reading data collection and processing is illustrated in Figure. 3.4a. The details of all components presented in the figure are discussed later in this section. The features utilised for the radar are obtained from the STFT of the radar signal which provide the spectrograms of radar doppler shift due to lip and mouth movements. The analysis of the spectrograms showed that different vowels resulted in different spectrograms due to the differences in lip and mouth movements. To classify vowels, pre-trained Visual Geometry Group (VGG) models was utilised due to their better performance on abstract images like spectrograms [143, 144].

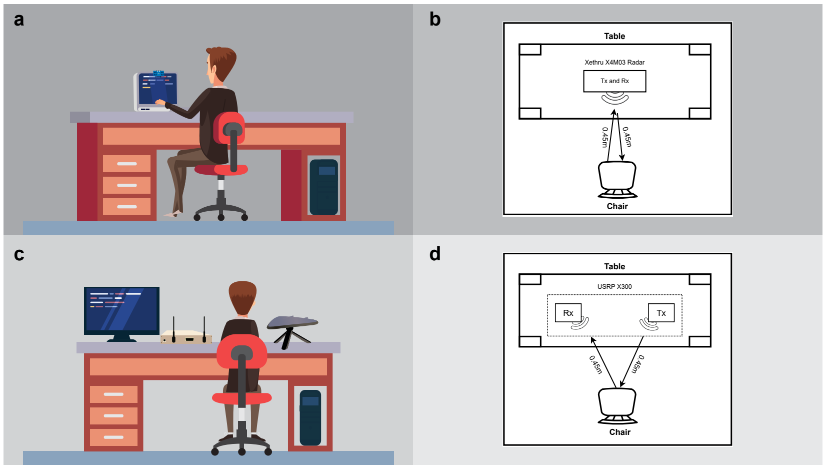


Figure 3.2: Experimental setup of the data collection through radar and Wi-Fi. (a) Front view of the data collection setup using Xethru UWB radar. (b) Top view of the radar based data collection. (c) Front view of Wi-Fi based data collection. (d) Top view of the Wi-Fi based data collection setup.

Parameter	Value
Platform	Xetru radar X4MO3
Instrumental range	9.6 meters
Target's distance from radar	0.45 meters
Operating frequency	7.29GHz
Transmitter power	6.3dBm
Activity duration	6 seconds
Collected samples in each class	50

Table 3.1: Configuration parameters of radar software and hardware.

Wi-Fi Based Experimental Setup

For the second set of experiments, Wi-Fi was used as a lip movement recognition platform. For this, a Universal Software Radio Peripheral (USRP) X300 was used, equipped with one directional antenna as a Tx and two omnidirectional antennas as Rx as shown in Figure. 3.2c. For experiments, monopole antennas, VERT2450, optimised at 2.45GHz frequency band, were used as Rx. A log-periodic antenna, HyperLOG 7040 X BPA was used as a Tx. Both Tx and Rx antenna gains were set to 35dB. The USRP was connected with a desktop having an Intel(R) Core (TM) i7-7700 3.60GHz processors with a 16GB RAM. Key parameter settings of the Wi-Fi based setup are indicated in Table. 3.2. GNU's Not Unix (GNU) radio was used to communicate with the USRP with the help of a virtual machine having Ubuntu 16.04 operating system. A python script was developed to send and receive data from USRP X300. The experiments were conducted at an operational frequency of Wi-Fi in 2.45GHz band. Both the Tx and Rx antennas were placed around 0.45 metres from the target as illustrated in Figure. 3.2d. Each activity was performed for 6 seconds. It is worth mentioning that Wi-Fi signals were tested with different features including time-frequency maps, etc. However, the CSI values of Wi-Fi signals performed best with variations in CSI amplitudes unlike the radar signals, where frequency shift was a major differentiating factor. The variations in one dimensional CSI amplitude showed clear patterns which could be attributed to a spoken vowel.

Parameter	Value
USRP Platform	X300
OFDM subcarriers	51
Operating frequency	2.45GHz
Transmitter Gain	35dB
Receiver gain	35dB
TX Antenna	Log periodic HyperLOG 7040, 700MHz to 4GHz
Rx Antenna	Monopole VERT2450, 2.45GHz
Target's distance from Tx and Rx antennas	0.45 meters
Activity duration	6 seconds
Collected samples in each class	50

Table 3.2: Configuration parameters of USRP software and hardware.

Data Collection

The conducted experiments were performed with two different technologies, *i.e.,* Wi-Fi and radar. Five vowels, A, E, I, O, and U were collected along with an empty letter, where subjects were not talking at all, and the lips were in normal closed position. An illustration of the lip movements to speak out all classes is shown in Figure. 3.3a, while the corresponding CSI samples and spectrograms are shown in Figure. 3.3b and Figure. 3.3c, respectively. The following section describes the experimental hardware setup used to collect data using both technologies.

For both experiments (radar and Wi-Fi), three participants, one male and two females, participated in the data collection process. The reason to include more participants was to make the dataset more realistic and diverse. A total of 3600 data samples were collected during both experiment for six classes, namely, A, E, I, O, U, and Emp, where Emp represents the lip posture of being silent. In each experiment, a total of 1800 data samples were collected from three

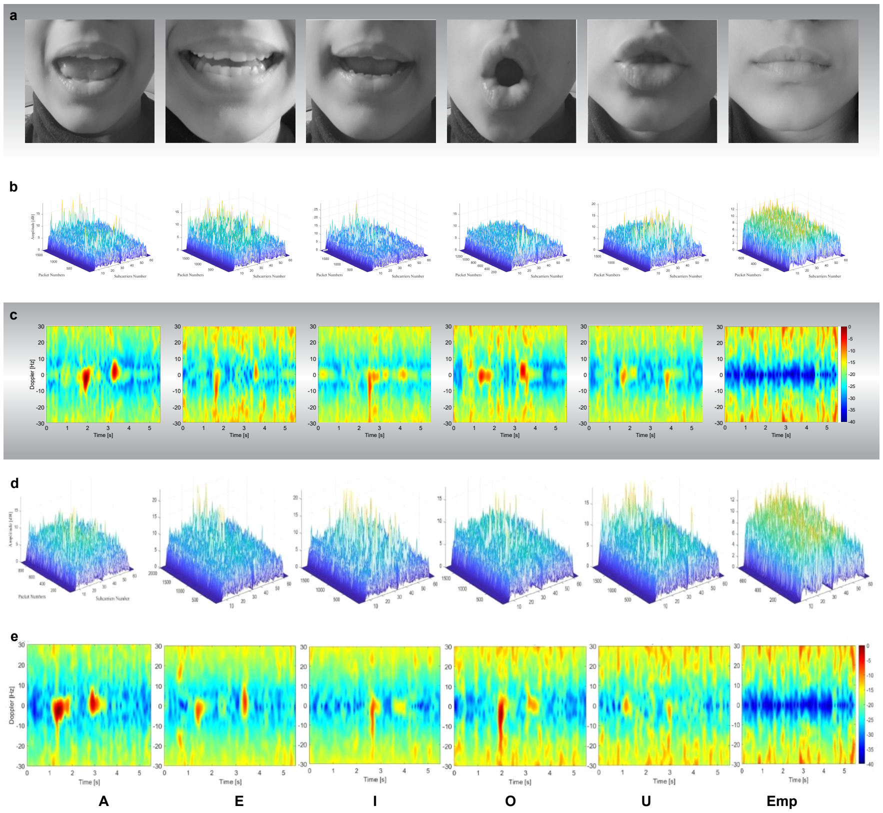


Figure 3.3: Pronounced vowels with their representation in Wi-Fi and radar signal. (a) A visual illustration of the pronounced vowels. (b) Wi-Fi data samples with mask representing various vowel classes. (c) Radar data samples with mask representing various vowel classes. (d) Wi-Fi data samples without mask representing various vowel classes. (e) Radar data samples without mask representing various vowel classes.

participants, 900 with face mask and 900 without face mask, where 50 samples were collected in each class. In particular, each participant repeated the speaking activity of each vowel 50 times with mask and 50 times without mask with the radar. Similarly, the same amount of data was collected from USRP with the same strategy. In this way, each participant contributed to collect 1200 data samples in total for six classes, two scenarios (with mask and without mask) and two technologies (radar and Wi-Fi). The ethical approval to conduct these experiments was obtained by the University of Glasgow's Research Ethics Committee (approval no.: 300200232, 300190109).

In the case of Wi-Fi, each instance of the data represents the CSI amplitudes, where 2000 packets were transmitted in a duration of six seconds. Figure. 3.3b illustrates the CSI patterns (amplitude) of considered Lip movements, *i.e.*, A, E, I, O, U and empty, in the case of face mask. The CSI patterns in the case of without face mask are illustrated in Figure. 3.3d Different colours in each figure represent the 51 subcarriers of the OFDM signal. Y-axis of each sub-figure represents the amplitude of the subcarriers while number of received packets are displayed on x-axis. The same data collection strategy was applied in radar, where a total number 1800 data samples were collected for three subject male and females with and without face mask, with 50 data samples in each class. In the case of radar, each instance of data sample is represented in the form of a spectrogram, displayed in Figure. 3.3c for with face mask. The spectrograms for without face mask scenario are represented in Figure. 3.3e.

Data Pre-Processing Radar Data

In the beginning, the radar chip was configured via the XEP interface with x4driver. Data were recorded from the module at 500 Frames Per Second (FPS) in the form of float message data. A loop was used to read the data file and save the data into a data stream variable, which was mapped into a complex range-time-intensity matrix. Thereafter, the Moving Target Indication (MTI) filter was applied to get the doppler range map. Afterward, the second MTI was used as a butterworth 4^{th} order filter to generate the spectrograms using the following parameters: window length, overlap percentage, and Fast Fourier Transform (FFT) padding factor. In particular, a window length of 128 samples, and a padding factor of 16 were used. In addition, a range profile was created by first converting each chirp to an FFT. A second FFT is then conducted on a defined number of consecutive chirps for a given range bin. Furthermore, an STFT was used to create these spectrograms, because, unlike Fourier Transform (FT), it offers both temporal and frequency information [145]. This is done by segmenting the data and then performing a fourier transform on each segment. When the window length is changed, both the temporal and frequency resolutions are altered inversely. For example, if one increases the other decreases. The level of doppler detail in radar data is determined by the hardware's sampling capability. The greatest unambiguous doppler frequency in radar is $F_d, max = \frac{1}{2}t_r$, where t_r is the chirp time. In this chapter, we look at lip-reading recognition at a distance D(t) from a specified location such as the mouth. V(t) represents the point of target movement in front of the radar, and T_s represents the transmitted signal,

$$T_s(t) = A\cos(2\pi ft). \tag{3.1}$$

The received signal is provided by Rs(t),

$$R_s(t) = \acute{A}\cos(2\pi f(t - \frac{2D(t)}{c})), \tag{3.2}$$

where A is the reflection coefficient, and c is the speed of light. The reflected signal can be expressed as $R_s(t)$, where the signal reflected off the target points at an angle θ to the direction of radar.

$$R_s(t) = \acute{A}\cos(2\pi f(1 + \frac{2v(t)}{c})(t - \frac{4\pi D(\theta)}{c})). \tag{3.3}$$

The Doppler shift that corresponds to it can be written as,

$$f_d = f\frac{2v(t)}{c}. \tag{3.4}$$

The returned signal becomes a composite of several moving elements such as the head, and lips. Each component moves at its own speed and acceleration. If we consider i to be the various moving components of the lip, we can write the received signal as

$$R_s(t) = \sum_{i}^{N} A_i\cos(2\pi f(1 + \frac{2vi(t)}{c})(t - \frac{4\pi D_i(0)}{c})). \tag{3.5}$$

The doppler shift is the result of a complex interaction of numerous doppler shifts induced by different moving face parts. Detection of lip-reading in a reliable fashion clearly depends upon the characteristics of the doppler signatures. After obtaining the spectrograms of various vowels and empty files from the participants, a dataset was constructed. The spectrogram is in the form of an image; therefore, we used DL models. The dataset consisted of two key modules: (i) system training and (ii) system testing. We implemented the proposed pre-trained DL classification algorithms on the spectrogram to recognize vowels and the empty dataset, as indicated in the high-level signal flow diagram in Figure 3.4a.

Data Pre-Processing Wi-Fi Data

The data was transmitted in the form of OFDM symbols comprising of 52 closely spaced subcarriers. Data were collected in the forms a matrix that contains frequency responses of all N = 51 subcarriers as shown in Eq. 3.6.

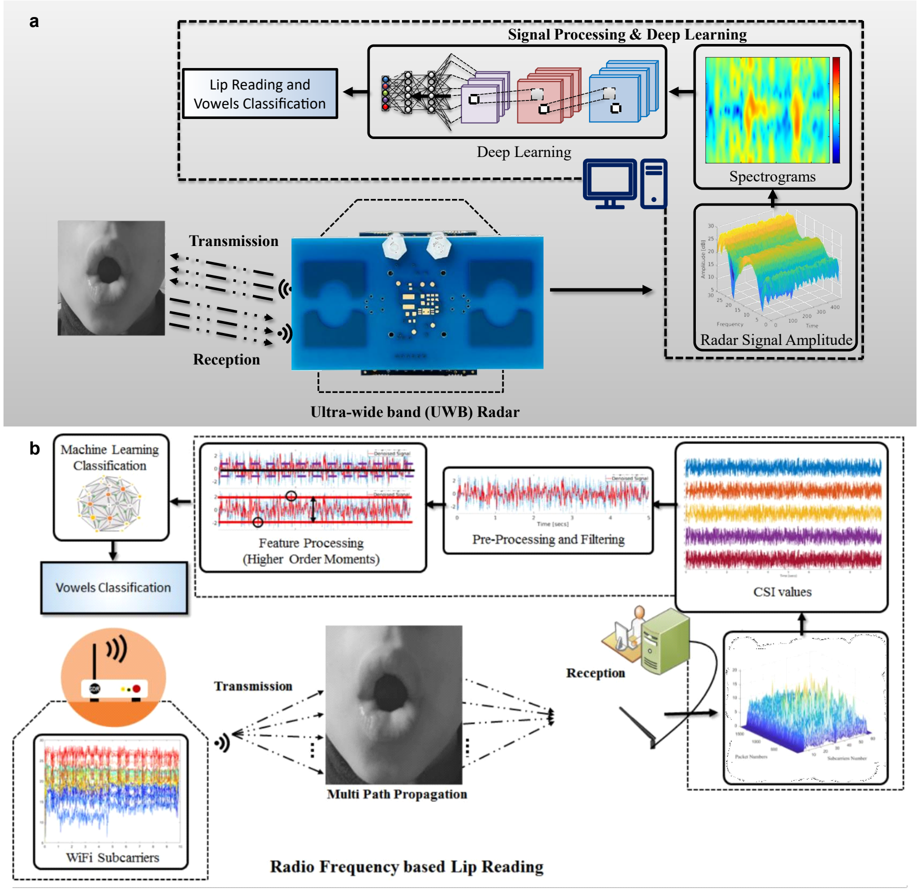

a
Signal Processing & Deep Learning
Lip Reading and Vowels Classification
Deep Learning
Spectrograms
Transmission
Reception
Radar Signal Amplitude
Ultra-wide band (UWB) Radar
b
Machine Learning Classification
Vowels Classification
Feature Processing (Higher Order Moments)
Pre-Processing and Filtering
CSI values
Transmission
Reception
Multi Path Propagation
WiFi Subcarriers
Radio Frequency based Lip Reading

Figure 3.4: Overall system overview. (a) Radar-based system overview and data collection for lip-reading. (b) Wi-Fi-based system overview and data collection for lip-reading.

$$H = [H_1(f), H_2(f), \cdots, H_N(f)]^T, \tag{3.6}$$

Here frequency of each subcarrier H_j can be represented as

$$H_j(f) = |H_j(f)|e^{j\angle H_j(f)}, \tag{3.7}$$

where $|H_j(f)|$ and $\angle H_j(f)$ are the amplitude and phase responses of the jth subcarrier. Each of these subcarrier responses is related to the system input and output as given in Eq. 3.8,

$$H_j(f) = \frac{Y_j(f)}{X_j(f)}, \tag{3.8}$$

where $Xj(f)$ and $Yj(f)$ are the fourier transforms of input and output of the system. Indeed, the received CSI samples are impaired due to environmental noise.

As a result, the collected samples are denoised by subtracting the mean received power from each subcarrier. To observe the maximum variation due to lip movements the subcarrier with highest variance was identified for the feature extraction. A total of 9 features were extracted namely, mean, median, standard deviation, variance, minimum, eight peaks and high order moments, such as skewness and kurtosis. The numerical extracted features were stored in a Comma-Separated Values (CSV) file, which was then utilised by various ML algorithms. For each ML model, we employed a standard train-test split of 80% for training and 20% for testing to ensure sufficient data for learning while retaining a representative subset for evaluation. To compute performance metrics and ensure the reliability and stability of our models, we utilised k-fold cross-validation, specifically with k set to 5, across the training data. This method partitions the data into k subsets, trains the model on k-1 of those subsets, and validates it on the remaining subset, repeating this process k times with each subset used exactly once for validation. This approach allows for a comprehensive evaluation of the model's performance across different segments of the data. Following this, training, testing, and validation were conducted using the test-train split evaluation method to accurately classify vowels and the empty class, as shown in the high-level signal flow diagram in Figure. 3.4b.

Parameter Settings of the Considered Algorithms

The proposed classification methodology to distinguish lip-reading activities is divided into two key stages: (i) system training and (ii) system testing. In the case of radar data, the DL pre-trained models VGG16, VGG19, and InceptionV3 [146] were used on the spectrogram images generated from the radar data. While ML algorithms NN pattern recognition, support vector machine (SVM, medium gaussian SVM), Ensemble (boosted trees) and Naïve Bayes (kernel Naïve Bayes) were used on Wi-Fi data.The parameter settings of ML and DL model are shown in Table. 3.3.

VGG16 Model: VGG16 has been used with 16 convolution layers and a Rectified Linear Unit (ReLU) activation function, with kernel sizes of 3×3. Following each convolution layer, a max-pooling layer with all kernel sizes of 2×2 was added. Final layer worked as three fully connected layers (FC). The convolution layer and FC hold the weight of the training results, which allows them to determine the number of parameters.

VGG19 Model: A 3×3 filter was used to capture image details, consisting of five stages of convolution layers, five pooling layers, and three fully connected layers. The depth of the convolution kernel in the VGG19 network has been raised from 64 to 512, allowing for improved image feature vector extraction. A pooling layer was applied after each stage of convolutional layers. Each pooling layer has the same size and step size, which is 2×2.

InceptionV3 Model: A 48-layered InceptionV3 DL model was also applied on the dataset. Three convolution layers were added first, followed by a max pooling layer, two more convolution layers, and another max pooling layer. The spectrograms were sent to various convolutions, which convoluted the input images using various filters, stacked the extracted data, and sent it forward, and this process was repeated multiple times across the network [147], rather than manually adjusting the filter size for each layer.

Neural Network Pattern Recognition Model: Data were passed through two-layer feed-forward networks with sigmoid hidden neurons, SoftMax output neurons, and scaled conjugate gradient back propagation. Meanwhile, weight and bias values are updated according to the scaled conjugate gradient method. Training, validation, and test sets of data were created. Network performance was measured using cross-entropy and miss-classification errors.

SVM (Medium Gaussian SVM) Model: SVM was used for classification of dataset by determining the optimum hyperplane for separating data points from one class to another. Training data, parameter values, prior probabilities, support vectors, and algorithmic implementation details were stored in trained SVM classifiers. The experimental data was modelled using a Gaussian kernel.

Ensemble (Boosted Trees) Model: Ensemble classifiers combined the results of a number of low-quality learners into a single high-quality ensemble model. Boosting ensemble method was used on the dataset to regulate the depth of tree learners by specifying the maximum number of splits or branch points. The experimental setup achieved better accuracy with 0.1 learning rate.

Naïve Bayes(Kernel Naïve Bayes) Model: Naïve Bayes classifier was used for lip-reading classification, which is based on Bayes theorem and assumes that predictors are conditionally independent in the given class. Specifically, a Gaussian Naïve Bayes kernel was used in this experiment

DL/ML Model	Parameters	Settings
VGG16	Number of Layers	16
	Initial learning rate	0.0001
	Mini-batch size	16
	Learning algorithm	Adam
	Loss function	Cross entropy
	Maximum epochs	25
	Iteration per epoch	500
VGG19	Initial Number of Layers	19
	learning rate	0.0001
	Mini-batch size	16
	Learning algorithm	Adam
	Loss function	Cross entropy
	Maximum epochs	25
	Iteration per epoch	500
InceptionV3	Number of Layers	48
	Initial learning rate	0.0001
	Mini-batch size	16
	Learning algorithm	Adam
	Loss function	Cross entropy
	Maximum epochs	25
	Iteration per epoch	500
NN	Initial Number of Layers	10
	Training Function	Scaled conjugate Gradient Backpropagation
	Number of Epoch	20
	Loss function	Cross entropy
SVM	Kernel Scale	3.9
	K-Fold Predict	5
	Kernel Function	Gaussion
	Loss Function	Classiferror
	Multiclass Method	One-vs-One
Ensemble	Ensemble Method	AdaBoost
	Learner type	Decision Tree
	Maximum Number of splits	20
	Learning rate	0.1
	Number of learners	30
	Loss Function	Classiferror
Naïve Bayes	Kernel Smooth Density	Unbounded
	K-fold Predict	5
	Kernel Function	Gaussian
	Loss function	Classiferror
	Predictor distribution	mvmn

Table 3.3: Parameter settings for the selected models.

Evaluation Metrics of Classification Models

The performance of the DL and ML models in the classification of vowels, consonants, and words is evaluated through accuracy, True Positive Rate (TPR), and False Positive Rate (FPR). TPR and FPR are calculated using equation 3.9 and 3.10, respectively. The equation 3.13 is used to calculate the F1 score, one of the most popular classification metrics in the literature. The F1−Score combines precision and recall, which are calculated using the equations 3.11 and 3.12. The equation 3.14 was used to calculate the Average accuracy, used to evaluate the performance of ML models.

$$TPR = \frac{TP}{TP+FN} \tag{3.9}$$

$$FPR = \frac{FP}{FP+TN} \tag{3.10}$$

$$Precision = \frac{\sum(TP)}{\sum(TP+FP)} \tag{3.11}$$

$$Recall = \frac{\sum(TP)}{\sum(TP+FN)} \tag{3.12}$$

$$F1-Score = 2\frac{(Precision.Recall)}{(Precision+Recall)} \tag{3.13}$$

$$Accuracy = \frac{\sum(TP+TN)}{\sum(TP+FP+TN+FN)} \tag{3.14}$$

where TP stands for true positive, i.e., both the truth and the predicted values are positive. FN is false negative, which represents the cases when the truth is positive and the prediction is negative.

3.2.2 Result and Discussion

In this section, a lip-reading RF-sensing based framework is proposed using both RF sensing technologies, *i.e.,* Wi-Fi and radar. Wi-Fi signals are generated using USRP X300, which uses CSI signals to identify human lip movements for all considered classes, i.e., A, E, I, O, U, and Empty. For radar, a UWB radar sensor, Xethru X4M03 was used, where reflected doppler signals (Hz) were plotted in the form of frequency-time diagrams, such as spectrograms. The proposed RF sensing system can either work as standalone or assist in sensing for hearing aids through reading of lip and mouth movements in the presence of face masks, which normally obstruct visual cues for hearing aids in vision-based systems. A diverse dataset of three participants (one male and two females) was collected for 5 vowels A, E, I, O, U, and Empty, where lips were not moving. The collected dataset was used to train different ML and DL algorithms. The work's major goal was to propose a secure lip-reading system that could identify the lip movements in the presence of a mask with different RF sensing technologies and ML/DL algorithms. In particular, four algorithms, NN, SVM, Ensemble, and Naïve Bayes, were evaluated using train-test evaluation methods on the Wi-Fi dataset, where the maximum classification test accuracy of 93.3% was observed on the male datasets without face mask. On the other hand, DL pre-trained models VGG16, VGG19, and InceptionV3, evaluated using train-test methods, achieved a maximum average test accuracy of 91.67% on male data without masks using radar. The proposed model's performance was assessed using TPR and FPR metrics derived from confusion matrices, ensuring reliable classification accuracy. Moreover, because the current system is a proof of concept with the goal of showing the importance and effectiveness of detecting lips using RF-sensing technology such as radar and Wi-Fi, future experiments will be conducted to detect different words or sentences in real time and perform activity from various angles using radar and Wi-Fi. Furthermore, and as mentioned earlier, the dataset used to achieve the previously reported results is made publicly available to encourage other researchers and the wider communities to take this system a step further.

Radar Data

The evaluation results of the considered DL algorithms (VGG16, VGG19, and InceptionV3) on the radar dataset are presented in Table. 3.4 and 3.5. VGG and InceptionV3 are CNN-based DL models (trained on ImageNet dataset [148]), which are commonly used in image classification. VGG16, VGG19, and InceptionV3 have 16, 19, and deep layers, respectively. A detailed description of these models is presented in references [149, 150]. Moreover, [151] provides the fundamental understanding of ML. It can be observed from Table. 3.4 and 3.5 that all algorithms produce comparable results with VGG16 slightly outperforming others on all individual subjects and combined dataset in terms of accuracy. Using VGG16, the classification accuracy of 91.7% is observed on S1 dataset without mask, which is reduced to a promising accuracy 83.3% when the subject wears the face mask. The other performance metrics, such as TPR and FPR are presented in Table. 3.4 and 3.5. It can be observed from the tables that they perform well on all individual classes. Almost all individual classes produce 100% TPR with mask and promising TPR on without mask dataset. Similarly, on combined dataset the same algorithm produces best results in terms of classification accuracy for both with mask and without mask. Overall, a classification accuracy of 85.94% is observed on without face mask combined dataset. On the other hand, the same algorithm classifies the vowels with 73.44% accuracy with face mask on the combined dataset. Moreover, other DL models, *i.e.,* VGG19 and Inception V3 also produce comparable results on the radar dataset.

Figure. 3.5 shows the accuracies of with mask and without mask scenarios for different DL algorithms examined on the radar data of male subject. It can be observed from the figure that InceptionV3 produces biggest accuracy difference between with mask and without mask

DL Model		TPR/FPR (%)	S1(Male)							S1(Female)						
			A	E	I	O	U	Emp	Accuracy (%)	A	E	I	O	U	Emp	Accuracy (%)
VGG16	With Mask	TPR (%)	90.0	50.0	80.0	90.0	90.0	100	83.3	70.0	50.0	100	90.0	100	100	85.0
		FPR (%)	10.0	50.0	20.0	10.0	10.0	0.0		30.0	50.0	0.0	10.0	0.0	0.0	
	Without Mask	TPR (%)	80.0	100	80.0	90.0	100	100	91.7	86.7	63.3	83.3	83.3	83.3	100	83.3
		FPR (%)	20.0	0.0	20.0	10.0	0.0	0.0		13.3	36.7	16.7	16.7	16.7	0.0	
VGG19	With Mask	TPR (%)	70.0	80.0	90.0	80.0	80.0	100	83.33	50.0	45.0	92.0	90.0	82.0	90.0	75.0
		FPR (%)	30.0	20.0	10.0	20.0	20.0	0.0		50.0	55.0	8.0	10.0	18.0	10.0	
	Without Mask	TPR (%)	80.0	90.0	70.0	90.0	90.0	100	86.67	90.0	70.0	70.0	100	60.0	100	81.67
		FPR (%)	20.0	10.0	30.0	10.0	10.0	0.0		10.0	30.0	30.0	0.0	40.0	0.0	
InceptionV3	With Mask	TPR (%)	80.0	70.0	90.0	50.0	90.0	100	80.0	100	60.0	60.0	30.0	70.0	100	70.0
		FPR (%)	20.0	30.0	10.0	50.0	10.0	0.0		0.0	40.0	40.0	70.0	30.0	0.0	
	Without Mask	TPR (%)	100	90.0	50.0	100	100	100	90.0	80.0	70.0	90.0	50.0	90.0	100	80.0
		FPR (%)	0.0	10.0	50.0	0.0	0.0	0.0		10.0	30.0	10.0	50.0	10.0	0.0	

Table 3.4: Comparative result of vowels with and without mask using radar dataset.

DL Model		TPR/FPR (%)	S3 (Female)							Combined						
			A	E	I	O	U	Emp	Accuracy (%)	A	E	I	O	U	Emp	Accuracy (%)
VGG16	With Mask	**TPR(%)**	50.0	80.0	100	70.0	60.0	100	**76.7**	75.0	67.0	44.0	67.0	67.0	100	**73.44**
		FPR(%)	50.0	20.0	0.0	30.0	40.0	0.0		25.0	33.0	56.0	33.0	33.0	0.0	
	Without Mask	**TPR(%)**	80.0	90.0	80.0	80.0	80.0	100	**85.0**	91.0	82.0	100	90.0	62.0	100	**85.94**
		FPR(%)	20.0	10.0	20.0	20.0	20.0	0.0		9.0	18.0	0.0	10.0	38.0	0.0	
VGG19	With Mask	**TPR(%)**	80.0	40.0	80.0	80.0	100	70.0	**75.0**	40.0	69.2	100	83.3	40.0	100	**68.9**
		FPR(%)	20.0	60.0	20.0	20.0	0.0	30.0		60.0	30.8	0.0	16.7	60.0	0.0	
	Without Mask	**TPR(%)**	50.0	80.0	100	70.0	60.0	100	**76.7**	88.0	73.0	100	36.0	78.0	100	**79.69**
		FPR(%)	50.0	20.0	0.0	30.0	40.0	0.0		12.0	27.0	0.0	64.0	22.0	0.0	
InceptionV3	With Mask	**TPR(%)**	80.0	40.0	80.0	80.0	100	70.0	**75.0**	76.0	60.0	36.0	46.0	82.0	90.0	**65.0**
		FPR(%)	20.0	60.0	20.0	20.0	0.0	30.0		24.0	40.0	64.0	54.0	18.0	10.0	
	Without Mask	**TPR(%)**	80.0	30.0	100	80.0	90.0	100	**80.0**	75.0	67.0	80.0	80.0	22.0	100	**73.44**
		FPR(%)	20.0	70.0	0.0	20.0	10.0	0.0		25.0	33.0	20.0	20.0	78.0	0.0	

Table 3.5: Comparative result of vowels with and without mask using radar dataset.

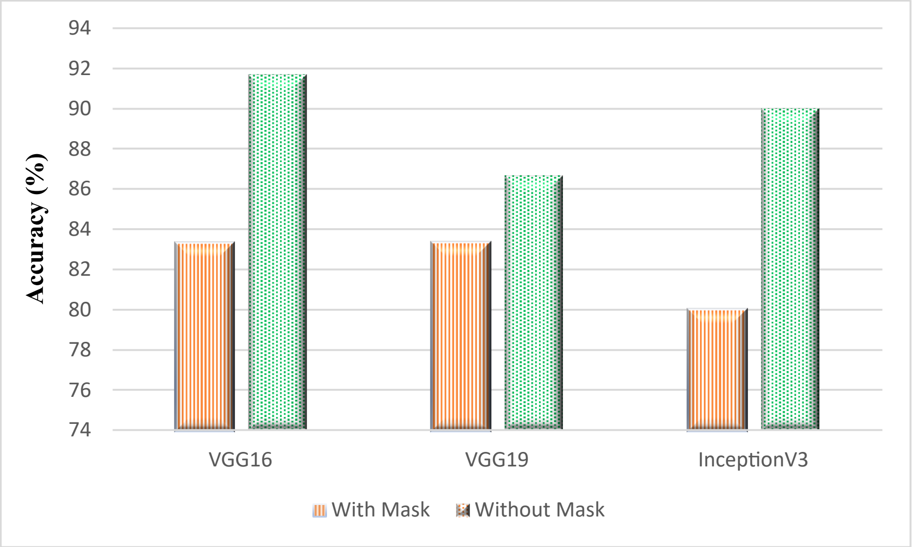


Figure 3.5: The accuracy improvement of male subject using DL algorithms between with mask and without mask using radar.

cases, which is around 12%, while VGG19 produces the least different, which is around just 4%. Overall, VGG16 performs better on both datasets with an accuracy difference of 7%.

Wi-Fi Data

Table. 3.6 and 3.7 represents the average accuracy of classifying the dataset collected from Wi-Fi using different ML and DL algorithms. Four different algorithms are considered, namely NN, Support Vector Machine (SVM), Ensemble and Naïve Bayes. The results are generated using test-train split evaluation method.

ML Model		TPR/FPR (%)	S1 (Male)							S2 (Female)						
			A	E	I	O	U	Emp	Accuracy (%)	A	E	I	O	U	Emp	Accuracy (%)
SVM (Medium Gaussian SVM)	**With Mask**	**TPR (%)**	54.0	40.0	42.0	38.0	46.0	88.0	**51.3**	38.0	72.0	58.0	70.0	36.0	96.0	**61.7**
		FPR (%)	46.0	60.0	58.0	62.0	54.0	12.0		62.0	28.0	42.0	30.0	64.0	4.0	
	Without Mask	**TPR (%)**	62.0	86.0	72.0	46.0	84.0	88.0	**73.0**	82.0	76.0	46.0	36.0	60.0	90.0	**65.0**
		FPR (%)	38.0	14.0	28.0	54.0	16.0	12.0		18.0	24.0	54.0	64.0	40.0	10.0	
Neural Network Pattern Recognition	**With Mask**	**TPR (%)**	60.0	100	60.0	60.0	75.0	100	**73.3**	50.0	75.0	71.4	85.7	50.0	100	**80.0**
		FPR (%)	40.0	0.0	40.0	40.0	25.0	0.0		50.0	25.0	28.6	14.3	50.0	0.0	
	Without Mask	**TPR (%)**	100	100	100	100	60.0	100	**95.6**	68.0	96.0	74.0	54.0	80.0	86.0	**76.3**
		FPR (%)	0.0	0.0	0.0	0.0	40.0	0.0		32.0	4.0	26.0	46.0	20.0	14.0	
Naive Bayes (Kernel Naive Bayes)	**With Mask**	**TPR (%)**	58.0	52.0	44.0	34.0	30.0	94.0	**52.0**	38.0	80.0	56.0	76.0	18.0	96.0	**60.7**
		FPR (%)	42.0	48.0	56.0	66.0	70.0	6.0		62.0	20.0	44.0	24.0	82.0	4.0	
	Without Mask	**TPR (%)**	60.0	88.0	78.0	40.0	76.0	98.0	**73.3**	76.0	66.0	58.0	22.0	58.0	96.0	**62.7**
		FPR (%)	40.0	12.0	22.0	60.0	24.0	20.0		24.0	34.0	42.0	78.0	42.0	4.0	
Ensemble (Boosted Trees)	**With Mask**	**TPR (%)**	58.0	80.0	42.0	66.0	16.0	96.0	**59.7**	58.0	80.0	42.0	66.0	16.0	96.0	**59.7**
		FPR (%)	42.0	20.0	58.0	34.0	84.0	4.0		42.0	20.0	58.0	34.0	84.0	4.0	
	Without Mask	**TPR (%)**	68.0	96.0	74.0	54.0	80.0	86.0	**76.3**	66.0	62.0	44.0	48.0	54.0	94.0	**61.3**
		FPR (%)	32.0	4.0	26.0	46.0	20.0	14.0		34.0	38.0	56.0	52.0	46.0	6.0	

Table 3.6: Comparative result of vowels with and without mask using Wi-Fi dataset.

ML Model		TPR/FPR (%)	S3 (Female)							Combined						
			A	E	I	O	U	Emp	Accuracy (%)	A	E	I	O	U	Emp	Accuracy (%)
SVM (Medium Gaussian SVM)	**With Mask**	**TPR (%)**	76.0	10.0	62.0	60.0	24.0	96.0	**54.7**	32.7	46.0	36.7	56.0	34.0	100	**50.9**
		FPR (%)	24.0	90.0	38.0	40.0	76.0	4.0		67.3	54.0	63.3	44.0	66.0	0.0	
	Without Mask	**TPR (%)**	56.0	56.0	52.0	56.0	60.0	90.0	**61.7**	48.0	40.0	44.7	42.0	67.3	100	**57.8**
		FPR (%)	44.0	44.0	48.0	44.0	40.0	10.0		52.0	60.0	55.3	58.0	32.7	0.0	
Neural Network Pattern Recognition	**With Mask**	**TPR (%)**	60.0	88.0	78.0	40.0	76.0	98.0	**76.7**	41.7	50.0	61.5	63.2	35.7	100	**61.1**
		FPR (%)	40.0	12.0	22.0	60.0	24.0	2.0		58.3	50.0	38.5	36.8	64.3	0.0	
	Without Mask	**TPR (%)**	100	100	83.3	83.3	66.7	100	**88.9**	60.0	100	60.0	60.0	75.0	100	**73.3**
		FPR (%)	0.0	0.0	16.7	16.7	33.3	0.0		40.0	0.0	40.0	40.0	25.0	0.0	
Naive Bayes (Kernel Naive Bayes)	**With Mask**	**TPR (%)**	80.0	8.0	64.0	72.0	14.0	90.0	**54.7**	36.7	45.3	39.3	46.0	27.3	100	**49.0**
		FPR (%)	20.0	92.0	36.0	28.0	86.0	10.0		63.3	54.7	60.7	54.0	72.7	0.0	
	Without Mask	**TPR (%)**	58.0	58.0	38.0	58.0	64.0	98.0	**62.3**	34.7	46.7	62.0	25.3	48.7	95.3	**52.1**
		FPR (%)	42.0	42.0	62.0	42.0	36.0	2.0		65.3	53.3	38.0	74.7	51.3	4.7	
Ensemble (Boosted Trees)	**With Mask**	**TPR (%)**	64.0	50.0	40.0	40.0	46.0	96.0	**56.0**	44.0	56.0	40.7	63.3	37.3	100	**56.9**
		FPR (%)	36.0	50.0	60.0	60.0	54.0	4.0		56.0	44.0	59.3	36.7	62.7	0.0	
	Without Mask	**TPR (%)**	60.0	56.0	56.0	54.0	58.0	92.0	**62.7**	50.0	51.3	48.0	33.3	64.0	100	**57.8**
		FPR (%)	40.0	44.0	44.0	46.0	42.0	8.0		50.0	48.7	52.0	66.7	36.0	0.0	

Table 3.7: Comparative result of vowels with and without mask using Wi-Fi dataset.

It can be noted from tables that NN algorithm outperforms others for individual male and female data and the combined dataset. Using NN algorithm, the classification accuracy of 95.6% is observed on S1 without face mask, while the same algorithm gives 73.3% classification accuracy on the same subject when he wears a face mask. Similarly, on the combined dataset, NN gives a premising accuracy of 73.3% without face mask and an accuracy of 61.1% on with-mask combined dataset. The other performance metrics, such as TPR and FPR are shown in Table. 3.6 and 3.7. It can be observed that these metrics perform well in all individual classes. Almost all individual classes produce 100% TPR with mask and promising TPR on without mask dataset. Interestingly, the classification accuracy of male dataset for all algorithms is higher than the females' dataset. This is due to the reason that the lip movements of male subject in pronouncing vowels were comparatively larger than females among the participants.

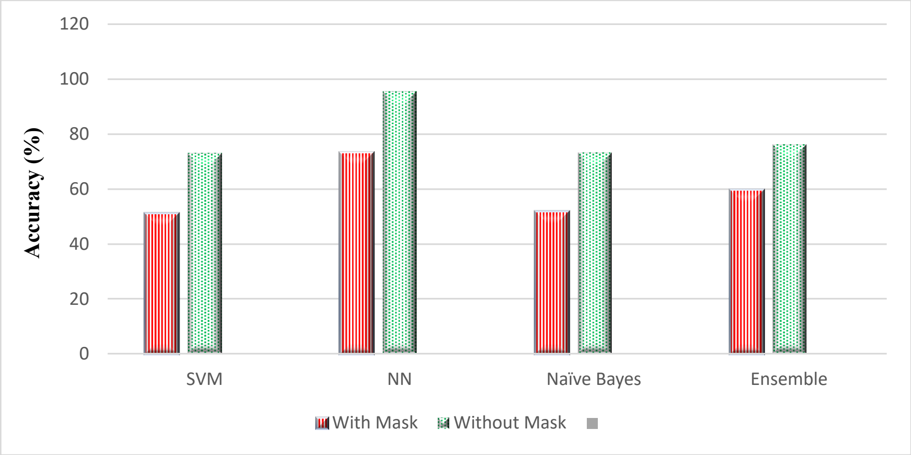


Figure 3.6: The accuracy improvement of male subject using ML algorithms between with mask and without mask using Wi-Fi.

Overall, the classification accuracy of with mask dataset is lower than without face mask. This is because of the reason that the lip movements are restricted due to the restraints caused by the face mask. For instance, a person may not be able to fully open the mouth while wearing face mask. The percentage accuracy difference in classifying with mask and without mask dataset is depicted in Figure. 3.6. The highest accuracy difference is observed for male subjects for NN algorithm where an accuracy difference of around 23%. The minimum difference observed is for ensemble algorithm on S1 dataset, where with mask and without mask accuracy difference is 12%.

3.3 RFID Based Speech Recognition

3.3.1 RFID Tag Performance Setup and Test Results

The passive UHF RFID tag used in our proposed smart mask underwent testing for reusability and rigor. It is a flexible, low-profile, linearly polarised textile laundry tag that offers versatile attachment methods and meets specific electrical specifications. The dimension of the tag is 58x15x1.5 mm. It is an EPC Gen2 compliance tag with a copper dipole antenna and Impinj Monza R6P Integrated Circuit (IC)/chip. A simplified model of the tag chip, consisting of lumped elements, is shown in Figure. 3.7a. The port model is derived using a source-pull method due to the nonlinear and time-varying nature of the tag's RF circuits. This model is an accurate mathematical representation of the chip's behavior over a wide range of frequencies. Table. 3.8, provides the values of the lumped elements for the Monza R6-P tag chip's port model, which are valid for all primary regions of operation within the UHF range (868-920 MHz). The lumped elements include C_{mount}, which represents the parasitic capacitance resulting from the overlap of the antenna trace with the chip surface, C_p, which is intrinsic to the chip and appears at the chip terminals, and R_p, which represents the energy conversion and absorption of the RF circuits.

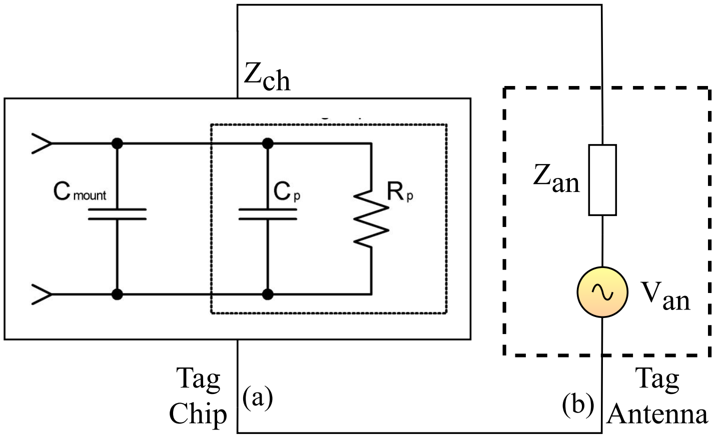


Figure 3.7: linearised RF-model of the tag. (a) Tag chip lumped element model. (b) Tag antenna lumped element model.

Symbol	**Parameter**	**Typical Value**
C_p	Chip Capacitance	1.23 *pF*
R_p	Chip Resistance	1.2 *kOhm*
C_{mount}	Capacitance due to adhesive and antenna mount parasitics	0.21 *pF*
Sensitivity	Chip Read Sensitivity	- 20 *dBm*

Table 3.8: Operating conditions and electrical characteristics of Monza R6-P chip port model.

The chip impedance Z_{ch} and antenna impedance Z_{an}, which vary with frequency, can be expressed according to [152, 153, 154], and the equivalent lumped circuit depicted in Figure. 3.7 as:

$$Z_{ch} = R_{ch} + jX_{ch} \tag{3.15}$$

$$Z_{an} = R_{an} + jX_{an} \tag{3.16}$$

The chip and antenna resistance is represented by R_{ch} & R_{an}, respectively, while the chip and antenna reactance is denoted by X_{ch} & X_{an}. *Vant* refers to the open-circuit RF voltage that arises from the electromagnetic field generated by the reader at the terminals of the tag antenna. The impedance of the chip, Z_{ch}, is affected by the power that the chip absorbs, P_{ch}, and this often has a draining effect on energy. To determine the power that is absorbed by the tag chip, P_{ch}, we utilise the maximum available power from the antenna, P_{an}, as well as the power transmission coefficient, P_{ch}, as shown below:

$$P_{ch} = P_{an}\tau \tag{3.17}$$

The maximum antenna power, P_{an}, is achieved when $Z_{ch} = Z_{an}$. The power transmission coefficient, τ, represents the degree of impedance matching between the IC and the antenna and is expressed as follows:

$$\tau = \frac{4 * R_{ch} R_{an}}{Z_{ch} + Z_{an}} \tag{3.18}$$

As τ approaches unity, the match between the tag chip and antenna impedance improves, with a perfect complex conjugate match achieved at τ=1. Thus, for a given chip-and-tag antenna setup, an ideal situation would be where Z_{ch}=Z_{an}, corresponding to τ=1. Moreover, in order for the chip to activate, the antenna is often matched to the minimum threshold power, P_{th}. The Friis free-space equation is utilised to compute the free-space tag antenna power, P_{an}, where:

$$P_{an} = P_{read} G_{ant} G_{read} (\frac{\lambda}{4\pi d})^2 \tag{3.19}$$

Here, P_{read} and G_{read} refer to the reader-transmitted power and antenna gain, respectively. G_{ant} represents the tag antenna gain, λ denotes the wavelength, and d represents the distance between the tag and reader. Substituting equation (3.17) and determining the read range, r, at which the tag receives the minimum P_{th} yields the following equation:

$$r = \frac{\lambda}{4\pi} \sqrt{\frac{P_{read} G_{ant} G_{read} \tau}{P_{th}}} \tag{3.20}$$

The tag's resonance, which represents the peak read range over a frequency range, is associated with the maximum power transmission coefficient, τ. Therefore, in order to achieve the maximum read range, it is essential to optimise the tag antenna to achieve the highest power transmission coefficient, τ, and then utilise (3.20) in conjunction with the reader system to calculate the corresponding read range, r. For a parallel circuit with a resistor and capacitor:

$$Q = R_p \times \omega(C_p + C_{mount}) \tag{3.21}$$

From 3.8, the parallel resistance of the tag chip, R_p = 1200 Ω, while the parallel capacitance, C_p = 1.23 pF with an additional parasitic capacitance of 0.21 pF. Furthermore, $\omega = 2\pi f$, where f is the central frequency of 900 MHz. As per circuit theory, the real component of the chip impedance, R_{ch}, at this frequency can be calculated as $R_{ch} = R_p/(1+Q^2)$. For this particular case, R_{ch} evaluates to 12.3 Ω. Additionally, the imaginary component of the chip impedance is $1/(\omega Cp)$ and equals -120.8 Ω. Utilising (3.15), we can express the chip impedance, Z_{ch} = 12 – j121 Ω. The matching of the antenna-chip impedance is validated through the read range tests, which are discussed in the subsequent section.

The chip employed in these tags is characterised by exceptional read sensitivity of up to -22.1 dBm when used with a dipole antenna. Additionally, the chip leverages autotune technology to maintain performance consistency across various dielectric materials. This technology's

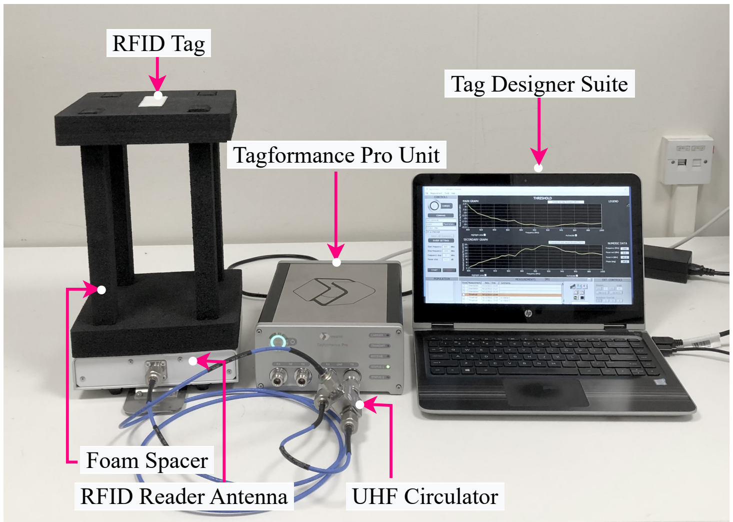


Figure 3.8: Experimental setup for tag measurements, using tagformance pro device.

primary advantage lies in its cost-effectiveness and high efficiency in achieving our research objectives. It has the capacity to store individualised data within the integrated IC and seamlessly integrates into Internet of Things (IoT) technologies. This functionality allows for the unique identification of items by leveraging the unique ID stored in its IC. However, a potential limitation of RFID technology is its restricted read range capability. The read range tests and other necessary measurements are presented in Section 3.3.1.

RFID Tag Measurement Results using Tagformance® Pro Unit

To evaluate the reliability of the RFID tag, we employed the Voyantic Tagformance® Pro device used in the industry. The measurement setup is shown in Figure. 3.8. The Tagformance device is composed of several components, including Tag Designer Suite (TDS) software, a Tagformance unit that comes with a UHF circulator and a foam spacer, and a linearly polarised RFID reader antenna that has a gain of 6 dBi and can be adjusted through the settings [155].

In the read range test, the sensitivity of the RFID tag is assessed across a frequency range of 800 MHz to 1000 MHz. At each frequency, the power of the forward and backscatter signals on the tag is analysed with various transmit-power levels. The test results are illustrated in Figure. 3.9. The read range measurements for the dry and wet tag are presented in Figure. 3.10. The dry tag achieves a read range of up to 6.5 m. whereas the wet tag can be read at up to 5m.

3.3.2 Methodology

The block diagram in Figure. 3.11 shows the methodology used in this chapter. There are three steps to the suggested framework. In the first step, we collected, build and annotated various lip-reading datasets. In the second step, the pre-processing phases are explained. Lastly, several ML models were used to classify the RFID-based lip-reading. The following subsections provide a

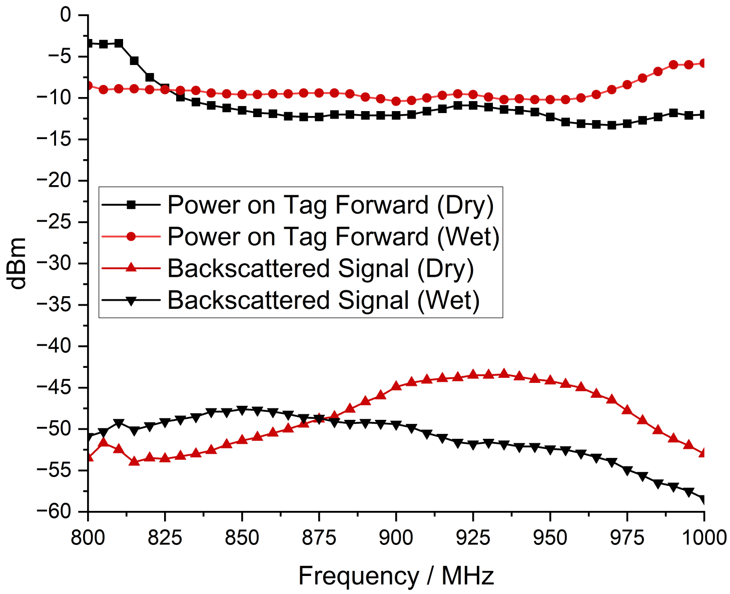


Figure 3.9: Analysed power on tag forward, and backscatter signal at 800-1000 MHz with multiple transmit-power levels for both the dry and wet tag.

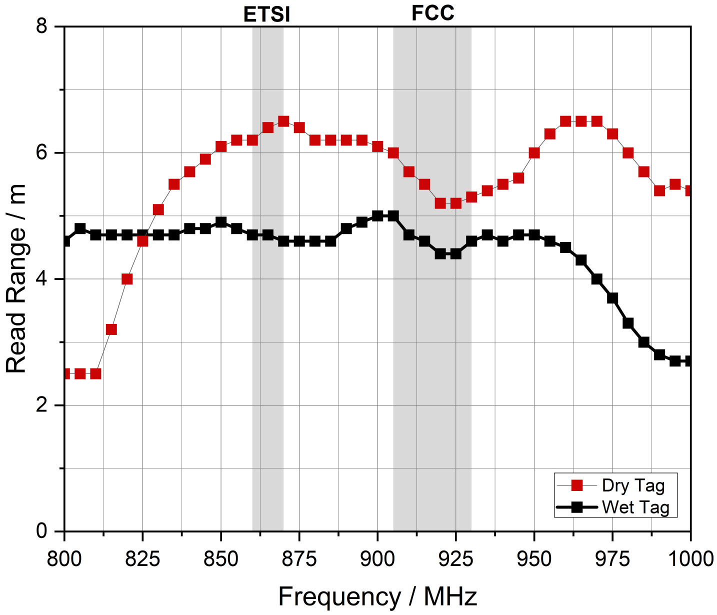


Figure 3.10: Read range measurements of the tag in both dry and wet conditions.

detailed description of each step of the proposed methodology.

Experimental Setup and Data Collection

In this section, we used an RFID-based smart mask to collect data on lip-reading. The experimental setup of the lip-reading using an RFID-based smart mask is shown in Figure. 3.12a. The RFID laundry tag was stitched on disposable face masks. The multiple color mask having different thicknesses were used for the experiments to check the authenticity of the system

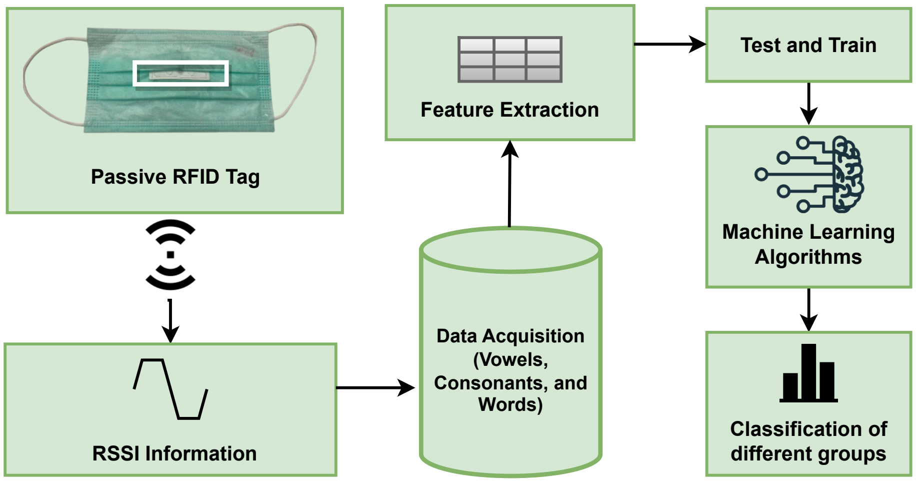

Passive RFID Tag
Feature Extraction
Test and Train
Machine Learning Algorithms
RSSI Information
Data Acquisition (Vowels, Consonants, and Words)
Classification of different groups

Figure 3.11: An overview of the proposed framework signal flow diagram highlighting the RFID technology, data collection, and ML models for lip-reading classification.

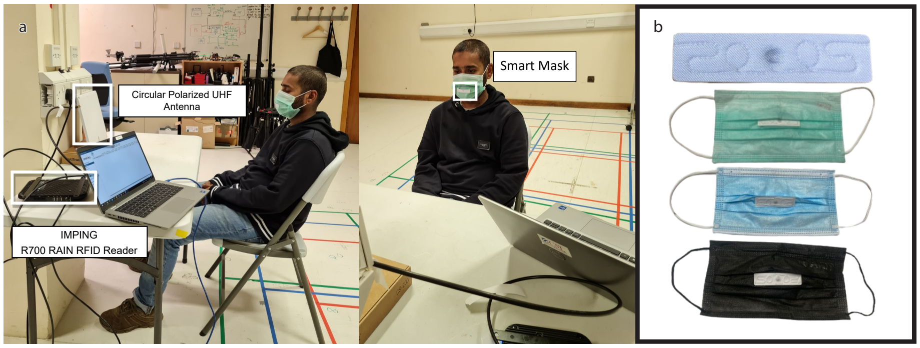
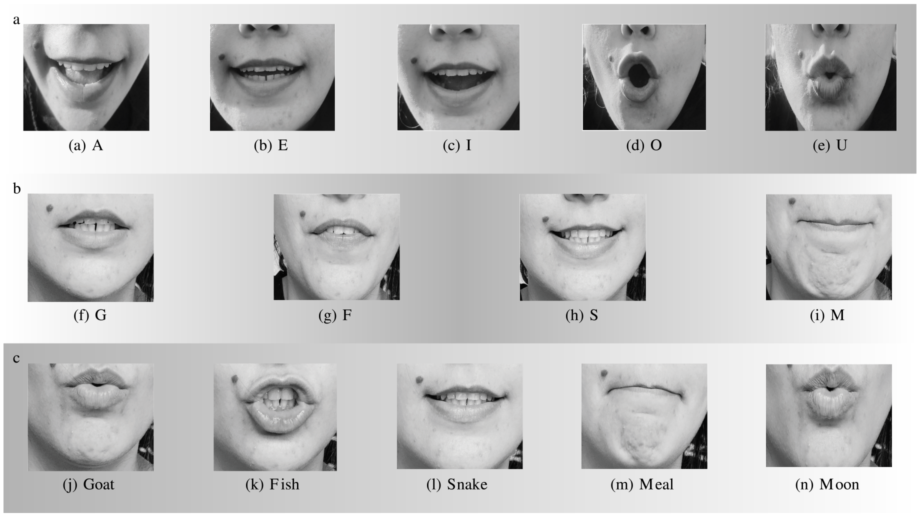


Figure 3.12: Experimental setup of lip-reading data collection using RFID-based smart mask. (a) Real experimental setup. (b) Color-thickness variants of smart masks used in the experimental setup.



(a) A (b) E (c) I (d) O (e) U

(f) G (g) F (h) S (i) M

(j) Goat (k) Fish (l) Snake (m) Meal (n) Moon

Figure 3.13: A visual illustration of the lip-reading. (a) Vowels. (b) consonants. (c) Words.

which is shown in Figure. 3.12b. The key parameter settings of the RFID lip-reading system are indicated in Table. 3.9. In this system, participants were asked to sit 0.50 meters away from the

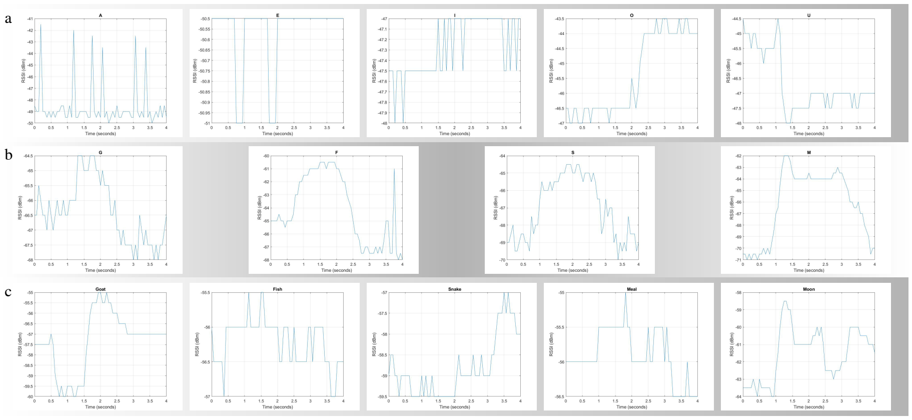


Figure 3.14: A graphical illustration of the received lip-reading signals: (a) vowels, (b) consonants, and (c) words.

Parameter	Value
Tag	Passive UHF RFID Laundry Tag
Frequency	868-920MHZ
Read Distance	Up to 6.5 meters
Reader	Impinj R7000 Rain RFID
Antenna	Circular Polarised UHF 6 dBi
Life Time	200 Wash Cycles
Mounting Methods	Sewing or Heating Sealing
Chip	Impinj Monza R6-P
Activity Duration	4 Seconds
Number Of Sample Per Class	50

Table 3.9: Selected hardware and software parameter settings.

RFID reader and antenna. The subject's body was in its regular position during data collection, with only head movements. Furthermore, each activity had a time limit of 4 seconds and the data collection process involved recording a single word/vowel/consonant from each subject. Figure 3.13 provides a visual illustration, and Figure 3.14 offers a graphical representation of the pronounced vowels, consonants, and words. A total of four participants, two males, and two females, participated in the data collection process. Multiple participants were invited to the data collection process to make the data more realistic and diverse. During the experiments, a total of 2800 data samples were collected, with 50 samples collected in each class. We distributed the dataset into three sub-classes (vowels, consonants, and words). Table. 3.10 provides a detailed overview of the collected dataset. In particular, each class is divided into two parts 80% data for training and 20% dataset for testing purposes. In each sub-set either vowels or words, a total of 1000 data samples were collected from participants, where 800 were utilised for training and 200 for testing purposes. In the case of consonants, a total of 800 data samples were collected from participants, where 640 were utilised for training and 160 for testing purposes.

Experimental Dataset					**Total**
Classes	**Subject(S1)**	**Subject(S2)**	**Subject(S3)**	**Subject(S4)**	
VOWELS					
A	50	50	50	50	200
E	50	50	50	50	200
I	50	50	50	50	200
O	50	50	50	50	200
U	50	50	50	50	200
CONSONANTS					
F	50	50	50	50	200
G	50	50	50	50	200
M	50	50	50	50	200
S	50	50	50	50	200
WORDS					
Fish	50	50	50	50	200
Goat	50	50	50	50	200
Meal	50	50	50	50	200
Moon	50	50	50	50	200
Snake	50	50	50	50	200
Total	420	420	420	420	**2800**

Table 3.10: An overview of the information gathered, the number of participants, and the activities performed.

Data Pre-Processing

The collected data was in the form of RSSI values stored in a single CSV file namely Scikit. The library was used to preprocess data and implement ML models. Additionally, CSV files are interpreted using the Python program, i.e., Pandas. The CSV files are then converted into data frames, which are then analysed with SciKit29. In the end, 14 labels were added in the first column of data frames. A total of 9 features were extracted namely, mean, median, mode, standard deviation, variance, minimum, maximum, and high order moments, such as skewness and kurtosis. The final data is fed to different ML algorithms, namely Random Forest, K-NN, SVM, Logistics Regression, and SVM RBF.

Classification via Machine learning Models

For classification, the RSSI information collected in the previous step is fed into ML models. Three different ML models are considered for this purpose: Random Forest, k-NN, and SVM(RBF). The high-level signal flow diagram of the proposed lip- reading recognition system is illustrated in Figure. 3.11. Our classification framework differentiates different groups of English structures such as vowels, consonants, and Words. The next subsections provide a detailed description of the ML models used in this chapter.

Random Forest: A random forest is a cutting-edge ML classifier for classifying numeric datasets[156]. In order to fit various decision tree classifiers on various subsamples of the dataset, it implemented a meta-estimator to increase predicted accuracy and manage over-fitting.

ML Model	Parameter	Setting
Random Forest	N Estimators	200
	CV	10
	Criterion	gini
	Min Sample Split	2
	Max Feature	Sqrt
	Min Sample	1
K-Nearest Neighbours	N Neigbors	3
	CV	10
	Weights	Uniform
	Leaf Size	30
	P	2
	Metric	Minowski
SVM RBF	Gamma	Auto
	Kernel	Rbf
	C	6.7
	Degree	3
	Cache Size	200

Table 3.11: Selected model parameter configurations.

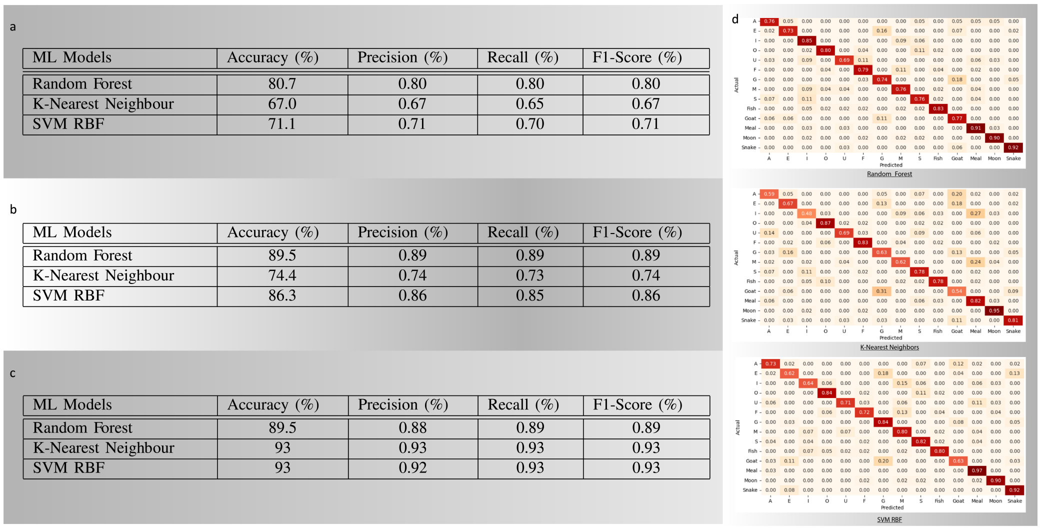
Table. 3.11 presents the hyper-parameter settings used for the Random-Forest model.
K-Nearest Neighbors(k-NN): This is a well-known decision rule that is commonly used in pattern classification [157]. In this technique, the ideal choice of the value of k was largely data-dependent; generally, a bigger k decreases the effects of noise but makes the classification boundaries less distinct. The hyper-parameter settings of k-nearest neighbor are shown in Table. 3.11.
SVM RBF: RBF kernel SVM used gamma and C parameters [158], where the gamma parameter defines how far a single training example's influence reaches, with low values indicating "far" and high values indicating "close." The C parameter trades off correct training example classification against maximisation of the decision function's margin. The hyper-parameter settings of SVM RBF are shown in Table. 3.11

3.3.3 Experiments and Results

This section focuses on the dataset description as well as system evaluation using the ML models previously discussed.

Dataset

A collection of RSSI values was produced as a result of the earlier described data collection and pre-processing phases. The dataset contains 2800 samples from 14 different categories/classes. These classifications are divided into three groups: (i) vowels, (ii) consonants, and (iii) words.



a

ML Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	80.7	0.80	0.80	0.80
K-Nearest Neighbour	67.0	0.67	0.65	0.67
SVM RBF	71.1	0.71	0.70	0.71

b

ML Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	89.5	0.89	0.89	0.89
K-Nearest Neighbour	74.4	0.74	0.73	0.74
SVM RBF	86.3	0.86	0.85	0.86

c

ML Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	89.5	0.88	0.89	0.89
K-Nearest Neighbour	93	0.93	0.93	0.93
SVM RBF	93	0.92	0.93	0.93

Figure 3.15: Experiment results of different ML models for the classification of lip-reading. (a) Vowels. (b) Consonants. (c) Words. (d) The combined result of all fourteen classes.

The vowels group consists of the five classes A, E, I, O, and U. The second group consonants are F, G, M, and S, while the last group is made up of words from the Fish, Goat, Meal, Moon, and Snake classes. Each of these groups has classes with an equal amount of samples. The dataset of each group was divided into two subsets: training and testing. In the vowels and words, dataset 300 samples were used as test and 700 samples for training. In the same way, the consonants dataset was divided into two parts, 240 samples for testing and 560 for training. All classes and subjects are represented equally in the training and testing sets. The University of Glasgow's Research Ethics Committee (permission numbers: 300200232, 300190109) received ethical approval for these experiments.

Results and Discussion

The experimentation in this work serves two purposes. In the first, we introduced RFID based smart mask for lip-reading recognition, and in the other, we compared the performance of various existing ML models. We collected and analysed the different sub-categories of English structure datasets such as vowels, consonants, and words taken from the diverse genders on the performance of RFID-based lip-reading frameworks. As a result, we conducted three distinct experiments on RSSI-captured data to evaluate the performances of the models. Table. 3.11 contains the hyper-parameter settings for all models. On the dataset, all of the models are fine-tuned. Furthermore, all experiments used fixed training and testing sets. Our training and testing sets contain 80% and 20% of total data, respectively. Figure. 3.15 displays the experimental results of experiments conducted with various English language structures in terms of precision, recall, and F1 Score. Overall better results were achieved for the combined and individual groups for all the models.

In the case of the vowels dataset, we calculated different diverse subjects' results which included females and males. In terms of subject (S1), Subject (S3), and Subject (S2) have high accuracy using the SVM RBF algorithm around 97.18%, 91.96%, and 83.13% as compared to other ML algorithms. Subject (S4) has also high classification accuracy of around 79.91% using Random Forest. In the combined RFID lip-reading vowels dataset of all females and males, we got high classification accuracy of 80.0% with precision, recall, and F1-score using Random Forest algorithms which are shown in Figure. 3.15a.

Similarly, consonant datasets namely F, G, M, and S were collected by diverse groups of subjects. In terms of Subject (S1) and Subject (S4) have high classification accuracy using Random Forest, k-NN, and SVM RBF algorithm around 97.18% and 97.48% as compared to other ML algorithms. Subject (S2) and Subject(S3) have a high accuracy of 86.43% using Random forest and k-NN, than other proposed ML algorithms. In combined RFID consonant datasets, we got high classification accuracy of 89.5% using Random Forest algorithm and has high precision, recall, and F1-score to other ML algorithms which are shown in Figure. 3.15b.

In the case of words datasets namely Fish, Goat, Meal, Moon, and Snake were collected by multiple subjects. K-NN algorithm has high classification accuracy using subject (S1) and subject (S3) datasets compared with other ML algorithms with around 89.15% accuracy. In terms of subject (S2) and subject (S3) a high classification accuracy is achieved using the SVM RBF algorithm which is around 95.58% and 97.18%. RFID words combined datasets got 93.0% classification accuracy along with high precision, recall, and F1-score using k-NN and SVM RBF algorithms which are shown in Figure. 3.15c.

Lastly, the confusion matrix of the combined dataset is shown in Figure. 3.15d. Three different ML models were applied to RSSI information namely Random Forest, k-NN, and SVM RBF. In the case of Random Forest most of the classes are correctly recognised except "U" because it performed similarly to "F". Here again, most of the classes are correctly classified using the k-NN algorithm except "I" which has misclassified with "S". Furthermore, the confusion matrix of SVM RBF mostly classifies except two classes "Goat" and "I". Overall, all three algorithms correctly classified 14 classes but Random Forest outperformed other with 80.0% classification accuracy.

3.4 Summary

This chapter presents results from a study on lip-reading using RF sensing technologies like Wi-Fi, radar, and RFID. Wi-Fi signals were generated using a USRP x300, which identifies human lip movements across various classes by utilising CSI signals. For radar technology, a UWB radar sensor, Xethru X4M03, was used to plot reflected Doppler signals (Hz) in frequency-time diagrams, such as spectrograms. Passive RFID technology was used to collect RSSI data, displaying time on the x-axis and power changes in dBm on the y-axis. The proposed RF sensing techniques can function independently or assist hearing aids by reading lip and mouth movements, especially in situations where face masks obstruct visual cues in vision-based systems. A diverse dataset, including male and female participants, was collected. The primary goal of the paper was to propose a secure lip-reading system capable of identifying lip movements with different RF sensing technologies and ML/DL algorithms, even in the presence of a mask. Four algorithms – NN, SVM, Ensemble, and Naïve Bayes – were evaluated on the Wi-Fi dataset using train-test methods. The highest classification accuracy observed was 93.3% on datasets involving males without face masks. Additionally, DL pre-trained models like VGG16, VGG19, and InceptionV3 were tested with radar data, achieving a maximum average accuracy of 91.67% on male data without masks. With RFID, RSSI data was input into various ML models, including Random Forest, k-Nearest Neighbors (k-NN), and SVM RBF. The system successfully classified lip movements, reaching a 100% accuracy rate. Furthermore, this current system serves as a proof of concept, demonstrating the importance and effectiveness of lip detection using RF-sensing technologies such as radar, Wi-Fi, and RFID. The next chapter explores 15 different British Sign Language recognition systems using UWB radar, providing an alternate perspective on communication with the deaf community through hearing aids.

Chapter 4

Deep Learning for British Sign Language Recognition: A Non-Contact and Privacy-Preserving Method

4.1 Introduction

The number of persons with hearing loss rises each day. Those who have trouble hearing often rely on SL for communication, depending on their level of disability. Sign language is used by millions of deaf people around the world to communicate and is imperative for their social integration. Similar to spoken languages, different parts of the world use different versions of SL. The importance of SL with its origin is discussed in [159, 160], with examples from American, British, Japanese, Chinese, and Arabic. Automatic sign language recognition is a challenging research area and is still in its early stages compared to speech recognition, despite of many efforts done in the literature using wearable or vision-based approaches. Wearable systems like hand gloves are intrusive and limit continuous human-computer interaction as they require users to wear sensors continuously. Camera-based systems also have fundamental flaws, such as privacy concerns due to recorded videos/images and the dependency on ambient lighting, which limits their effectiveness in SLR. Recently, RF sensing has emerged as a viable alternative in assisted living and contactless activity monitoring as it does not require end-users to wear or carry devices. The use of existing communication devices and infrastructure, like Wi-Fi routers, further enhances its applicability. To overcome the drawbacks of camera-based systems, radar-based sensing systems have been proposed. These systems are not affected by ambient lighting variations and ensure user privacy. They function by exploiting the unique doppler signatures created by hand movements.

Although RF sensing has been partly discussed in the literature to recognise SL, the existence of a diverse dataset that includes samples from a wide range of subjects (diverse age and sex) and covers diverse number of classes is missing in the literature. To bridge this gap, this work focuses on identifying different emotions, most common verbs, and family groups in BSL using micro-doppler signatures of the data collected using a radar sensor. Fifteen different types of doppler signatures are considered that include verbs (Drink, Eat, Help, Stop and Walk), emotions (Confused, Depressed, Happy, Hate and Sad) and Family Group (Family, Brother, Father, Mother and Sister). These categories of BSL signs are represented through dynamic SL, which utilises mobility or movements of the hands to represent various signs. An UWB Radar, XeThru X4M03 was used for recording the dataset. We note that the dataset is recorded at two different distances and angles. These characteristics make the dataset a better choice for training and evaluation of ML algorithms for BSL signs recognition and translation. The recorded data is represented in the form of spectrograms and further spatio-temporal features are extracted using GoogleNet and squeezenet CNNs.

The main contributions of the work are summarised as follows:

- We propose a contactless BSL recognition system that automatically recognises and translates BSL signs into verbs and emotions.
- The idea presented is of a future hearing aid device capable of capturing sign language to facilitate non-verbal communication for deaf individuals.
- We also collect a large-scale benchmark dataset containing a total of 1950 samples from 15 different types of BSL signs captured at distances of 141 and 154 cm. Moreover, the data samples are captured from two different angles. To ensure diversity, the data was collected from four deaf participants (1 male and 3 females) having ages between 16 and 82.
- We report experimental results of several state-of-the-art DL models on the dataset, providing a baseline for future research in this domain.

4.2 Methodology

Figure. 4.1 provides the block diagram of the methodology adopted in this work. The proposed framework is divided mainly into three phases. In the first phases, we collected and annotated a large collection of BSL signs. In the second phase, we used some signal processing techniques to extract spectrograms of various signs captured in the first phase. Finally, several DL models are employed to classify the signs.

In the next sub-sections, we provide a detailed description of each of the phases of the proposed methodology.

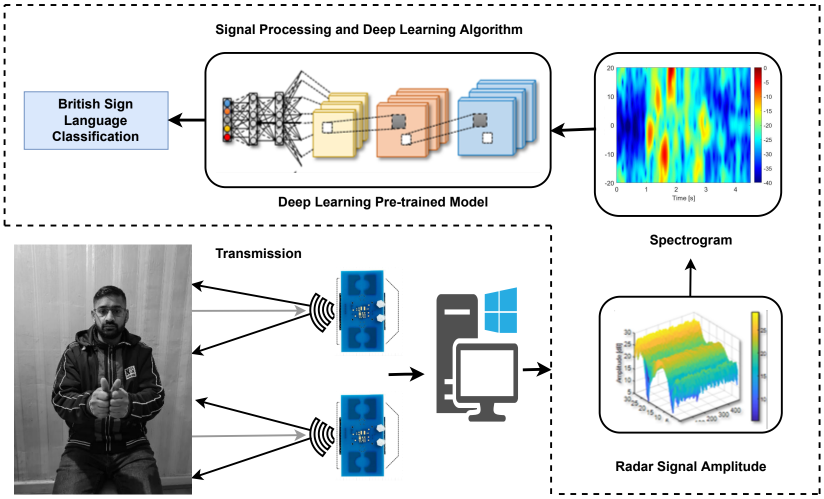
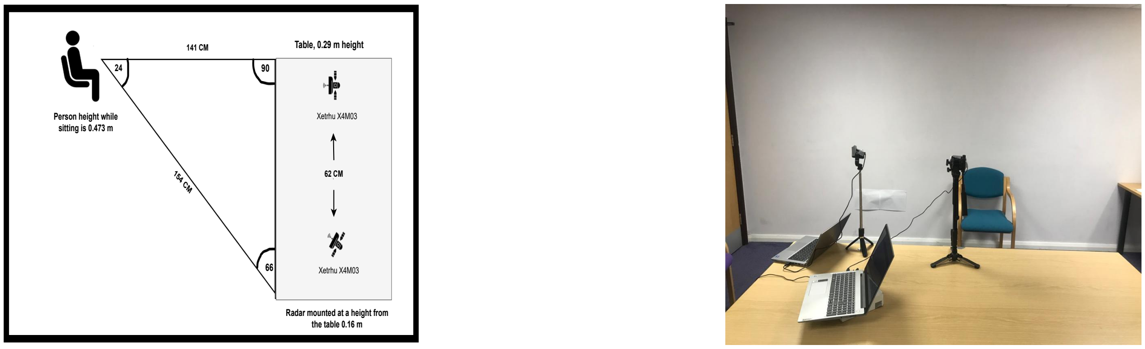

Signal Processing and Deep Learning Algorithm
British Sign Language Classification
Deep Learning Pre-trained Model
Spectrogram
Time [s]
Transmission
Radar Signal Amplitude

Figure 4.1: Block diagram of the proposed framework highlighting the UWB radar-based system, data collection, and the DL models for the classification of BSL signs.

4.2.1 Data Collection

In this phase, we collected BSL sign data through UWB radar. Figure. 4.2 provides an overview of the hardware setup of the radar-based BSL data collection system. To this aim, an UWB radar sensor, namely Xethru X4M03 is used. The Xethru X4M03 is a UWB radar sensor with built-in Tx and Rx antennas, providing a maximum detection range of 9.6 metres.


141 CM
Table, 0.29 m height
24
90
Person height while sitting is 0.473 m
Xetrhu X4M03
62 CM
154 CM
66
Xetrhu X4M03
Radar mounted at a height from the table 0.16 m

(a) Measurable Experimental-Setup

(b) Real Experimental Setup

Figure 4.2: Experimental setup of data collection using Xethru UWB radar sensor.

As shown in the figure, the radar sensor is placed on top of the screen of the laptop. The key parameter settings of the radar are indicated in Table. 4.1. In order to differentiate complexity levels in the dataset, the radar sensor was placed at two different distances including a distance of 141 and 154 cm from the subject (i.e., person). Moreover, the sign gestures of the subject are recorded at two different angles. The variations in the distance and viewpoint are expected to help in training distance and viewpoint invariant DL models that are able to recognise gestures of the subjects from different distances and angles. During the data collection, the body of the subject was in normal position with head and hands movements only. Moreover, the duration of each activity was set to 5 seconds involving the data collection of a single gesture from a single subject. Figure. 4.3 provides visual illustration of the pronounced BSL.

Parameter	Value
Platform	Xetru radar X4MO3
Instrumental range	9.6 meters
Target's distance from radar	141 cm and 154 cm
Operating frequency	7.29GHz
Transmitter power	6.3dBm
Activity duration	5 seconds
Collected samples in each class	15

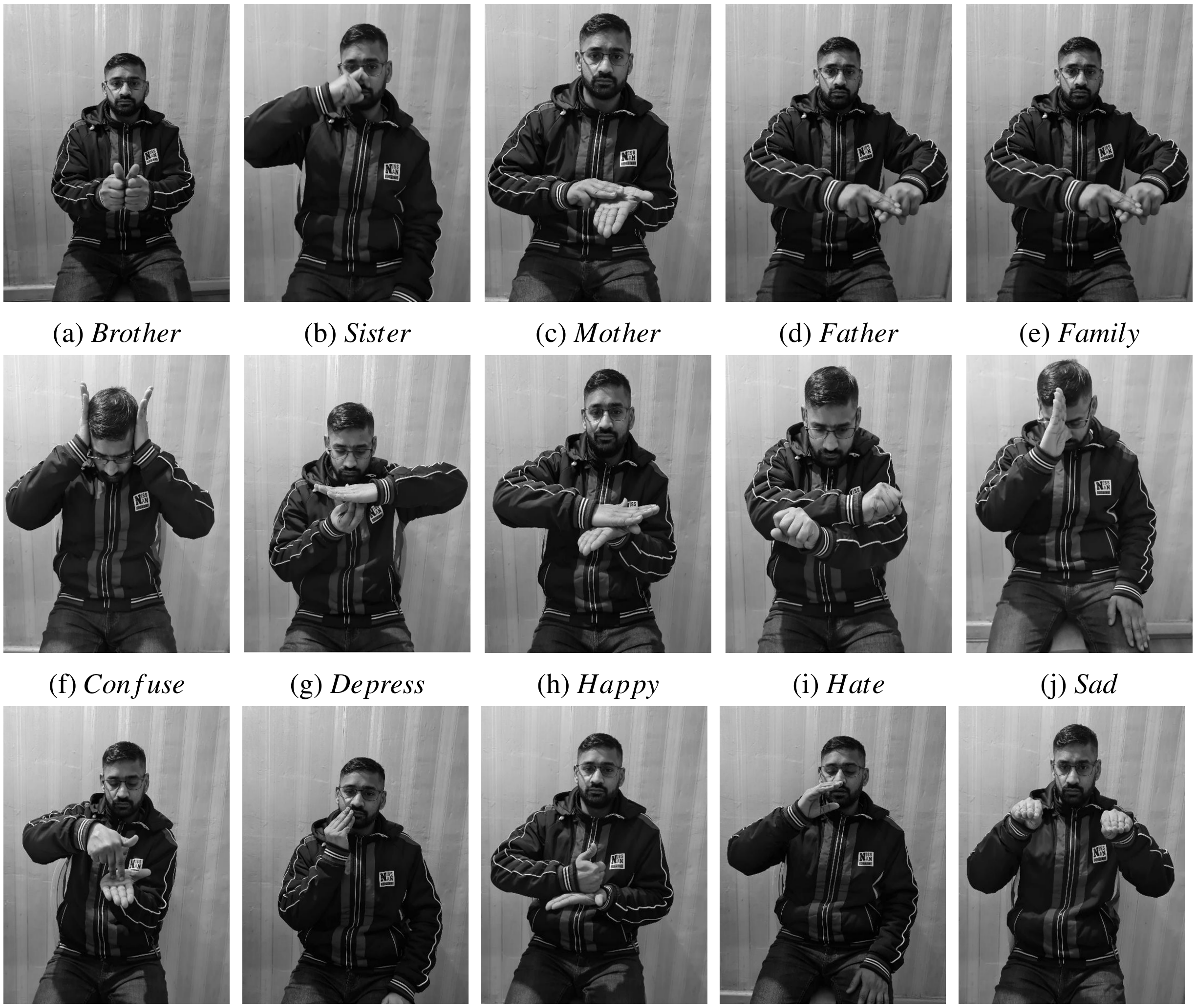
Table 4.1: Parameters configuration of radar software and hardware.

Moreover, four deaf subjects/participants, one male and three females participated in the data collection process. The reason to include more participants was to make the dataset more realistic and diverse. A total of 1950 data samples were collected during experiment for 15 different categories at distances of 141 cm and 154 cm. The details of the collected dataset are highlighted in Table. 4.2. In each experiment, a total of 975 data samples were collected from four deaf participants, where 15 samples were collected in each class. In particular, each participant repeated the speaking activity of each gesture 15 times with the radar. In this way, each participant contributed to collect 225 or 300 data samples in total for fifteen classes. In each case, A total of 975 spectrograms were categorised as fifteen BSL signs, with 750 being utilised for training and 225 for testing.

After capturing the data, the RF signal was transmitted and received from the radar within this duration. The details of all components presented in the figure are discussed later in this section.

4.2.2 Pre-Processing

In this section, we describe the pre-processing steps carried out to extract the required spectrograms. In the beginning, the radar chip was configured via the XEP interface with x4driver. Data were recorded from the module at 500 frames per second (FPS) (in the form of the float message data), where each value is a 32-bit floating point number. A loop was used to read the data file and save the data into a DataStream variable, which was mapped into a complex (range-time-intensity) Range-Time Intensity (RTI) matrix. Thereafter, MTI filter was applied to get the doppler range map. Afterwards, the second MTI was used as a Butterworth 4^{th} order filter to generate the spectrograms using the following parameters: window length, overlap percentage, and FFT padding factor. In particular, a window length of 128 samples and a padding factor of 16 was used. In addition, a range profile was created by first converting each chirp to an FFT. A second FFT is then conducted on a defined number of consecutive chirps for a given range bin.



(a) *Brother* (b) *Sister* (c) *Mother* (d) *Father* (e) *Family*

(f) *Confuse* (g) *Depress* (h) *Happy* (i) *Hate* (j) *Sad*

(k) *Walk* (l) *Eat* (m) *Help* (n) *Drink* (o) *Stop*

Figure 4.3: A visual illustration of the pronounced british sign language.

Classes	**Experimental Dataset**								
	141 cm				**154 cm**				**Total**
	Subject (S1)	**Subject (S2)**	**Subject (S3)**	**Subject (S4)**	**Subject (S1)**	**Subject (S2)**	**Subject (S3)**	**Subject (S4)**	
Brother	15	15	20	15	15	15	20	15	**130**
Sister	15	15	20	15	15	15	20	15	**130**
Mother	15	15	20	15	15	15	20	15	**130**
Father	15	15	20	15	15	15	20	15	**130**
Family	15	15	20	15	15	15	20	15	**130**
Confused	15	15	20	15	15	15	20	15	**130**
Depressed	15	15	20	15	15	15	20	15	**130**
Happy	15	15	20	15	15	15	20	15	**130**
Hate	15	15	20	15	15	15	20	15	**130**
Sad	15	15	20	15	15	15	20	15	**130**
Stop	15	15	20	15	15	15	20	15	**130**
Walk	15	15	20	15	15	15	20	15	**130**
Eat	15	15	20	15	15	15	20	15	**130**
Help	15	15	20	15	15	15	20	15	**130**
Drink	15	15	20	15	15	15	20	15	**130**
Total	**225**	**225**	**300**	**225**	**225**	**225**	**300**	**255**	**1950**

Table 4.2: An overview of the data collected, number of subjects and the activities performed.

Furthermore, a STFT was used to create these spectrograms, because, unlike fourier transform, it offers both temporal and frequency information [161]. This is done by segmenting the data and then performing fourier transform on each segment. When the window length is changed, both the temporal and frequency resolutions are altered inversely. For example, if one increases the other decreases. The level of doppler detail in radar data is determined by the hardware's sampling capability. The greatest unambiguous doppler frequency in radar is $F_d, max = \frac{1}{2}t_r$, where t_r is the chirp time. In this chapter, we look at BSL gestures recognition at a distance D(t) from a specified location, such as the subject's hand and faces. V(t) represents the point of target movement in front of the radar, and T_s represents the transmitted signal,

$$T_s(t) = A\cos(2\pi ft). \tag{4.1}$$

The received signal is provided by Rs(t),

$$R_s(t) = \acute{A}\cos(2\pi f(t - \frac{2D(t)}{c})), \tag{4.2}$$

where A is the reflection coefficient, and c is the speed of light. The reflected signal can be expressed as $R_s(t)$, where the signal reflected off the target points at an angle θ to the direction of radar.

$$R_s(t) = \acute{A}\cos(2\pi f(1+\frac{2v(t)}{c})t - \frac{4\pi D(\theta)}{c})). \tag{4.3}$$

The doppler shift that corresponds to it can be written as,

$$f_d = f\frac{2v(t)}{c}. \tag{4.4}$$

The returned signal becomes a composite of several moving elements such as the head and hand. Each component moves at its own speed and acceleration. If we consider i to be the various moving components of the hands, we write the received signal as

$$R_s(t) = \sum_{i}^{N} A_i\cos(2\pi f(1+\frac{2vi(t)}{c})t - \frac{4\pi D_i(0)}{c})). \tag{4.5}$$

The doppler shift is the result of a complex interaction of numerous doppler shifts induced by moving hands and head. Detection of SL in a reliable fashion clearly depends upon the characteristics of the doppler signatures. After obtaining the spectrograms of various signs from the participants, datasets were labeled with 15 different signs. As indicated in the high-level signal flow diagram in Figure. 4.1, the dataset is consisted of two key modules: (i) system training and (ii) system testing. The proposed pre-trained DL classification algorithms were implemented on spectrograms to recognise the BSL dataset.

4.2.3 Classification via Deep Learning Models

The data in the form of spectrograms considered as images is why we used DL models. The spectrograms generated and labeled in the previous step are now fed into DL models for classification purposes. For this purpose, three different pre-trained models, namely VGGNet, GoogLeNet, and SqueezeNet are considered. Our classification framework, designed to differentiate between BSL signs/activities, primarily relies on fine-tuning pre-trained models. Here, multiple state-of-the-art CNN architectures, which were pre-trained on ImageNet [162], are fine-tuned on the spectrogram images generated from radar data. In fine-tuning the pre-trained models, we used data augmentation to overcome the overfitting along with modifying the top layers of the models to classify the collected data into fifteen considered classes, namely Verbs (Drink, Eat, Help, Stop and Walk), Emotions(Confused, Depressed, Happy, Hate and Sad) and Family Group (Family, Brother, Father, Mother and Sister).

In the following subsections, we provide a detailed description of the CNN architectures used in this work.

GoogLeNet Model: GoogLeNet [163] is one of the state-of-the-art and commonly used CNN architecture for different image classification tasks [164]. The architecture is composed of 22 layers including convolutional, pooling layers, inception modules, and a fully connected layer. The inception module is made up of 6 convolutional layers and a pooling layer. The module consists of patches or filters of sizes 1 × 1, 3 × 3 and 5 × 5. These filters of different sizes help to obtain different patterns of the input image. The feature maps obtained from various filters are concatenated at the output of each module. Furthermore, 1 × 1 convolutions are performed prior to convolutions by large filters. The use of 1 × 1 convolution filter decreases the number of parameters required by GoogLeNet. The hyper-parameter settings of GoogleNet are shown in Table. 4.3.

SqueezeNet Model: Our second pre-trained model is based on SqueezeNet architecture [165], which is composed of 18 layers. This architecture has shown comparable results with fifty times fewer parameters, which makes it a preferable choice for applications with fewer data and low computational resources. Squeezenet adapts to three major strategies. The first strategy reduces the 3×3 filters to 1×1 filters given in the squeeze layer. The second strategy uses expand layer in which 1×1 and 3×3 filters are fed with less input parameters from the squeeze layer. The third strategy down-samples late (having smaller stride values), so that the last layer has larger activation maps which results in better accuracy.The parameter settings of SqueezeNet are shown in Table. 4.3.

VGG16 Model: Another pre-trained model is based on VGG16 architecture [166], which is composed of 16 layers. This architecture contains a total of 138 million parameters, which used a 3x3 filter size with a stride 1 and always use the same padding and max-pooling layer of a 2x2 filter with stride 2. The arrangement of the layers in this architecture is as follows convolutional layers, ReLU layers, and max pool layers. ReLU is more computationally efficient because it results in faster learning and it also decreases the likelihood of vanishing gradient problems. The end of the model has 3 fully connected layers followed by a softmax for output. The parameter settings of VGG16 are shown in Table. 4.3.

4.3 Experiments and Results

This section elaborates on the data description with its distribution, system evaluation, and the obtained classification results from all considered pre-trained models.

4.3.1 Dataset

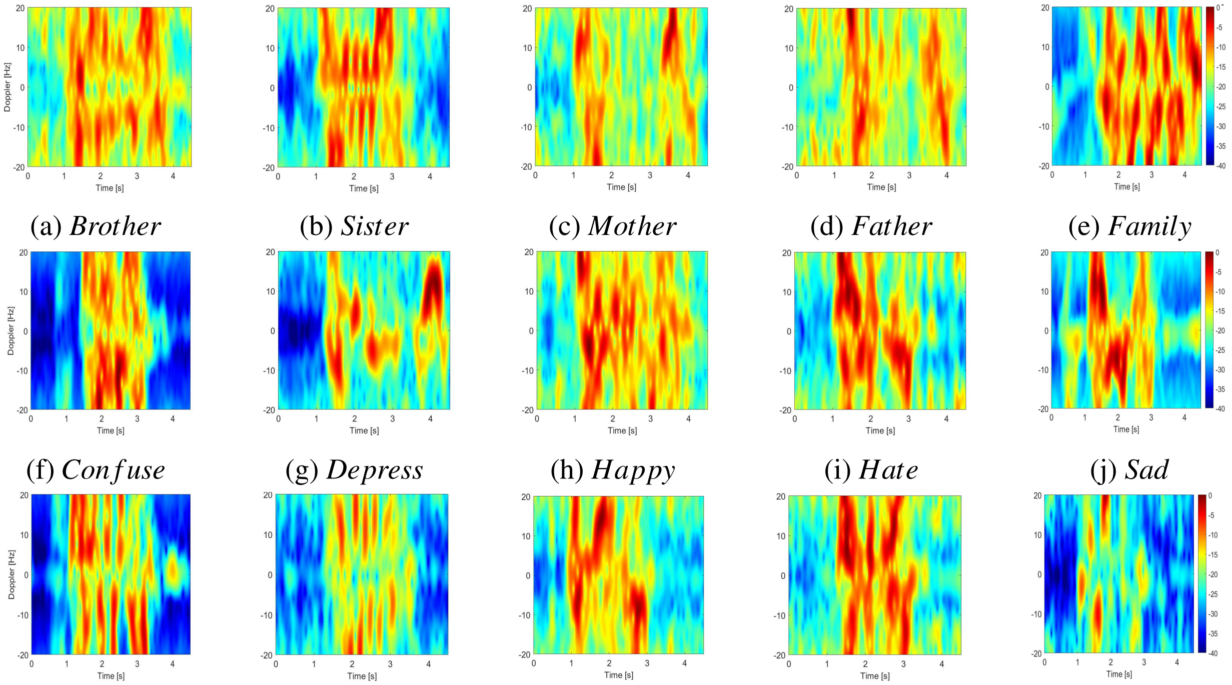
The data collection and pre-processing phases described earlier resulted in a collection of spectrograms. In total, the dataset is composed of 1950 samples from 15 different categories/classes. These classes are sub-grouped into three groups, namely (i) verbs, (ii) emotions, and (iii) family. The verbs group includes five classes namely Drink, Eat, Help, Stop and Walk. The emotions include Confused, Depressed, Happy, Hate and Sad while the final group is made of Family,

DL Model	Parameters	Settings
GoogleNet	Initial learning rate	0.0001
	Mini-batch size	128
	Learning algorithm	Adam
	Loss function	Cross entropy
	Maximum epochs	100
	Iteration per epoch	500
	Elapsed time for 141 cm	04:35:57
	Elapsed time for 154 cm	06:15:21
SqueezeNet	Initial learning rate	0.0001
	Mini-batch size	128
	Learning algorithm	Adam
	Loss function	Cross entropy
	Maximum epochs	100
	Iteration per epoch	500
	Elapsed time for 141 cm	01:41:53
	Elapsed time for 154 cm	02:03:28
VGG16	Initial learning rate	0.0001
	Mini-batch si	16
	Learning algorithm	Adam
	Loss function	Cross entropy
	Maximum epochs	100
	Iteration per epoch	46
	Elapsed time for 141 cm	41:32:11
	Elapsed time for 154 cm	45:03:58

Table 4.3: Parameter settings for the selected models.

i.e., Brother, Father, Mother, and Sister classes. Each of these fifteen classes contain an equal number of sample. Figure. 4.4 provides some samples images/spectrograms from the dataset.

The dataset has been divided into two subsets namely training and testing set. The training set is composed of 1560 samples while the testing set provides a total of 390 samples. These subset have equal representation from all the classes and subjects.



(a) *Brother* (b) *Sister* (c) *Mother* (d) *Father* (e) *Family*

(f) *Confuse* (g) *Depress* (h) *Happy* (i) *Hate* (j) *Sad*

(k) *Walk* (l) *Eat* (m) *Help* (n) *Drink* (o) *Stop*

Figure 4.4: Obtained spectrum's sample of (a) Brother, (b) Sister, (c) Mother, (d) Father, (e) Family, (f) Confused, (g) Depress, (h) Happy, (i) Hate, (j) Sad, (k) Walk, (l) Eat, (m) Help, (n) Drink, (o) Stop signs.

4.3.2 Evaluation Matrics for Classification Model

In this chapter, the performance of the DL models in classification of BSL signs is evaluated in terms of weighted average accuracy, precision, recall, and F1 Score. F1 Score is one of the most commonly used metrics in the literature for classification, which is calculated using Equation 6.2. F1 Score is a combination of precision and recall, which are calculated using equation 6.3 and 6.4, respectively.

$$F1 = 2\frac{Precision.Recall}{Precision + Recall} \tag{4.6}$$

$$Precision = \frac{\sum TruePositive}{\sum TruePositive + \sum FalsePositive} \tag{4.7}$$

Model	141 cm			154 cm		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
GoogleNet	0.81	0.87	0.83	0.78	0.87	0.80
SqueezNet	0.75	0.83	0.77	0.69	0.76	0.71
VGGNet16	0.90	0.92	0.91	0.86	0.89	0.87

Table 4.4: An evaluation of the pre-trained models in terms of macro-recall, macro precision, and macro-F1-score on the datasets captured at distances of 141 and 154 cm.

$$Recall = \frac{\sum TruePositive}{\sum TruePositive + \sum FalseNegative} \tag{4.8}$$

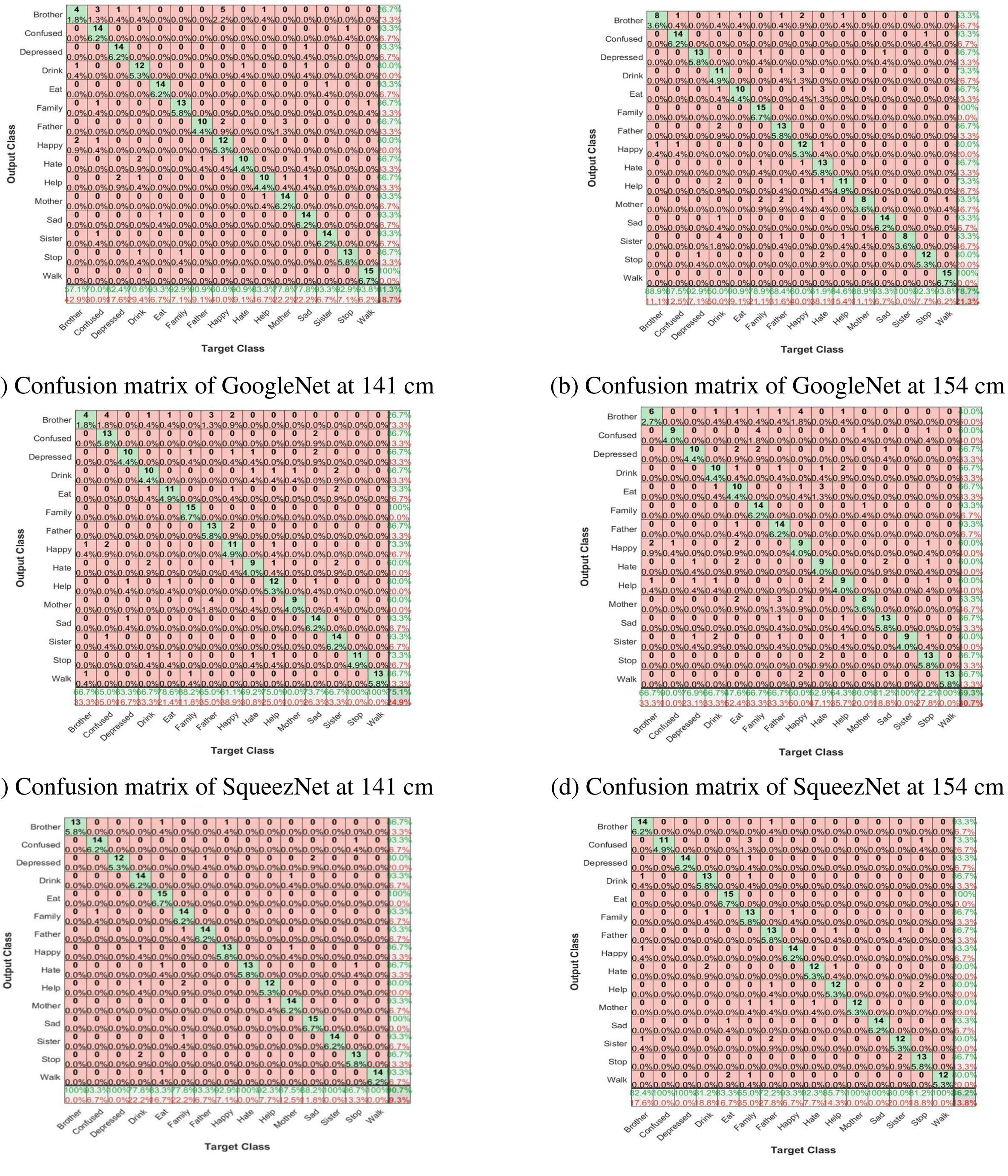
4.3.3 Results and Discussion

The objectives of the experimentation in this chapter are two-fold. On one side, we analyse the performance of different existing pre-trained models on the newly collected BSL dataset. On the other hand, we analyse the impact of variations in viewpoint and distances from the subject on the performance of BSL frameworks. Therefore, we conducted two different experiments by evaluating the performances of the models on spectrograms captured at distances of 141 and 154 cm.

The hyper-parameter settings for all the models are provided in Table. 4.3. All the models are fine-tuned on the dataset for a maximum of 100 epochs. Moreover, in all the experiments, fixed training and testing sets are used. Our training and testing sets are composed of 80% and 20% of the total data, respectively.

Table. 4.4 provides the experimental results conducted at distances of 141 and 154 cm in terms of precision, recall, and F1 Score. As expected, overall better results for all the models are obtained when the radar sensor is placed at a distance of 141 cm compared to 154 cm. This is due to the reason that the 141cm radar is placed exactly in front of the subject and the micro-doppler signature at this view points are more sensitive to hands movements as compared to what is received at 151cm radar for the same movements. As a result, the ML model better classify the hand movements as this viewpoint. As far as the performances of the pre-trained models is concerned, overall better results are obtained with VGGNet achieving an F1 Score of 0.87.

In order to better analyse the performances of the models, we also provide confusion matrices for each model at each distance in Figure. 4.5. Figure. 4.5a illustrates the confusion matrix of the GoogleNet model with 141 cm distance. It can be observed from the figure that most of the classes are correctly classified having a lowest classification accuracy of 0.26% for Brother class, which is mostly confused with the class Happy. Similarly, the confusion matrix of the GoogleNet at a distance of 154 cm is presented in Figure. 4.5b, where the classification accuracy is nearly 100% for all classes except the classes Brother, Mother, and Sister. The samples from the class Brother are mostly (around 0.13%) confused with the class Happy. The samples from



(a) Confusion matrix of GoogleNet at 141 cm

(b) Confusion matrix of GoogleNet at 154 cm

(c) Confusion matrix of SqueezNet at 141 cm

(d) Confusion matrix of SqueezNet at 154 cm

(e) Confusion matrix of VGGNet at 141 cm

(f) Confusion matrix of VGGNet at 154 cm

Figure 4.5: The confusion matrices of all the models at two different distances.

the class Mother confused with the samples from the classes Father and Family. This may be because of the reason that both right and left-hand fingers move in the same manner. Similarly, the class Sister has resemblance with class Drink as the participants used their right-hand index finger and thumb to touch the nose.

Similarly, the confusion matrix of SqueezeNet with a distance of 141 cm is presented in Figure. 4.5c. Here again, most of the samples from all the classes are correctly classified with the exception of Brother, which shows similarities with class Confused gestures because both the classes gesture right-hand and left-hand rub with each other between the head or without a head. Furthermore, the confusion matrix of SqueezeNet at a distance of 154 cm is presented in Figure. 4.5d. In this case, most of the classes are correctly classified with an exception of class Brother and Mother, which show similarities with the classes Happy and Father. However, 80% of test samples are correctly classified for the class Sad with only 20% matching confusion with the class Depressed.

Likewise, the confusion matrix of classifying the considered emotions using VGG16 with a distance of 141 cm is presented in Figure. 4.5e. Here again, most samples from all classes are correctly classified, with the exception of Depressed, which shows similarities with the Sad class because in both classes, the front mouth makes a downward movement. On the other side, Help is similar to the Family because both signs have upturned hands while performing the activity. Furthermore, the confusion matrix of VGG16 at a distance of 154 cm is presented in Figure. 4.5f. In this case, most of the classes are correctly classified with an exception of class Confused, which shows similarities with the classes Family and Stop. However, 70% of test samples are correctly classified for the class Confused with only 20% matching confusion with the class Family and 10% with the class Stop.

4.4 Summary

This chapter presented a privacy-preserving BSL recognition system using state-of-the-art XeThru X4M03 UWB radar sensors and DL algorithms. In BSL, the fifteen most common gestures are categorised into classes, namely Verbs (Drink, Eat, Help, Stop, and Walk), Emotions(Confused, Depressed, Happy, Hate, and Sad), and Family Group (Family, Brother, Father, Mother, and Sister). For each class, micro-doppler unique features were stored in the form of spectrograms. These were used to train two DL models, namely GoogleNet and SqueezeNet models. The classification accuracy for most of the classes was close to 100% with the GoogleNet model outperforming others, giving an overall accuracy of 81.33% on all fifteen classes. The next chapter presents another application of hearing impairment devices designed to assist deaf people in identifying the behavior of others, both with and without wall obstacles.

Chapter 5

Wi-Fi and Radar Fusion for Head Movement Sensing Through Walls Leveraging Deep Learning

This chapter highlights the potential of non-invasive head movement recognition for various applications, including human behavior identification, controlling assistive technologies like wheelchairs for quadriplegics, virtual/augmented reality, and assistive driving. Wearable and vision-based devices face specific limitations, including challenges with ambient lighting, line of sight, and privacy concerns. Additionally, wearing these devices can often be uncomfortable. Wi-Fi and radar offer contactless sensing applicable in different applications through-wall scenarios. We propose a contactless system using UWB radar and Wi-Fi signals, combined with ML and DL techniques to detect human head movements. Our study focuses on six common head gestures (head down, head up, head left 90, head right 90, head right 45, and head right 90) using time-frequency multi-resolution analysis based on wavelet scalograms for feature extraction from channel state information values and radar signal spectrograms. By fusing features from both radar and Wi-Fi signals and employing DL models like VGG16 and InceptionV3, we achieve high classification accuracies of 83.33% and 91.8% for head movement detection with and without walls, respectively.

5.1 Introduction

Head movement [167] carries important information related to human behavior which is an integral part of non-verbal communication and has a wide range of applications for human-computer interaction, such as assistive technologies, virtual and augmented reality, and assistive driving systems. Head movement detection has been widely utilised for assistive driving of wheelchairs for patients suffering from paralysis, driver drowsiness detection and alert systems. Intelligent assistive driving systems can reduce the number of road accidents by monitoring driver's behavior through head movements and generate alerts accordingly. Mental tiredness impairs focus when driving and has major safety implications [168, 169]. Poor sleep and tiredness are major causes of poor driving performance, steering mistakes, loss of vehicle control, and deadly accidents [170, 171, 172, 173]. Driving assistance systems rely heavily on the detection of driver attentiveness. The orientation of the driver's head may reflect his degree of attention. Head movement is getting high popularity in assistive driving since an estimated 1,560 reported road deaths in 2021 in the UK [2].

In recent years, there has been an increase in assistive technologies in healthcare and many other domains that benefit from smart technology concepts. Head movement detection has proven to be effective in many applications such as the detection of driver's fatigue [174], human visual focus [175], behavior recognition [176], vitals monitoring [177], healthcare cognitive assistance [178], in figuring out the human head kinematics [179] to estimate and predict possible head collision injuries in athletes, and in clinical depression monitoring [180], etc.

The current head movement detection systems primarily utilise camera-based and wearable technologies, each with significant limitations. Camera-based systems face privacy concerns due to the necessity of recording the target, legal implications that may limit their use in both public and private spaces, and the potential for being perceived as photographing someone without consent, which is illegal in many jurisdictions. Additional challenges include obstructions to the line of sight and difficulties in training with long video sequences. Wearable devices, on the other hand, disrupt daily routines due to their intrusive nature.

RF head movement sensors, on the other hand, can fulfill the demand for next-generation technologies. By recognizing head motions using RF sensing, ML, and DL techniques, various applications can be benefited from very accurate cues. Moreover, unlike vision-based systems, RF sensing-based head movements are unaffected by opaque barriers or walls separating the target and the transponder. RF signals can pass through walls to detect visual cues, such as head and lip movements. Head movements provide additional functionality for the next generation of MM hearing aid devices for understanding the behavior of people. In this chapter, we designed, developed, and tested an RF sensing-based method for detecting head motions with and without a wall. Activity monitoring through walls or barriers via Wi-Fi and radar devices is a great breakthrough in the field. Since cameras are limited to line-of-sight visuals they can not detect/sense any object or humans through walls/barriers. Therefore in this work, we proposed a radar and Wi-Fi-based novel system that can perform head-movements monitoring through walls and other opaque barriers.

This chapter focuses on recognizing different head movements and collecting data using micro-doppler signatures and CSI amplitude using a radar sensor and Wi-Fi signals. The existing dataset is diverse in nature that includes samples from a wide variety of subjects (ages and genders) and a diverse number of classes that cover all essential aspects of head movements. Head up, Head down, Right 90, Left 90, Right 45, and Left 45 are the six types of doppler signatures and CSI data considered for this work. These types of movements include dynamic gestures in which mobility or head are used to represent various movements. The dataset was recorded using two separate methods, i.e., using a Radar sensor and Wi-Fi signals with and without a wall. These features make the dataset a better option for the training and assessment of ML and DL algorithms for the recognition of head movements. In order to visualize the recorded data, spectrograms and CSI amplitudes were used.

The following presents the main contributions:

- This RF-based system provides the idea of a flexible hearing aid because it tracks lip movements along with the individual's head in different positions.

- We proposed a contact-less head recognition system that automatically recognises and translates head movements with and without a wall in between the target and transponder setup.

- In addition, we collect a dataset of 2400 samples from 6 different types of head movements captured at 0.50 centimeters distance away from the target. Furthermore, the data samples are collected using 2 different techniques (Radar sensor and Wi-Fi signals) with and without a wall. To ensure diversity, data was collected from four participants (two males and two females) ranging in age from 20 to 40 years.

- For the radar dataset, various DL models including VGG16, VGG19, InceptionV3, and SqueezeNet were applied. When tested on a combined dataset of four subjects, VGG16 outperformed the other algorithms, achieving 80% accuracy with the presence of a wall and 79.2% without it.

- For the Wi-Fi dataset, VGG16, VGG19, InceptionV3, SqueezeNet, Neural network pattern recognition, Tree(Medium Tree), and Ensemble(Boosted Tree) was applied on the individual subject, a combined dataset of four subjects InceptionV3 outperformed as compared to another algorithm 80% Accuracy with the wall and 89% without the wall.

- The fusion of features for different DL models was tested. The highest accuracy values of 91.8% without the wall was achieved with feature fusion of VGG16 and InceptionV3 DL models. Furthermore, the highest accuracy of 83.33% was achieved through the walls with the feature fusion of VGG16 and InceptionV3 DL models.

- In this work, we presented the experimental results from several state-of-the-art DL and ML models applied to our benchmark dataset, which can serve as a foundation for future research in the domain of detecting head movements through walls.

This chapter proposes novel head movement gestures using micro-doppler signatures using radar-sensor with and without walls. Six different gestures are considered, Head 45L, Head

45R, Head 90L, Head 90R, and Head Down. An ultra-wideband radar, XeThru X4M03 is used to record experimental data. The received data is represented in the form of spectrograms while spatiotemporal features were extracted using fusion of two different models. We achieved 91.8% of classification accuracy without a wall. The possible use cases of the proposed technology are illustrated in Figure. 5.1.

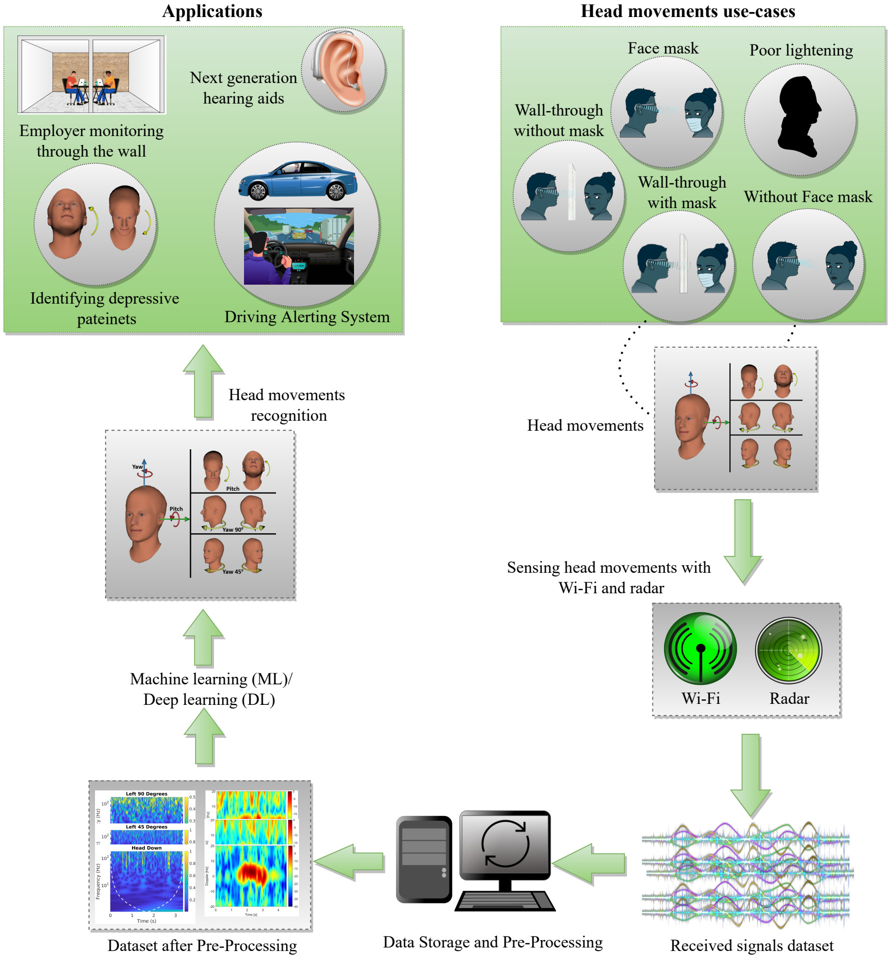
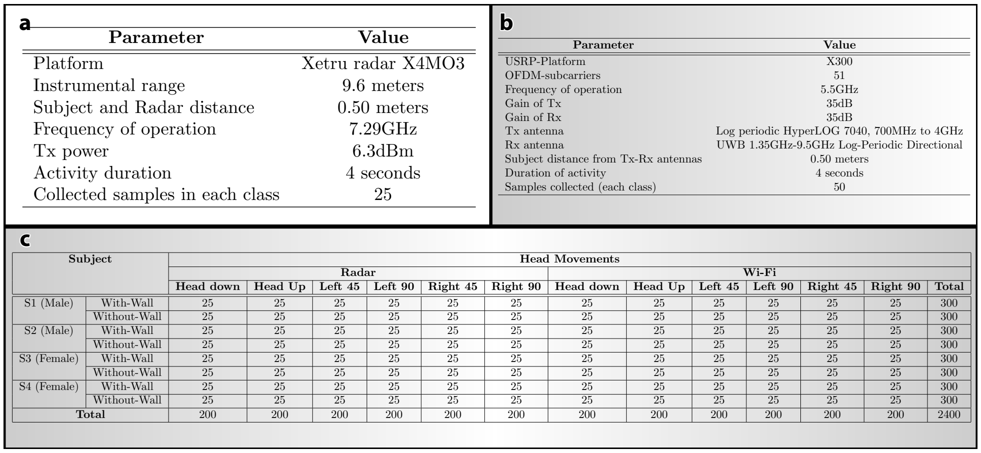

Applications
Employer monitoring through the wall
Next generation hearing aids
Identifying depressive pateinets
Driving Alerting System
Head movements use-cases
Face mask
Poor lightening
Wall-through without mask
Wall-through with mask
Without Face mask
Head movements recognition
Head movements
Yaw
Pitch
Yaw 90°
Yaw 45°
Sensing head movements with Wi-Fi and radar
Wi-Fi
Radar
Machine learning (ML)/ Deep learning (DL)
Left 90 Degrees
Left 45 Degrees
Head Down
Dataset after Pre-Processing
Data Storage and Pre-Processing
Received signals dataset

Figure 5.1: Conceptual representation of the suggested methodology for head movements.

5.2 Experimental Setup

5.2.1 Radar based setup

The experimental setup and configuration parameters for the radar-based head movement system are illustrated in Figure. 5.2a. The sensor has a 1.5 GHz sensor bandwidth and a detection range of about 9.6 meters. It utilizes a UWB radar sensor, specifically the Xethru-X4M03 model, equipped with both Tx and Rx antennas. To derive valuable insights from the radar data, we employed the STFT on the radar signal. This process resulted in the creation of spectrograms



a

Parameter	Value
Platform	Xetru radar X4MO3
Instrumental range	9.6 meters
Subject and Radar distance	0.50 meters
Frequency of operation	7.29GHz
Tx power	6.3dBm
Activity duration	4 seconds
Collected samples in each class	25

b

Parameter	Value
USRP-Platform	X300
OFDM-subcarriers	51
Frequency of operation	5.5GHz
Gain of Tx	35dB
Gain of Rx	35dB
Tx antenna	Log periodic HyperLOG 7040, 700MHz to 4GHz
Rx antenna	UWB 1.35GHz-9.5GHz Log-Periodic Directional
Subject distance from Tx-Rx antennas	0.50 meters
Duration of activity	4 seconds
Samples collected (each class)	50

c

Subject		Head Movements												
		Radar						Wi-Fi						
		Head down	Head Up	Left 45	Left 90	Right 45	Right 90	Head down	Head Up	Left 45	Left 90	Right 45	Right 90	Total
S1 (Male)	With-Wall	25	25	25	25	25	25	25	25	25	25	25	25	300
	Without-Wall	25	25	25	25	25	25	25	25	25	25	25	25	300
S2 (Male)	With-Wall	25	25	25	25	25	25	25	25	25	25	25	25	300
	Without-Wall	25	25	25	25	25	25	25	25	25	25	25	25	300
S3 (Female)	With-Wall	25	25	25	25	25	25	25	25	25	25	25	25	300
	Without-Wall	25	25	25	25	25	25	25	25	25	25	25	25	300
S4 (Female)	With-Wall	25	25	25	25	25	25	25	25	25	25	25	25	300
	Without-Wall	25	25	25	25	25	25	25	25	25	25	25	25	300
Total		200	200	200	200	200	200	200	200	200	200	200	200	2400

Figure 5.2: Head movements activity with their representation in Wi-Fi and radar signal. (a) The configuration parameters of radar software and hardware without and with through the wall experiment. (b) The configuration parameters of Wi-Fi software and hardware with and without the wall experiment. (c) An overview of the gathered data, the total number of participants, and the conducted activities.

that effectively captured the radar doppler shift corresponding to various head movements. Examination of these spectrograms revealed that different head movements produced distinct spectrogram patterns.

Scenario 1 - Line-of-sight: With no wall in between target and transponder setup

The sensor was placed in front of the participants/subject at around a half-meter distance. The experimental data recording activity for head movements was carried out by placing the radar 0.5 meters away from the subject sitting on a chair. The only movements performed here by the subjects were the head movements with slight shoulder movements which naturally arises while talking. The rest of the body was in a normal sitting position. Each activity was performed in a 4 seconds time frame. In these 4 seconds, the RF signal was transmitted and received back by the radar. The data collection and processing using UWB radar setup are shown in Figure. 5.3a.

Scenario 2 -Non-line-of-sight: With wall/opaque barrier in between target and transponder setup

The sensor was placed in the line of sight of the participants/subject at around a half-meter distance. A plasterboard/drywall wall was placed between the target and the radar. The experimental data recording activity for each head movement was carried out for 4 seconds and during these four seconds, the radar sent and received the RF signals. The subject was sitting on the chair in a normal position while performing head movements activity. The data collection and processing using UWB radar setup are shown in Figure. 5.3b.

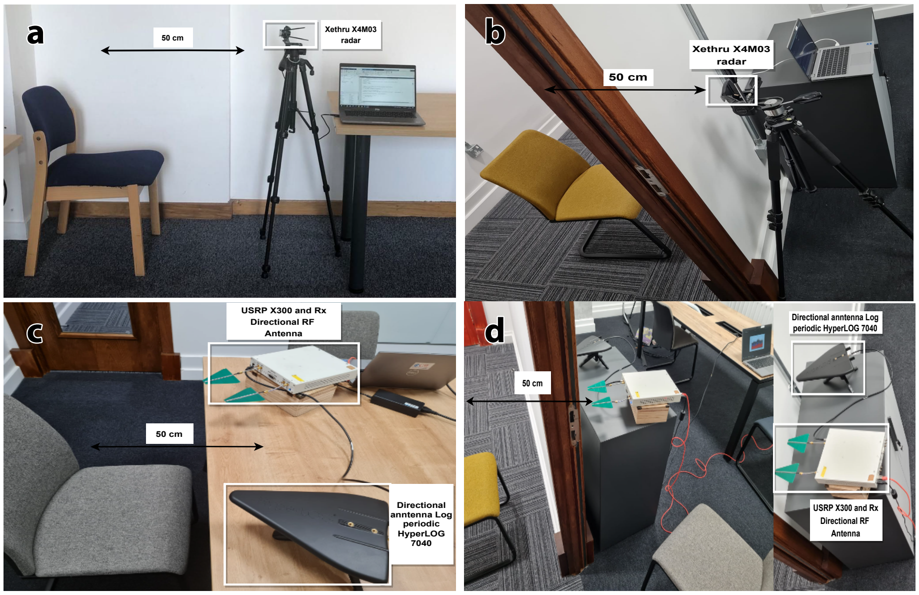

a
50 cm
Xethru X4M03 radar
b
50 cm
Xethru X4M03 radar
c
USRP X300 and Rx Directional RF Antenna
50 cm
Directional anntenna Log periodic HyperLOG 7040
d
50 cm
Directional anntenna Log periodic HyperLOG 7040
USRP X300 and Rx Directional RF Antenna

Figure 5.3: Head movements activity with their representation in Wi-Fi and radar signal. (a) An experimental setup of the radar signal without a wall. (b) An experimental setup involving radar signal penetration through a wall in a closed-door environment. (c) An experimental setup of Wi-Fi signal without a wall. (d) An experimental setup of Wi-Fi signal through a wall in a closed-door environment.

5.2.2 Wi-Fi based setup

The second set of experiments was performed using Wi-Fi. The experimental setup and parameter configuration for Wi-Fi based head movement system is given in Figure. 5.2b. The main equipment of this setup is USRP-X300 with a single transmitter antenna (directional) and two antennas at the receiver side which are (directional) in nature. On the transmitter side, the Rx antenna UWB 1.35GHz-9.5GHz Log-Periodic Directional was used as a transmitter whereas two monopole antennas (VERT2450) optimized at an operating frequency of 5.5 GHz were used as a receiver. The gain of both the Tx/Rx antennas was set to 35 dB. The USRP and desktop were connected using an Intel(R)-Core(TM) i7-7700 processor operating at 3.60 GHz with 16GB of RAM. Communication between the USRP and GNU Radio was established using a virtual machine running Ubuntu 16.04. A Python script was employed to transmit and receive data from the USRP-X300. The experiments were conducted within the 5.5 GHz Wi-Fi frequency band.

Scenario 1 - Line-of-sight: With no wall in between target and transponder setup

The Tx and Rx antennas were situated approximately 0.50 meters away from the subject, and each head movement was performed continuously for 4 seconds. The data collection process and subsequent processing using Wi-Fi equipment are depicted in Figure. 5.3c. It is important to note that Wi-Fi signals were evaluated based on a range of characteristics, including time-frequency maps, among others. Unlike radar signals, where frequency shift was the primary distinguishing factor, Wi-Fi CSI values were most effective when variations in CSI amplitudes were observed. These fluctuations in one-dimensional CSI amplitude revealed distinctive patterns of head movement.

Scenario 2 - Non-line-of-sight: With wall/opaque barrier in between target and transponder setup

The Tx and Rx antennas were positioned around 0.50 meters away from the subject. Plasterboard or drywall was placed between Tx/Rx, and target. The experimental data recording activity for each head movement was carried out for 4 seconds and during these four seconds, the Tx signal hit the target and was received back to the receiver. The subject was sitting on the chair in a normal position while performing head movements activity. The data collection and processing using Wi-Fi setup are shown in Figure. 5.3d.

5.3 Methods

The main illustration of head movement activity is shown in Figure. 5.4a. In the case of Wi-Fi, 2000 packets were transmitted within four seconds, where each data instance represented the CSI amplitudes. The CSI patterns (amplitude) of considered head movements, namely, Head down, Head up, Head left 90, Head Right 90, Head Right 45, and Head Left 45, are depicted in Figure. 5.4b without wall and Figure. 5.4d with wall experiments. In each figure, the 51 subcarriers of the OFDM signal are represented by different colors. The amplitude of the subcarriers is represented on the y-axis of each sub-figure, while the number of received packets is displayed on the x-axis. In the radar scenario, the same approach was used for data collection with a total of 600 data samples, four subjects participated including two males and two females, with 25 data samples in each class. Data is in the form of a spectrogram, which is shown in Figure. 5.4c without wall and Figure. 5.4e with wall experiments. Each figure's different colors represent a change in frequency. In each spectrogram, y-axis represents the doppler shift (Hz), while the x-axis represents time in seconds.

5.3.1 Radar Data Processing

The Xethru X4M03 radar chip was configured using the XEP interface and X4driver. Data was recorded at a rate of 500 Frames Per Second (FPS) in the form of float message data. A loop was implemented to read the data file, and the values were subsequently stored in a DataStream variable, which was then converted into a complex range-time-intensity matrix. To generate a doppler range map, an MTI filter was applied. Spectrograms were created using the following parameters: overlap percentage, window length set to 128, FFT, and padding factor set to 16. A Butterworth 4th-order filter, serving as the second MTI filter, was used.

Each chirp underwent an initial FFT transformation to produce a range profile. Subsequently, a second FFT was performed on a specific number of chirps in a sequence for each range bin. Spectrograms were created using the STFT, which segments the data and applies the fourier transform to each segment, providing information about both time and frequency. The radar data's doppler information depends on the hardware sampling rate, and the highest unambiguous doppler frequency in radar is determined by the chirp time, given by the formula $F_d, max = \frac{1}{2t_c}$.

Head movements recognition at a distance D(t) from a specified location such as the head is the focus of this chapter. Ts is the transmitted signal, while $V(t)$ is the target position in front of the RADAR,

$$T_s(t) = E\cos(2\pi ft). \tag{5.1}$$

The signal received is provided by Rs(t),

$$R_s(t) = \acute{E}\cos(2\pi f(t - \frac{2D(t)}{c})), \tag{5.2}$$

where the speed of light is c and E is the reflection coefficient. The signal that is reflected off the target points at an angle $theta$ to the direction of the RADAR and is denoted by the symbol $Rs(t)$.

$$R_s(t) = \acute{E}\cos(2\pi f(1 + \frac{2v(t)}{c})(t - \frac{4\pi D(\theta)}{c})). \tag{5.3}$$

The corresponding doppler shift can be expressed as,

$$f_d = f\frac{2v(t)}{c}. \tag{5.4}$$

The signal that is received back is composed of a number of moving parts, including the head and other small motions of the body. Each component moves with its own acceleration and speed. The received signal can be written as if i shows the various moving parts of the head. We can write as

$$R_s(t) = \sum_{k}^{N} A_k \cos(2\pi f(1 + \frac{2vk(t)}{c})(t - \frac{4\pi D_k(0)}{c})). \tag{5.5}$$

The doppler shift is the result of a complex interaction of multiple doppler shifts due to various head movements. The feature of doppler signatures depends on the detection of head movements. After getting the spectrogram of different subjects, It was divided into two datasets: (i) data training and (ii) data testing. The spectrogram fed into the proposed pre-trained DL classification algorithm for the classification of the head movements dataset.

5.3.2 Wi-Fi Data Processing

The data was transmitted using OFDM symbols with 52 subcarriers that were tightly spaced. According to Eq. 5.6, data were collected in a matrix form having frequency responses of subcarriers N=51.

$$H = [H_1(f), H_2(f), \cdots, H_M(f)]^K, \tag{5.6}$$

Here, the Hl-frequency subcarrier is expressed as

$$H_l\,(f) = |H_l\,(f)\,|e^{l\angle H_l(f)}, \tag{5.7}$$

where, amplitude $|H_l(f)|$ and phase $\angle H_l(f)$ are responses of the lth subcarrier. All subcarrier responses correlated with system input and output as shown in Eq. 5.8,

$$H_k(f) = \frac{Y_l(f)}{X_l(f)}, \tag{5.8}$$

where input and output fourier transformations are denoted by $Xl(f)$ and $Yl(f)$, respectively. The received CSI data often contain environmental noise. Therefore, the collected data is processed by eliminating the mean received power for each subcarrier from every sample. To observe the maximum variation due to head movements, the subcarrier with the highest variance was identified for feature extraction. These 10 features were extracted from the dataset namely minimum, median, variance, eight peaks, standard deviation, high order moments, mode, skewness, kurtosis, and moments. Features extracted into a CSV file were used to train various ML algorithms, as described in another section. After that, to accurately classify the head movement classes, training, and testing were carried out using the test-train split evaluation method.

5.3.3 Evaluation Metrics of Classification Models

The performance of DL and ML models was evaluated through TPR, FPR, and accuracy using the head movements dataset. Equations 3.9 and 3.10 are used to determine TPR and FPR, respectively. Additionally, accuracy was calculated using the equation, which is one of the most commonly used metrics in the literature for classification 3.14.

5.3.4 Parameter Settings of the ML and DL Algorithms

The presented approach for classifying head movements was divided into two parts: (i) system training and (ii) system testing. For the Wi-Fi numeric dataset, ML algorithms such as Neural Network (NN) pattern recognition, Medium Tree, and Boosted Trees were applied. The

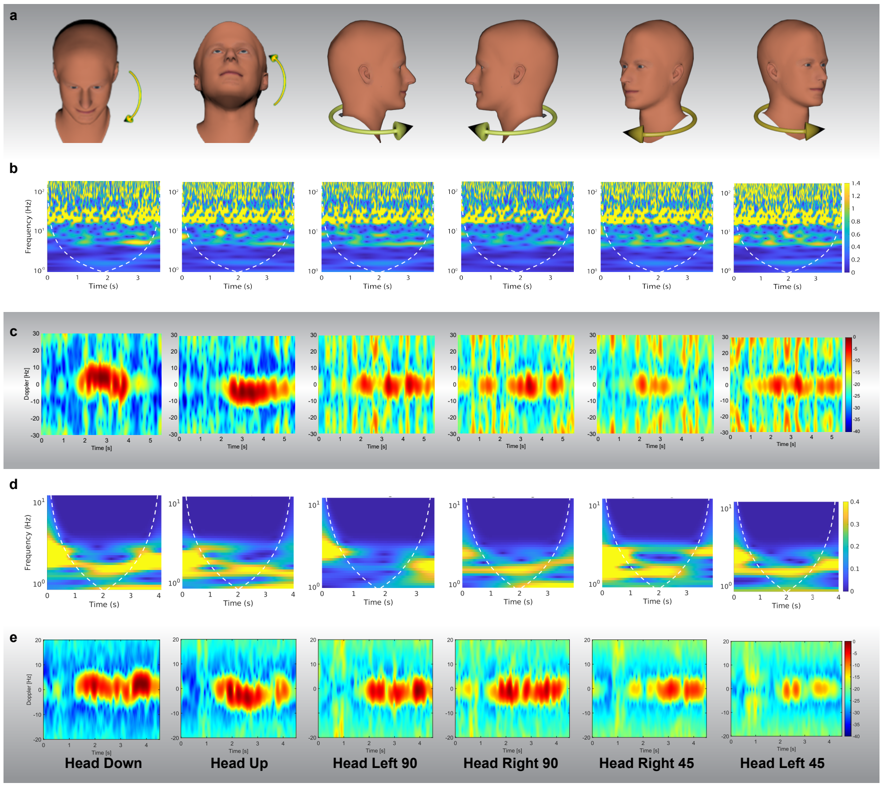
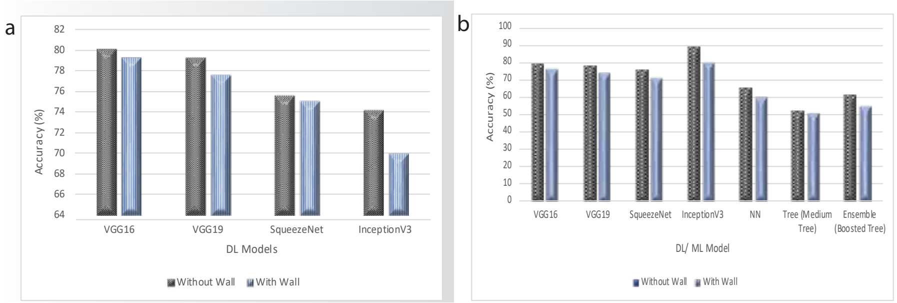


Figure 5.4: Wi-Fi and radar signal representation of head movement activity. (a) A visual representation of head movements from various angles. (b) Wi-Fi data samples representing various classes of head movements without walls. (c) Radar data samples representing various head movement classes without the wall. (d) Wi-Fi data samples representing various classes of head movements with walls. (e) Data samples from radar that represent different head movement classes with the presence of a wall.



c

DL Model	TPR/FPR (%)	Data Fusion						
		Head down	Head Up	Left 45	Left 90	Right 45	Right 90	Accuracy (%)
VGG16 + InceptionV3	TPR (%)	100	100	86.0	81.0	80.0	87.0	91.8
	FPR (%)	0.0	0.0	14.0	29.0	20.0	13.0	

d

DL Model	TPR/FPR (%)	Data Fusion						
		Head down	Head Up	Left 45	Left 90	Right 45	Right 90	Accuracy (%)
VGG16 + InceptionV3	TPR (%)	96.0	95.0	89.0	86.0	76.0	58.0	83.33
	FPR (%)	4.0	5.0	11.0	14.0	24.0	42.0	

Figure 5.5: Overall system overview and the results. (a) The comparative result of radar-based system with and without wall using DL models. (b) The comparative result of a Wi-Fi-based system with and without wall using DL and ML models. (c) The data fusion result of Wi-Fi and radar data without wall using DL models. (d) The data fusion result of Wi-Fi and radar data through the wall using DL models.

pre-trained DL models—VGG16, VGG19, InceptionV3, and SqueezeNet were utilised on spectrogram images generated from radar data. To achieve better results and improve future MM hearing impairment systems, we converted the Wi-Fi numeric dataset into scalograms. After getting features extracted from the spectrogram and scalgroam at fully connected layers fused it into a unified feature vector. This unified vector is then fed into the fully connected layer(s) of a neural network for further processing and ultimately for making predictions or classifications. The ML and DL model parameter settings are shown in Table. 5.1.

VGG16 Model: The data was input into the VGG16 models convolution layers with rectified linear unit (ReLU) activation functions and 3×3 kernel sizes. Each convolution layer was followed by a max-pooling layer with 2×2 kernel sizes. The final layer comprised three FC. The convolution layer and FC layers contained the training weights, which determined the number of parameters.

VGG19 Model: The data was passed through a different layer which consists of 3×3 filters with five stages of convolutional layers, five pooling layers, and three fully connected layers to get image information. The convolution kernel depth has been increased from 64 to 512 of the VGG16 network for better image feature vector extraction. Every stage of convolutional layers was followed by pooling layers which have the size and step size of 2×2.

InceptionV3 Model: The dataset was processed using the InceptionV3 DL model, which consists of 48 layers. The architecture of the model involves a sequence of three convolution layers, followed by a max pooling layer, two more convolution layers, and another max pooling layer.

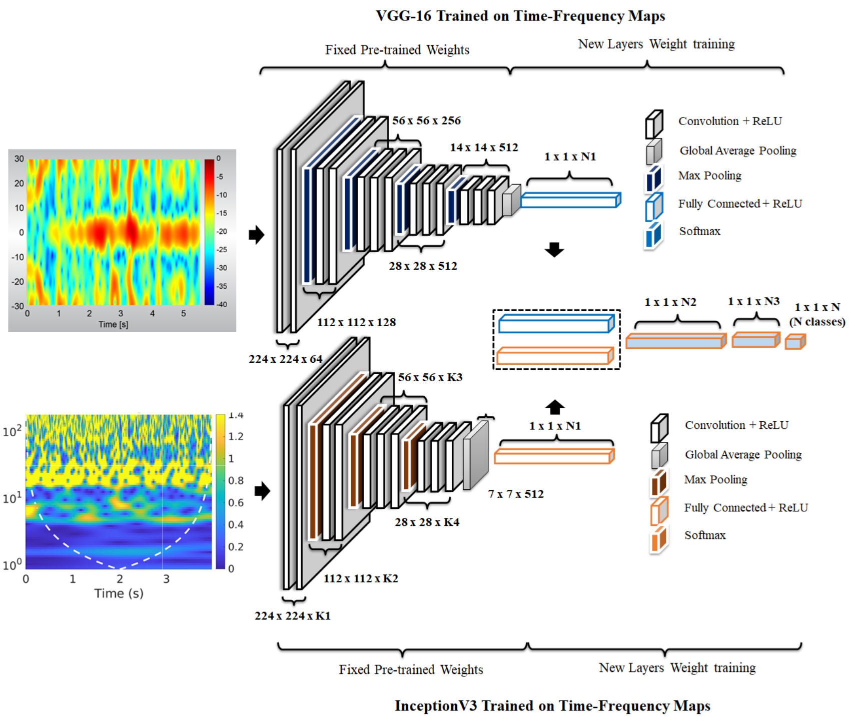


Figure 5.6: Feature fusion of radar and Wi-Fi time-frequency maps.

Spectrograms were input into the model, which then underwent multiple convolutions using various filters. This process was repeated several times across the entire network to facilitate image classification.

SqueezeNet Model: SqueezeNet is an 18-layer deep CNN. Spectrograms of the input were sent to the layers. The last convolution layers were added as follows the dropout layer was set to 50%, convolution layers with stride, Relu as activation function, Global average pooling, and softmax layer were added before the classification output layer.

NN (Neural Network Pattern Recognition) Model: The pattern recognition neural network used in this chapter comprises two-layer feed-forward networks with hidden neurons using sigmoid activation functions. SoftMax activation functions were applied to the output layer neurons. The network was trained using the scaled conjugate gradient backpropagation algorithm, which involved updating the weight and bias values as data passed through these layers. Subsequently, the dataset was partitioned into training, validation, and testing subsets. The network's performance was assessed based on cross-entropy and misclassification error metrics.

Tree (Medium Tree) Model: Data were fed to decision trees, classification trees, and regression trees for classification. It followed the decisions in the trees down to a leaf node in order to forecast a reaction. The response was located in the leaf node. Classification trees provided nominal answers, such as "true" or "false".

Ensemble (Boosted Tree) Model: The classifier has the ability to combine the results of multiple low-quality learners into a single high-quality model. The data were input to the booting ensemble algorithm, which identified the highest breakpoints or branch points to handle the depth of tree learners. The experimental setup achieved improved precision with a learning rate of 0.1.

5.4 Results and Discussion

Two RF sensing technologies were used in two different experiments with and without a wall, *i.e.,* Wi-Fi and radar. Data collection involved the capture of six head movements: Head up, Head down, Head Right 90, Head Left 90, Head Right 45, and Head Left 45. These movements were recorded with subjects in a stationary position and their bodies in a typical posture. To enhance the dataset's authenticity, four participants (two males and two females) took part in both the radar and Wi-Fi experiments. A total of 2400 data samples were collected from both experiments using radar and Wi-Fi, with and without a wall as shown in Figure. 5.2c. In each experiment with wall and without wall using radar, a total of 600 data samples were collected from four participants, where 25 samples were taken from each class. Specifically, each participant repeated each head movement activity 25 times with the radar. Likewise, the same number of data was acquired from USRP using the same strategy. The University of Glasgow's Research Ethics Committee granted ethical approval for these experiments (approval no.: 300200232, 300190109). In the context of radar datasets, both with and without a wall, the evaluation outcomes for the considered DL algorithms (VGG16, VGG19, SqueezeNet, and InceptionV3) are presented in Figure. 5.5a. Notably, all the algorithms produced comparable results, with VGG16 slightly outperforming the others in both scenarios, whether with or without a wall, when using a combined dataset. Specifically, when employing the VGG16 algorithm, a classification accuracy of 80.0% is achieved on the combined dataset without a wall, which is marginally reduced to a promising accuracy of 79.2% in the presence of a wall.

The evaluation results for various DL and ML algorithms (including VGG16, VGG19, SqueezeNet, InceptionV3, Neural Network Pattern (NNP) recognition, Tree (Medium Tree), and Ensemble (Boosted Tree)) for Wi-Fi signals both with and without a wall are displayed in Figure. 5.5b, using the combined dataset. It is evident from the graph that the InceptionV3 algorithm surpasses the others on the combined dataset. Specifically, when utilising the InceptionV3 algorithm, a classification accuracy of 89.0% is achieved without a wall, whereas the same algorithm yields an 80.0% classification accuracy with the presence of a wall. The fusion of different DL models was tested which is illustrated in Figure. 5.6 . The highest accuracy values of 91.8% without the wall were achieved with feature fusion at the fully connected layers of VGG16 and InceptionV3 DL models shown in Figure. 5.5c. Furthermore, the highest accuracy of 83.33% was achieved through the walls with the feature fusion of VGG16 and InceptionV3 DL models shown in Figure. 5.5d.

DL/ML Model	Parameters	Settings
VGG16	Number of Layers	16
	Learning rate	0.0001
	Batch size	16
	Learning algorithm	Adam
	Loss function	Cross entropy
	Number of epochs	25
VGG19	Number of Layers	19
	Learning rate	0.0001
	Batch size	16
	Learning algorithm	Adam
	Loss function	Cross entropy
	Number of epochs	25
InceptionV3	Number of Layers	48
	Learning rate	0.0001
	Batch size	16
	Learning algorithm	Adam
	Loss function	Cross entropy
	Number epochs	25
SqueezeNet	Number of Layers	18
	Learning rate	0.0001
	Batch size	16
	Learning algorithm	Adam
	Loss function	Cross entropy
	Number of epochs	25
NN	Number of Layers	10
	Training Function	Scaled conjugate Gradient Backpropagation
	Number of epochs	20
	Loss function	Cross entropy
Tree (Medium Tree)	SplitCriterion	gdi
	MaxNumSplits	20
	Surrogate	off
	KFold	5
	Loss Function	Classiferror
Ensemble	Learner type	Decision Tree
	Ensemble Method	AdaBoost
	Loss Function	Classiferror
	Learning rate	0.1
	Number of learners	30
	Maximum Number of splits	20

Table 5.1: Parameter settings for the selected DL and ML models.

5.5 Summary

The work of this chapter has presented an RF sensing-based head movement recognition system proposed using Wi-Fi and radar, and state-of-the-art DL and ML algorithms. All directions of head movements were covered, such as Head up, Head down, Head left 90, Head right 90, Head left 45, and Head right 45. Wi-Fi data was passed to the InceptionV3 model and radar data to VGG16 models and the features of the two models were fused for the highest performance results of 91.8% without the walls and 83.33% accuracy was achieved through the walls. Furthermore, the proposed system preserves the privacy concerns of users, which may exist in vision-based systems. The next chapter focuses on another aspect of future MM hearing impairments, centering on RF-based facial expression analysis. This technique helps deaf individuals to recognise the expressions of another individual via visual cues.

Chapter 6

RF Sensing Enabled Tracking Of Human Facial Expressions Using Machine Learning Algorithms

This chapter presents facial expressions as crucial indicators for understanding human behavior, enabling the identification and assessment of positive and negative emotions. Moreover, facial expressions provide insights into various aspects of mental activities, social connections, and physiological information. Currently, most facial expression detection systems rely on cameras and wearable devices. However, these methods have drawbacks, including privacy concerns, issues with poor lighting and line of sight blockage, difficulties in training with longer video sequences, computational complexities, and disruptions to daily routines. To address these challenges, this chapter proposes a novel and privacy-preserving human behavior recognition system that utilises FMCW radar combined with ML techniques for classifying facial expressions. Specifically, the study focuses on five common facial expressions: Happy, Sad, Fear, Surprise, and Neutral. The recorded data is obtained in the form of a micro-doppler signal, and state-of-the-art ML models such as Super Learner, Linear Discriminant Analysis, Random Forest, K Nearest Neighbour, Long Short-Term Memory, and Logistic Regression are employed to extract relevant features. These extracted features from the radar data are then fed into ML models for classification. The results show a highly promising classification accuracy of 91%.

6.1 Introduction

Facial recognition technology has witnessed substantial advancements in recent years, finding widespread applications across various domains. These applications include healthcare and mental health, human-robot interaction, biometric identification, human-computer interaction, security and surveillance, as well as entertainment and gaming. It is also treated as a nonverbal sign used by people to convey emotions, intentions, and social signals [181]. The paper

[182] presents that understanding and monitoring facial expressions has piqued the interest of researchers in a variety of fields, including psychology, neuroscience, and computer vision. Face expressions are ubiquitous and may be used to communicate effectively across cultures and languages [183]. They are essential in social interactions because they let humans express emotions such as happiness, sadness, anger, fear, surprise, disgust, and contempt [184]. Facial expressions are caused by the coordinated movement of facial muscles, which results in recognised and interpretable patterns by human observers [185]. Affective computing, human-computer interaction, healthcare, and social robotics all benefit from the capacity to reliably monitor and comprehend facial emotions [186]. The current facial movement detection systems are based on wearable and camera-based technologies. These techniques have limitations, such as the need to record the target, which restricts their practical use due to privacy concerns. The current facial detection systems are based on wearable and camera-based technologies. These techniques have limitations, such as the need to record the target, which restricts their practical use due to privacy concerns. The legal implications of such aids may restrict their wider use in public and private settings; for example, video-in-head motions may be viewed as photographing someone without their consent, which is illegal in many countries. The main drawbacks of existing camera-based and wearable-based technology include serious privacy concerns, poor lighting, obstructions to the line of sight, training difficulties with longer video sequence data, and computational complexities, and wearable devices disrupt daily routines.

In contrast, RF facial expression sensors present a promising solution to meet the requirements of next-generation technologies. These sensors utilise RF sensing and ML techniques to accurately detect and recognise facial expressions, offering valuable cues that can benefit a wide range of applications.

The following presents the main contributions of our research work in the field:

- We proposed a unique RF sensing-based facial expression monitoring system that integrates powerful ML algorithms for accurate facial expression recognition and is applicable for different applications such as healthcare and mental health, human-robot interaction, biometric identification, human-computer interaction, security and surveillance, as well as entertainment and gaming.

- A dataset comprising 1000 samples was gathered, encompassing five distinct facial expressions. The data collection took place at a consistent distance of 1.50 meters from the target. To ensure variability and inclusiveness, four participants were involved in the data collection process, consisting of two males and two females within the age range of 20 to 40 years.

- We used extensive experiments and comparisons with existing camera-based and wearable device-based technologies to assess the performance and usefulness of the proposed RF-based system, demonstrating the benefits and prospective uses of RF sensing in facial expression monitoring.

- In this study, we have presented the experimental results of several advanced ML models applied to our benchmark dataset. These findings provide valuable insights and can serve as a fundamental reference for future research in the domain of facial expression detection.

This research introduces innovative facial expression gestures using radar-sensor and micro-doppler signatures. The study focuses on five distinct gestures: Neutral, Happy, Sad, Fear, and Surprise. Experimental data is collected using an FMCW radar, and the recorded data is represented as a micro-doppler signal. The classification accuracy achieved is 91.0%. The potential applications of this technology are illustrated in Figure. 6.1. Detailed information regarding the setup, data collection process, ML algorithms, and experimental results are provided in the subsequent sections.

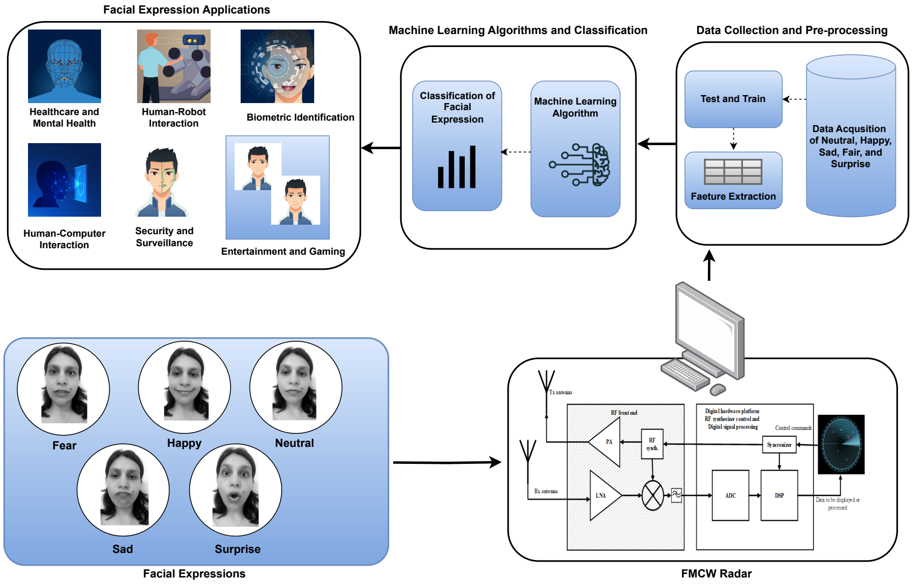

Facial Expression Applications
Healthcare and Mental Health
Human-Robot Interaction
Biometric Identification
Human-Computer Interaction
Security and Surveillance
Entertainment and Gaming
Machine Learning Algorithms and Classification
Classification of Facial Expression
Machine Learning Algorithm
Data Collection and Pre-processing
Test and Train
Faeture Extraction
Data Acqusition of Neutral, Happy, Sad, Fair, and Surprise
Fear
Happy
Neutral
Sad
Surprise
Facial Expressions
Tx antenna
Rx antenna
RF front end
PA
RF synth.
LNA
Digital hardware platform RF synthesizer control and Digital signal processing
Control commands
Syncronizer
ADC
DSP
Data to be displayed or processed
FMCW Radar

Figure 6.1: The overall flow diagram of proposed facial expressions system.

6.2 Methodology

The methodology used in this study is depicted in Figure. 6.1 as a block diagram. The framework comprises three main steps. Firstly, diverse facial expression datasets were collected, constructed, and annotated using FMCW radar. Subsequently, the pre-processing phases are explained in the form of an algorithm, as depicted in Figure. 6.3. Finally, a range of ML models was employed for the classification of facial expressions. The following subsections provide a detailed discussion of each stage in the proposed methodology.

Parameter	Value
Sensor Type	FMCW Radar
Center Frequency	9.5 GHz
Bandwidth	1GHz
Chirp Width	50 us
Chirp Repitation Period	100 us
Transmitter Power	23 dBm
Reciever Gain	20 dB
Tx and Rx Type	Horn Antenna
Antenna Gain	13 dB
Number Of Samples Per Class	30
Distance From Rdar to Target	1.5 meter
Activity Duration	4 sec

Table 6.1: Selected hardware and software parameter settings

6.2.1 Experimental Setup and Data Collection

The hardware configuration of the radar-based facial expression system is illustrated in Figure. 6.2. The experiments were conducted in two separate rooms, as shown in Figure. 6.2a and Figure. 6.2b, representing Room 1 and Room 2, respectively. The experimental setup included a FMCW radar sensor, positioned in front of the user. The FMCW radar sensor comprised a Tx, an Rx, and two horn antennas for transmitting and receiving, allowing for a maximum detection range of 20 meters. The key parameter settings of the radar system are outlined in Table. 6.1. During the facial expression tasks, the subject was positioned at a distance of 1.50 meters from the radar. The subject's body remained in a natural position, with the only movements being those of the face and slight head movements, which are typical during the conversation. A single subject facial expression was recorded for each action in the study for a total of 4 seconds. The radar system transmitted and received the RF signal during this time period. Figure. 6.4b presents the range time output captured during the measurement of various facial expressions. This output is utilised for signal processing and distance measurement purposes. To extract features from the radar data, the FFT is applied, generating spectrograms that illustrate the radar doppler shift caused by facial movements as depicted in Figure. 6.4c. The analysis of these spectrograms reveals variations corresponding to different facial expressions due to the distinct movements of the face and mouth. The use of a trigger in a radar system is essential to ensure precise timing, synchronisation, and control over the data acquisition process, as shown in Figure. 6.5.

6.2.2 Data Pre-Processing and Machine Learning Models

The FMCW radar was used for data collection, many parameters are important for target detection in radar systems such as Tx power, Carrier frequency, antenna gain, receiver sensitivity,

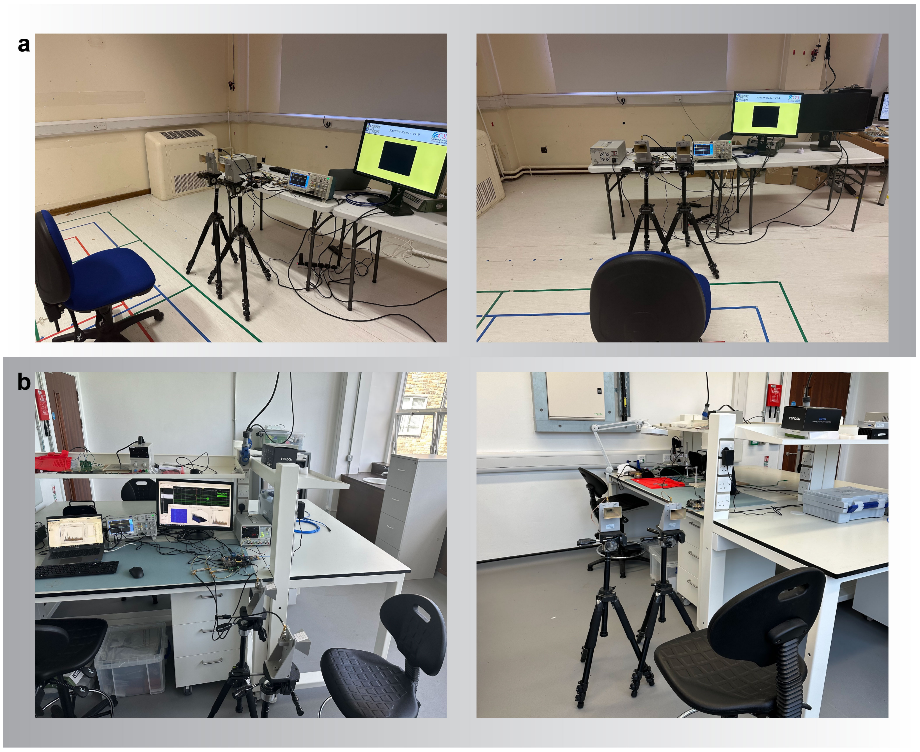


Figure 6.2: The experimental setup. (a) The experimental setup for room 1. (b) The experimental setup for room 2.

target cross-section area, and receiver noise. To extract facial expression accurately, first detect the target location is needed. The target range in radar systems is defined by the delay between the transmitted and received echoes. Equation 6.1 shows the radar range equation:

$$R_{max} = 4\sqrt{\frac{P_t G_t G_r^2 \sigma}{(4\pi)^3 K T_0 B F_n \frac{S}{N}}} \tag{6.1}$$

Where P_t is the transmitted power, G_t is the Tx antenna gain, G_r is the receiver antenna gain, is the wavelength, is the target cross-section area, K is Boltzmann constant, T is system temperature, B is the Bandwidth, F is the receiver noise Figure, and S/N is signal to noise ratio[187],[188]. Firstly, we generated of the FMCW ramp which is the fundamental step in the operation of FMCW radar systems, enabling the measurement of target range based on the frequency difference between transmitted and received echoes. After that, acquired the received echoes using an Analog-to-Digital Converter (ADC) with a sampling rate of 250 MSPS (Mega Samples Per Second) and applied signal processing techniques on collected samples such as low pass filtering, target extraction, and range calculation. Signal integration technique is applied for improving SNR to enhance the signal quality and increase the probability of detection. SNR was measured of the strength of the desired signal compared to the background noise level. Downsampling technique was used to reduce the sampling rate of a signal while retaining the essential information. We did downsampling to 10 MSPS. For data pre-processing and ML approaches,

Algorithm 1 Algorithm for Range-Time Map Generation and Classification

```
procedure RangeTimeMapGeneration($F_b$)
    Input: Beat frequency $F_b$
    Output: Range-time map $X$ (number of chirps × samples per chirp)
end procedure
procedure ModelRepresentation
    Denoising algorithm
    FFT (Fast Fourier Transform) for range estimation (Vectorized)
    Target detection
    Doppler processing for detected targets (Parallelized)
    Send results through Ethernet for data processing
end procedure
procedure Classification($X$)
    Input: Input space $X$
    Output: Output space (Classes) $Y$
    Training data: Parallelized data loading and processing
    Hypothesis space
    Learning algorithm: $A$
end procedure
procedure LearningProcess($h(x)$)
    Parallelized hypothesis selection by $A$
    Loss function: $L(y, h(x))$ (Measures classification error)
end procedure
procedure TrainingPhase
    Parallelized data batching and processing
    Minimize empirical risk: $R_{\text{emp}}(h) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, h(x_i))$
end procedure
procedure EvaluationAndTuning($x_{\text{test}}, y_{\text{test}}, \Theta$)
    Initialize best hyperparameters $\theta_{\text{best}}$
    Initialize best evaluation metric (e.g., accuracy) $e_{\text{best}}$
    for each $\theta \in \Theta$ do
        Perform $k$-fold cross-validation with hyperparameters $\theta$
        Compute the average cross-validation performance metric
        if cross-validation metric is better than $e_{\text{best}}$ then
            Update $e_{\text{best}}$ with the new metric value
            Update $\theta_{\text{best}}$ with the current hyperparameters $\theta$
        end if
        Perform other technique (e.g., grid search, random search) with hyperparameters $\theta$
        Compute the performance metric for the other technique
        if other technique metric is better than $e_{\text{best}}$ then
            Update $e_{\text{best}}$ with the new metric value
            Update $\theta_{\text{best}}$ with the current hyperparameters $\theta$
        end if
    end for
    Select the best hyperparameters $\theta_{\text{best}}$ based on both techniques
    Train the final model using the entire dataset and $\theta_{\text{best}}$
    Save the trained model to a file
end procedure
```

Figure 6.3: Pseudo code for the proposed facial expression systems.

we employed the Scikit-learn library, widely recognised for its comprehensive suite of data analysis tools and supervised ML algorithms in Python, which are already optimised [189, 190, 191]. Additionally, we employed Pandas, a Python library, for parsing CSV files and converting them into data frames. Labels were assigned to the first column of the data frames. In the combined dataset, produced by merging the data frames of each sample, NaN (Not a Number) values may arise due to slight mismatches in the micro-doppler signal. To handle these NaN values, we used the SimpleImputer function from Scikit-learn to replace them with the mean of each row. It is important to note that this data cleansing process does not alter the overall data patterns. Following the data cleansing step, the processed data, which was in numerical form, was input into several ML algorithms. This study's proposed facial expression recognition system underwent evaluation using six distinct ML methods. The system's performance was assessed based on the test accuracy with which it could correctly classify various facial expressions. Each ML algorithm's accuracy is assessed independently using two approaches to ensure robust analysis: (i) k-fold cross-validation and (ii) train-test split. K-fold cross-validation is a widely adopted approach in ML testing, where the dataset is divided into k groups. In this experiment, we set k to 10, resulting in the dataset being divided into 10 groups or folds. Each group is then used as a test set while the remaining groups serve as training sets. This process is repeated k times, with each group acting as the test set once. The results obtained from each group of classifications collectively represent the performance of the ML algorithms on the entire dataset. In addition to k-fold cross-validation, the train-test split technique is employed. This technique involves dividing the dataset into training and testing subsets. The training data is used to train the ML models, enabling them to learn from the provided inputs and corresponding labels.

6.3 Experiments and Results

In this section, we will provide a detailed description of the dataset used in the study, along with an evaluation of the system using the ML models mentioned earlier.

6.3.1 Dataset

In this phase, facial expression data was collected using FMCW radar. Figure. 6.2 illustrates the hardware setup of the radar-based system employed for facial expression data collection. The FMCW radar sensor was equipped with two horn antennas, one for Tx and the other for Rx, enabling a maximum detection range of 20 meters. As shown in the figure, the radar sensor is placed on the table. The key parameter settings of the radar are indicated in the Table. 6.1. In order to encompass different complexity levels in the dataset, it was recorded in two different room environments. The data collection is significantly impacted by a variety of environmental circumstances, ensuring the reliability and authenticity of the proposed system in a variety of locations. After collecting the dataset, it is proved that the system has the same behaviour in all environments. During the data collection process, the subjects maintained a neutral body position, focusing solely on facial movements. Additionally, each activity had a fixed duration of 4 seconds, allowing for the collection of data corresponding to a single gesture performed by an individual subject. Figure. 6.4 provides a visual illustration of the face expression datasets. The ethical approval for these experiments was obtained from the Research Ethics Committee at the University of Glasgow (approval no. 300200232, 300190109). The data collection process involved the participation of four individuals, consisting of two males and two females. The inclusion of multiple participants aimed to enhance the realism and diversity of the dataset. A total of 1000 data samples were collected during the experiment, encompassing five distinct categories across two different rooms. The details of the collected dataset are highlighted in the Table. 6.2. In each experiment, a total of 1000 data samples were collected from four participants, with 30 samples collected in each class. In particular, each participant repeated the facial expression activity of each gesture 30 times with the radar. In this way, each participant contributed to collecting 250 data samples in total for the fifth class. In each case, a total of 1000 data were categorised as fifth facial gestures, of which 800 were utilised for training and 200 for testing.

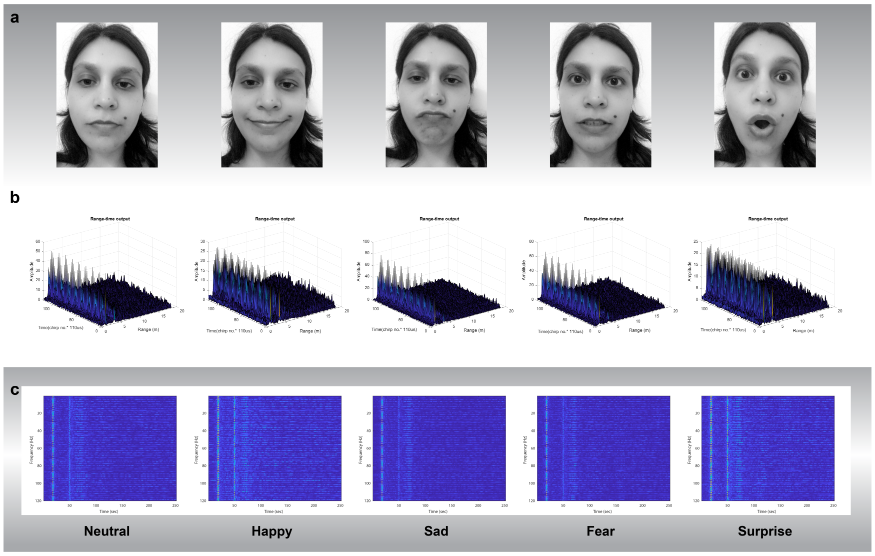


Figure 6.4: The radar signal representation of facial expression activities. (a) A visual representation of facial expression activities. (b) The range-time output of various activities (c) Spectogram representing various facial expression activities.

6.3.2 Performance Metrics for Evaluating the Classification Model

In this study, the performance evaluation of ML models for facial expression dataset classification involves several metrics, including weighted average accuracy, precision, recall, f1-score, accuracy, and a 95% confidence interval (CI). The F1-score is a widely used metric in classifi-

Classes	**Experimental Dataset**				
	Subject (S1)	**Subject (S2)**	**Subject (S3)**	**Subject (S4)**	**Total**
Happy	50	50	50	50	200
Sad	50	50	50	50	200
Surprise	50	50	50	50	200
Fair	50	50	50	50	200
Neutral	50	50	50	50	200
Total	250	250	250	250	1000

Table 6.2: A summary of the collected data, the participant count, and the conducted activities of facial expressions.

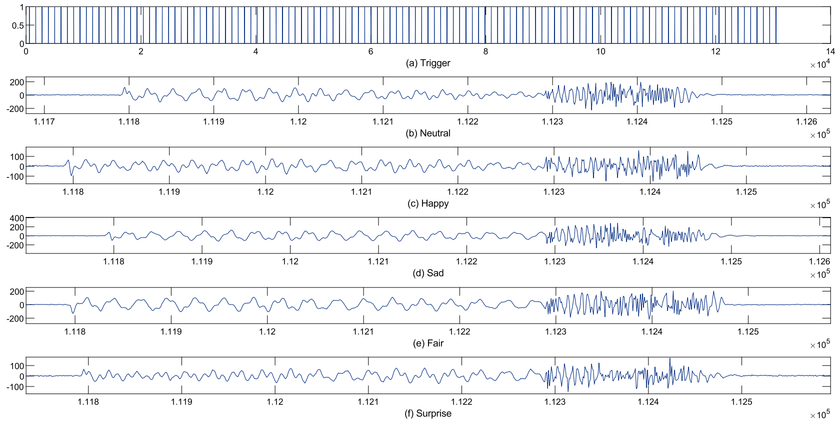


Figure 6.5: A visual representation of variation in frequency for different facial expressions (a) Trigger. (b) Neutral. (c) Happy. (d) Sad. (e) Fair. (f) Surprise.

cation literature, serving as a measure of classification performance. It combines precision and recall, which are calculated using Equations 6.3 and 6.4, respectively. The F1-score, calculated using Equation 6.2, provides a comprehensive evaluation of the model's ability to balance both precision and recall in classification tasks. The overall accuracy of the combined dataset is calculated using 6.5 and verified the interval using 6.6. Where, the 95% asymptotic CI measures the statistical significance of experimental results. It represents the radius, with n = 1000 samples (20% of the dataset), and uses k as the number of standard deviations. The CI has a 95% probability of containing the true classification result. A value of k = 1.96 from the Gaussian distribution establishes this 95% confidence level.

$$F1 - Score = 2\frac{Precision.Recall}{Precision + Recall} \tag{6.2}$$

$$Precision = \frac{\sum TP}{\sum TP + \sum FP} \tag{6.3}$$

$$Recall = \frac{\sum TP}{\sum TP + \sum FN} \tag{6.4}$$

Models	Precision	Recall	F1-Score	Accurcay (%)	95% CI
Super Learner	0.90	0.90	0.90	90.0	0.84-0.96
Linear Discriminant Analysis	0.81	0.79	0.79	79.0	0.77-0.82
Random Forest	0.86	0.85	0.85	85.0	0.83-0.87
K Nearest Neighbour	0.82	0.80	0.80	80.0	0.78-0.83
Long Short-Term Memory	0.92	0.91	0.91	91.0	0.89-0.93
Logistic Regression	0.75	0.75	0.74	75.0	0.72-0.78

Table 6.3: The evaluation of the ML models on the facial expression dataset involved measuring weighted average recall, weighted average precision, weighted average F1-score, accuracy, and determining a 95% confidence interval.

$$Accuracy = \frac{\sum(TP+TN)}{\sum(TP+FP+TN+FN)} \tag{6.5}$$

$$Interval = k \times \sqrt{\frac{Accuracy \times (1-Accuracy)}{n}} \tag{6.6}$$

6.3.3 Results and Discussion

This experimentation serves two purposes. In the first step, we introduced radar-based facial recognition, and in the next step, we compared the performance of various existing ML models such as Super Learner, LDA, RF, KNN, LSTM, and Logistic Regression. We collected and analysed the performance of facial expression frameworks using different facial expression datasets such as Neutral, Happy, Sad, Fair, and Surprise from different genders. As a result, we performed experiments on micro-doppler signal data to evaluate the model's performance. The hyper-parameter settings for all models are listed in the Table. 6.4. All of the models on the dataset have been fine-tuned. Additionally, the training and testing sets were fixed throughout all studies. The percentages of the entire data in our training and testing sets are 80% and 20%, respectively Table. 6.3 shows the outcomes of studies with various facial expression structures in terms of precision, recall, f1-score, accuracy, and interval which help with decision-making and comparisons. The Figure. 6.6 shows the confusion matrix of all proposed models on collected datasets. Overall, better outcomes were obtained for the combined and individual datasets using all models.

In the super learner algorithm, the combined dataset includes males and females. We got a high classification accuracy of 90% with precision, recall, and F1-score and accurate interval, which are shown in Figure. 6.6a. All the classes are correctly classified except Fair because 11% of expressions are similar to Happy.

Similarly, Linear Discriminant Analysis is well-performed on the combined dataset with 80% accuracy, precision, recall, f1-score and valid interval which are shown in Figure. 6.6b. All the classes are correctly classified except Fair which has been misclassified with Happy, Sad, and Surprise with a ratio of 0.08, 0.08, and 0.16.

Using Random Forest, the combined dataset includes males and females. We got a high classification accuracy of 85% with precision, recall, f1-score, and accurate interval which are shown in Figure. 6.6c. All the classes are correctly classified except Fair and Surprise. The Fair has similarities with Happy with a ratio of 0.16. Here again, Surprise has similarities with Fair, Happy, Neutral, and Sad with ratios of 0.095, 0.048, 0.048, and 0.048.

In the case of the K Nearest Neighbour algorithm, we got a high classification accuracy of 80%, precision, recall, f1-score, and interval, which are shown in Figure. 6.6d. Except for Surprise all classes are correctly classified because it has similarity with Fair and Sad with a ratio of 0.14 and 0.095.

Using the Long Short-Term Memory algorithm on the combined dataset, we got high classification accuracy of 91%, precision, recall, f1-score, and interval, which are shown in Figure. 6.6e. All classes are correctly classified except Fair which has been misclassified with Sad and Surprise with ratios of 0.15 and 0.15.

Logistic Regression, the combined dataset includes males and females. We got a high classification accuracy of 75% with precision, recall, and F1-score and an accurate interval, which are shown in Figure. 6.6f. All the classes are correctly classified except Fair because it has similarities with Happy, Neutral, Sad, and Surprise with ratios of 0.12, 0.08, 0.08, and 0.16.

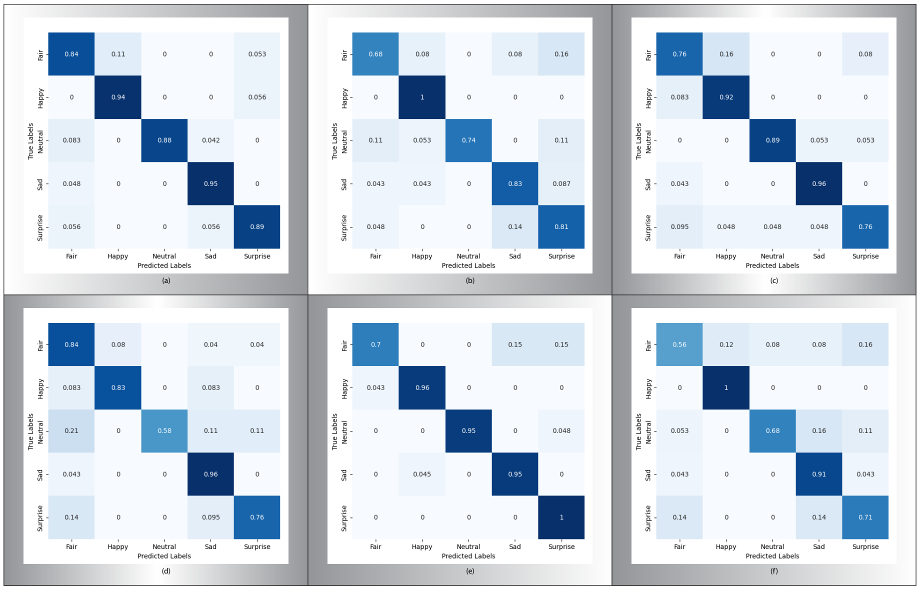


Figure 6.6: The confusion matrix of well-known ML algorithms on the combined dataset. (a) Super Learner. (b) Linear Discriminant Analysis. (c) Random Forest. (d) K Nearest Neighbour. (e) Long Short-Term Memory. (f) Logistic Regression.

6.4 Summary

This chapter presents a contactless and privacy-preserving facial recognition framework. The diverse dataset is taken from different users in the form of micro-doppler signals and fed into

ML Model	Parameters	Settings
Super Learner	N_estimator Type N_split Solver Gamma	20 Multi-threading 10 Liblinear Scale
Linear Discriminant Analysis	Solver Shrinkag Store_covariance Tol CV Covariance_estimator	Svd None False 0.0001 10 None
Random Forest	N_estimator CV Criterion Mini_samples_Split Max_features	200 10 gini 2 sqrt
K Nearest Neighbour	N_neighbors CV Weights Leaf Size Metric	5 20 Uniform 30 minikowski
Long Short-Term Memory	Learning rate Batch size Learning algorithm Loss function Number of epochs	0.0001 128 Adam Binary_crossentropy 100
Logistic Regression	Penalty tol Solver CV C	l2 0.0001 lbfgs 10 1.0

Table 6.4: Parameter settings for the ML models

well-known ML models. The collected data consists of five different classes: neutral, happy, sad, fair, and surprise. The experiment included four participants, two male and two female, aged 20 to 40 years. The micro-doppler data is fed into various ML models, including Super Learner, Linear Discriminant Analysis, Random Forest, K Nearest Neighbour, Long Short-Term Memory, and Logistic Regression. The face movements were mostly classified correctly, achieving a 100% accuracy rate. Among the models tested, the Long Short-Term Memory algorithm performed the best, with an overall accuracy of 91% for all five classes. The next chapter discusses the future direction of MM hearing impairment devices.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

The main purpose of this thesis is to propose future MM hearing aid devices based on RF sensing capable of communicating deaf communities with normal individuals with the assistance of ML and DL. The thesis details the current literature in the field of hearing impairments by using contact and contactless methods. Contact-based methods include wearable sensors. Although contact-based sensors provide good accuracy, they have disadvantages such as discomfort, battery capacity issues, and human forgetfulness to wear the device. The regular use of assistive technologies becomes especially critical in the context of cognitive disorders. This is because deaf people have a higher risk of memory problems, making continuous engagement with these devices become necessary for their efficiency. When people discomfort to utilise these devices daily, their usefulness decreases significantly. Contactless methods include cameras, radar, Wi-Fi, and RFID technology. Camera technology is effective for monitoring lip-reading recognition, sign language, head movements, and facial expressions, however, it has privacy issues. Using cameras in hearing aids could be seen as recording without consent, which is legally problematic in many regions. Moreover, the widespread use of face masks during the COVID-19 pandemic has further limited the effectiveness of vision-based hearing aids.

RF sensing-based hearing aid technologies, including Wi-Fi, Radar, and RFID, effectively address the limitations of wearable and camera-based devices. These RF sensing devices offer numerous advantages for future hearing aids: they maintain privacy, operate effectively even with face masks on since RF signals can penetrate masks to detect vital visual cues, and are commonly present in homes with Wi-Fi networks. Wi-Fi transmits data between devices wirelessly using RF signals. Micro-movements of lips, head, hands, and face between an RF transmitter and receiver affect signal propagation, visible through CSI analysis. The radar framework utilises doppler shift spectrograms and requires only a single Tx and Rx on a single chip. A DL model interprets these spectrograms to classify various micro-movements for hearing aids. Additionally, the another RF-based framework incorporates a passive RFID tag, similar to those used in UHF Textile Laundry products, integrated into standard masks for data collection. The collected data represented as RSSI values, are analysed using different ML models to detect these micro-movements. In order to assist deaf communities, this thesis offers a monitoring system that uses RF signals to detect micro-movements. It investigates the identification of RF patterns linked to certain spoken and nonverbal signals through the use of ML and DL techniques. The thesis analyses various ML and DL algorithms focusing on their accuracy to ensure that the future hearing device meets the needs of deaf people and facilitates all perspectives.

The thesis chapters elaborate on how a multi-functional, future-oriented hearing aid was proposed using RF sensing technology. The initial applications of the system successfully detected human lip movements (vowels, consonants, words, and sentences), recognising the primary vocabulary of BSL, identifying human behaviors through head movement recognition(Head up, Head down, Head left 90, Head right 90, Head left 45, and Head right 45), and facial recognition(Neutral, Happy, Sad, Fear, and Surprise) which are accurately visible in motions of the CSI data stream, Doppler shift, and RSSI information. This progression has enabled the application of ML and DL techniques to precisely classify the specific lip, hand, head, and face motions. This offers a proof of concept demonstrating that RF signals can detect such micro-movements.

This thesis investigates the potential of cutting-edge MM assistive technology to improve verbal and nonverbal communication within the deaf community. The work investigates how these RF-based MM hearing aids outperform previous technologies and introduces novel concepts. It opens the door for future advances in assistive hearing equipment while recognising significant success in this field. It demonstrates the growth and significance of hearing aid technology in overcoming communication obstacles and shows how these technologies can help the deaf communicate more effectively through in-depth analysis, making the environment easier and more accessible.

7.2 Limitation

Herein, there are some limitations of RF sensing-based technology, such as those listed in the following section.

1. **Limited Number Of Data:** The low number of participants and the limited amount of data can significantly affect the generalisation of the machine learning algorithm, whether it involves classical or deep learning methods.

2. **Regulatory Constraints:** The operation of RF sensors is subject to regulatory constraints, including limits on transmission power and frequency band usage. These regulations can limit the design and deployment of RF sensing technologies, particularly in regions with strict wireless communication standards.

3. **Susceptibility To Interference:** RF sensors can be highly susceptible to interference from other wireless devices operating in the same frequency band. This interference can degrade the quality of the sensed data, affecting the accuracy and reliability of the applications that depend on it.

7.3 Future Work

Future research efforts will focus on the many aspects of future MM hearing impairments using AI techniques, such as the following.

1. **Diverse Environmental Data Collections:** This study proposes a concept for a future MM hearing aid device, aimed at benefiting deaf individuals through the use of RF signals, while also pinpointing areas for further enhancement. The initial research was conducted with meticulous attention, focusing on the collection of CSI, Doppler shift, and RSSI data from a predetermined set of locations. The next phase of research will aim to enhance the detection process by compiling a more extensive dataset, which will include data gathered from a broader range of positions and orientations. This expanded data collection will be instrumental in enabling future MM hearing aids to function effectively in diverse environments.

2. **Advancements in Signal Processing:** Future research should prioritise advancements in signal processing, a key factor in reducing noise within CSI, Doppler shift, and RSSI datasets. These improvements are essential for achieving more accurate movement classification. Moreover, minimising external noise is vital for optimal system performance across a variety of settings. This includes adapting to different indoor environments with varying room configurations, furniture arrangements, and architectural designs, as well as tackling the unique challenges presented by outdoor conditions. Such modifications are imperative to ensure the effective functioning of future MM hearing aids in both indoor and outdoor environments.

3. **Multi-Target Activity Monitoring:** Future research should focus on advancing multi-target activity monitoring, especially in enhancing the accuracy of CSI, Doppler shift, and RSSI datasets. Such improvements are essential for precise movement classification in environments with varied noise sources. The next steps will involve developing more sophisticated algorithms and sensor technologies to effectively differentiate and track multiple targets. These advancements are not only crucial for achieving higher accuracy in controlled settings but are also key to enhancing the system's adaptability and reliability in dynamic, real-world scenarios. The ultimate goal is to establish a robust multi-target activity monitoring system that can efficiently function in a range of environments, from densely populated urban areas to complex indoor spaces.

4. **Real-Time Activity Monitoring:** Future research efforts will focus on significantly enhancing real-time monitoring of all micro-movements, covering activities ranging from lip reading and hand gesture recognition to head and facial expression tracking. The goal is to establish a foundational support system for the development of advanced MM hearing aid devices. These devices, utilising RF-based technology, are anticipated to be cost-effective and highly versatile, functioning seamlessly across various environments. A key aspect of our advancement strategy involves integrating sophisticated algorithms with state-of-the-art sensor technology. Such integration is essential for ensuring that MM hearing aids can accurately interpret and respond in real time to the subtlest human gestures and expressions, thus providing a more intuitive and natural user experience.

5. **Advancements in ML and DL Algorithms:** Further improvements will be made in ML and DL to enhance the accuracy of RF sensing, particularly as the complexity of CSI, Doppler shift, and RSSI increases with more varied data collection. Our comprehensive analysis will assess ML and DL effectiveness in RF sensing, focusing on model accuracy and rapid data processing, critical in real-time applications. Additionally, we will focus on developing adaptive algorithms to continually improve performance in dynamically changing environments.

Bibliography

[1] WHO. *Deafness and hearing loss*. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. Accessed: 10 Jan 2024.

[2] 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.

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

[4] Kamil S Talha et al. “Speech analysis based on image information from lip movement”. In: *IOP Conference Series: Materials Science and Engineering*. Vol. 53. 1. IOP Publishing. 2013, p. 012016.

[5] Leon Rothkrantz. “Lip-reading by surveillance cameras”. In: *2017 Smart City Symposium Prague (SCSP)*. IEEE. 2017, pp. 1–6.

[6] Fatemeh Sadat Lesani, Faranak Fotouhi Ghazvini, and Rouhollah Dianat. “Mobile phone security using automatic lip reading”. In: *2015 9th International Conference on e-Commerce in Developing Countries: With focus on e-Business (ECDC)*. IEEE. 2015, pp. 1–5.

[7] Dimitris Kastaniotis, Dimitrios Tsourounis, and Spiros Fotopoulos. “Lip Reading modeling with Temporal Convolutional Networks for medical support applications”. In: *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE. 2020, pp. 366–371.

[8] Meiqi Zhuang et al. “Highly robust and wearable facial expression recognition via deep-learning-assisted, soft epidermal electronics”. In: *Research* (2021).

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

[10] K Neeraja, K Srinivas Rao, and G Praneeth. “Deep Learning based Lip Movement Technique for Mute”. In: *2021 6th International Conference on Communication and Electronics Systems (ICCES)*. IEEE. 2021, pp. 1446–1450.

[11] Guanhua Wang et al. “We can hear you with Wi-Fi!” In: *Proceedings of the 20th annual international conference on Mobile computing and networking*. 2014, pp. 593–604.

[12] M Mahbubur Rahman et al. “Word-level sign language recognition using linguistic adaptation of 77 GHz FMCW radar data”. In: *2021 IEEE Radar Conference (RadarConf21)*. IEEE. 2021, pp. 1–6.

[13] James McCleary et al. “Sign Language Recognition using micro-Doppler and Explainable Deep Learning”. In: *2021 IEEE Radar Conference (RadarConf21)*. IEEE. 2021, pp. 1–6.

[14] Yukang Yan et al. “Privatetalk: Activating voice input with hand-on-mouth gesture detected by bluetooth earphones”. In: *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 2019, pp. 1013–1020.

[15] Jingxian Wang et al. “Rfid tattoo: A wireless platform for speech recognition”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.4 (2019), pp. 1–24.

[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] Wenjie Xu et al. “Distributed Localization of a RF Target in NLOS Environments”. In: *IEEE Journal on Selected Areas in Communications* 33.7 (2015), pp. 1317–1330. DOI: `10.1109/JSAC.2015.2430152`.

[18] Neal Patwari and Joey Wilson. “RF Sensor Networks for Device-Free Localization: Measurements, Models, and Algorithms”. In: *Proceedings of the IEEE* 98.11 (2010), pp. 1961–1973. DOI: `10.1109/JPROC.2010.2052010`.

[19] Syed Aziz Shah et al. “Rf sensing for healthcare applications”. In: *Backscattering and RF Sensing for Future Wireless Communication* (2021), pp. 157–177.

[20] Mandar Gogate, Kia Dashtipour, and Amir Hussain. “Solving the cocktail party problem using Multi-modal Hearing Assistive Technology Prototype”. In: *The Journal of the Acoustical Society of America* 154.4_supplement (2023), A36–A36.

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

[22] 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.

[23] 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).

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

[25] Yu Gu et al. “WiFE: WiFi and Vision based Unobtrusive Emotion Recognition via Gesture and Facial Expression”. In: *IEEE Transactions on Affective Computing* (2023).

[26] Sijia Xu et al. “Force-induced ion generation in zwitterionic hydrogels for a sensitive silent-speech sensor”. In: *Nature Communications* 14.1 (2023), p. 219.

[27] Taemin Kim et al. “Ultrathin crystalline-silicon-based strain gauges with deep learning algorithms for silent speech interfaces”. In: *Nature communications* 13.1 (2022), p. 5815.

[28] Penghao Dong et al. “Electromyogram-Based Lip-Reading via Unobtrusive Dry Electrodes and Machine Learning Methods”. In: *Small* 19.17 (2023), p. 2205058.

[29] Yijia Lu et al. “Decoding lip language using triboelectric sensors with deep learning”. In: *Nature Communications* 13.1 (2022), pp. 1–12.

[30] Vasileios Tsouvalas, Aaqib Saeed, and Tanir Ozcelebi. “Federated self-training for semi-supervised audio recognition”. In: *ACM Transactions on Embedded Computing Systems* 21.6 (2022), pp. 1–26.

[31] Ahsan Adeel et al. “Lip-reading driven deep learning approach for speech enhancement”. In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 5.3 (2019), pp. 481–490.

[32] Da Wu, Ling Ding, Shejie Lu, et al. “Research on speech recognition acceleration technology based on embedded platform”. In: *2019 Chinese Control Conference (CCC)*. IEEE. 2019, pp. 3663–3668.

[33] 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.

[34] Pingchuan Ma, Stavros Petridis, and Maja Pantic. “Visual speech recognition for multiple languages in the wild”. In: *Nature Machine Intelligence* 4.11 (2022), pp. 930–939.

[35] Hanan A. Mahmoud, Fahad Bin Muhaya, and Alaaeldin Hafez. “Lip Reading Based Surveillance System”. In: *2010 5th International Conference on Future Information Technology*. 2010, pp. 1–4. DOI: 10.1109/FUTURETECH.2010.5482688.

[36] Fatemeh Sadat Lesani, Faranak Fotouhi Ghazvini, and Rouhollah Dianat. “Mobile phone security using automatic lip reading”. In: *2015 9th International Conference on e-Commerce in Developing Countries: With focus on e-Business (ECDC)*. 2015, pp. 1–5. DOI: 10.1109/ECDC.2015.7156322.

[37] Gerasimos Potamianos et al. “Audio-Visual Automatic Speech Recognition: An Overview”. In: *Issues in audio-visual speech processing* (Jan. 2004).

[38] Kamil S. Talha et al. “Speech Analysis Based On Image Information from Lip Movement Speech Analysis Based On Image Information from Lip Movement”. In: vol. 53. July 2013. DOI: 10.1088/1757-899X/53/1/012016.

[39] Dimitris Kastaniotis, Dimitrios Tsourounis, and Spiros Fotopoulos. “Lip Reading modeling with Temporal Convolutional Networks for medical support applications”. In: *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. 2020, pp. 366–371. DOI: 10.1109/CISP-BMEI51763.2020.9263634.

[40] Praneeth Nemani et al. “Deep learning based holistic speaker independent visual speech recognition”. In: *IEEE Transactions on Artificial Intelligence* (2022).

[41] Yue Ma et al. “Speech Recovery Based On Auditory Radar and Webcam”. In: *2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)*. Vol. 1. 2019, pp. 1–3. DOI: 10.1109/IMBIOC.2019.8777840.

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

[43] Guanhua Wang et al. “We can hear you with Wi-Fi!” In: *Proceedings of the 20th annual international conference on Mobile computing and networking*. 2014, pp. 593–604.

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

[45] Jingxian Wang et al. “Speech Recognition Using RFID Tattoos”. In: *IJCAI*. 2021, pp. 4849–4853.

[46] Yutong Gu et al. “American Sign Language Alphabet Recognition Using Inertial Motion Capture System with Deep Learning”. In: *Inventions* 7.4 (2022), p. 112.

[47] Liufeng Fan et al. “Smart-Data-Glove-Based Gesture Recognition for Amphibious Communication”. In: *Micromachines* 14.11 (2023), p. 2050.

[48] M Neela Harish and S Poonguzhali. “Gesture Recognition Glove for Speech and Hearing Impaired People”. In: *Renewable Energy Optimization, Planning and Control: Proceedings of ICRTE 2022*. Springer, 2023, pp. 81–93.

[49] Celestine Preetham et al. “Hand talk-implementation of a gesture recognizing glove”. In: *2013 Texas Instruments India Educators’ Conference*. IEEE. 2013, pp. 328–331.

[50] Md Ahasan Atick Faisal et al. “Exploiting domain transformation and deep learning for hand gesture recognition using a low-cost dataglove”. In: *Scientific Reports* 12.1 (2022), p. 21446.

[51] Jieming Pan et al. “Hybrid-flexible bimodal sensing wearable glove system for complex hand gesture recognition”. In: *ACS sensors* 6.11 (2021), pp. 4156–4166.

[52] Huihui Wang et al. “MEMS Devices-Based Hand Gesture Recognition via Wearable Computing”. In: *Micromachines* 14.5 (2023), p. 947.

[53] Saba Jadooki et al. “Fused features mining for depth-based hand gesture recognition to classify blind human communication”. In: *Neural Computing and Applications* 28.11 (2017), pp. 3285–3294.

[54] Britta Bauer and Hermann Hienz. “Relevant features for video-based continuous sign language recognition”. In: *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. IEEE. 2000, pp. 440–445.

[55] Mohamed Mohandes, Mohamed Deriche, and Junzhao Liu. “Image-based and sensor-based approaches to Arabic sign language recognition”. In: *IEEE transactions on human-machine systems* 44.4 (2014), pp. 551–557.

[56] Lionel Pigou et al. “Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video”. In: *International Journal of Computer Vision* 126.2 (2018), pp. 430–439.

[57] Natalia Neverova et al. “Multi-scale deep learning for gesture detection and localization”. In: *European conference on computer vision*. Springer. 2014, pp. 474–490.

[58] RAHUL D RAJ and ASHISH JASUJA. “British sign language recognition using HOG”. In: *2018 IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS)*. IEEE. 2018, pp. 1–4.

[59] Wisnu Aditya et al. “Novel Spatio-Temporal Continuous Sign Language Recognition Using an Attentive Multi-Feature Network”. In: *Sensors* 22.17 (2022). ISSN: 1424-8220. DOI: 10.3390/s22176452. URL: https://www.mdpi.com/1424-8220/22/17/6452.

[60] Sarah Alyami, Hamzah Luqman, and Mohammad Hammoudeh. “Isolated Arabic Sign Language Recognition Using A Transformer-based Model and Landmark Keypoints”. In: *ACM Transactions on Asian and Low-Resource Language Information Processing* (2023).

[61] James McCleary et al. “Sign Language Recognition using micro-Doppler and Explainable Deep Learning”. In: *2021 IEEE Radar Conference (RadarConf21)*. IEEE. 2021, pp. 1–6.

[62] M Mahbubur Rahman et al. “Word-level sign language recognition using linguistic adaptation of 77 GHz FMCW radar data”. In: *2021 IEEE Radar Conference (RadarConf21)*. IEEE. 2021, pp. 1–6.

[63] Sevgi Z Gurbuz et al. “American sign language recognition using rf sensing”. In: *IEEE Sensors Journal* 21.3 (2020), pp. 3763–3775.

[64] Xing Wang et al. "Chinese sign language recognition based on multi-view deep neural network for millimeter-wave radar". In: *Conference on Infrared, Millimeter, Terahertz Waves and Applications (IMT2022)*. Vol. 12565. SPIE. 2023, p. 1256502.

[65] Beichen Li et al. "Sign language/gesture recognition based on cumulative distribution density features using UWB radar". In: *IEEE transactions on instrumentation and measurement* 70 (2021), pp. 1–13.

[66] Gavin MacLaughlin, Jack Malcolm, and Syed Ali Hamza. "Multi Antenna Radar System for American Sign Language (ASL) Recognition Using Deep Learning". In: *arXiv preprint arXiv:2203.16624* (2022).

[67] Jiacheng Shang and Jie Wu. "A robust sign language recognition system with multiple Wi-Fi devices". In: *Proceedings of the Workshop on Mobility in the Evolving Internet Architecture*. 2017, pp. 19–24.

[68] Sijie Ji et al. "Construct 3D Hand Skeleton with Commercial WiFi". In: (2023).

[69] Guiping Lin et al. "Human activity recognition using smartphones with WiFi signals". In: *IEEE Transactions on Human-Machine Systems* 53.1 (2022), pp. 142–153.

[70] Zhongjian Gao et al. "A Multitask Sign Language Recognition System Using Commodity Wi-Fi". In: *Mobile Information Systems* 2023 (2023).

[71] Nengbo Zhang et al. "Wi-Phrase: deep residual-multihead model for wifi sign language phrase recognition". In: *IEEE Internet of Things Journal* 9.18 (2022), pp. 18015–18027.

[72] Huanyuan Xu et al. "RF-CSign: A Chinese Sign Language Recognition System Based on Large Kernel Convolution and Normalization-Based Attention". In: *IEEE Access* 11 (2023), pp. 133767–133780.

[73] Yongpan Zou et al. "Grfid: A device-free rfid-based gesture recognition system". In: *IEEE Transactions on Mobile Computing* 16.2 (2016), pp. 381–393.

[74] Zijing Ma et al. "RF-Siamese: approaching accurate rfid gesture recognition with one sample". In: *IEEE Transactions on Mobile Computing* (2022).

[75] Chuanxin Zhao et al. "RFID-Based Human Action Recognition through Spatiotemporal Graph Convolutional Neural Network". In: *IEEE Internet of Things Journal* (2023).

[76] Cao Dian et al. "Towards domain-independent complex and fine-grained gesture recognition with RFID". In: *Proceedings of the ACM on Human-Computer Interaction* 4.ISS (2020), pp. 1–22.

[77] Minghan Liu et al. "Preliminary Results on Sensing Pillow to Monitor Head Movement using strain sensing threads". In: *2022 IEEE Sensors*. IEEE. 2022, pp. 1–4.

[78] Aura Ximena González-Cely, Mauro Callejas-Cuervo, and Teodiano Bastos-Filho. “Wheelchair prototype controlled by position, speed and orientation using head movement”. In: *HardwareX* 11 (2022), e00306.

[79] Dongwook Lee et al. “Drowsy Driving Detection Based on the Driver’s Head Movement using Infrared Sensors”. In: *2008 Second International Symposium on Universal Communication*. 2008, pp. 231–236. DOI: 10.1109/ISUC.2008.76.

[80] Yiwen Jiang et al. “Head motion classification using thread-based sensor and machine learning algorithm”. In: *Scientific reports* 11.1 (2021), pp. 1–11.

[81] Amer Al-Rahayfeh and Miad Faezipour. “Eye tracking and head movement detection: A state-of-art survey”. In: *IEEE journal of translational engineering in health and medicine* 1 (2013), pp. 2100212–2100212.

[82] Thanarat Horprasert, Yaser Yacoob, and Larry S Davis. “An anthropometric shape model for estimating head orientation”. In: *3rd International Workshop on Visual Form, Capri, Italy*. 1997.

[83] Euclides N Arcoverde Neto et al. “Real-time head pose estimation for mobile devices”. In: *International Conference on Intelligent Data Engineering and Automated Learning*. Springer. 2012, pp. 467–474.

[84] Euclides N Arcoverde Neto et al. “Enhanced real-time head pose estimation system for mobile device”. In: *Integrated Computer-Aided Engineering* 21.3 (2014), pp. 281–293.

[85] Fairouz Merrouche and Nadia Baha. “Fall detection using head tracking and centroid movement based on a depth camera”. In: *Proceedings of the International Conference on Computing for Engineering and Sciences*. 2017, pp. 29–34.

[86] Yafei Wang et al. “Continuous driver’s gaze zone estimation using rgb-d camera”. In: *Sensors* 19.6 (2019), p. 1287.

[87] Muneeba Raja et al. “3D head motion detection using millimeter-wave Doppler radar”. In: *IEEE Access* 8 (2020), pp. 32321–32331.

[88] Chuanwei Ding et al. “Inattentive driving behavior detection based on portable FMCW radar”. In: *IEEE Transactions on Microwave Theory and Techniques* 67.10 (2019), pp. 4031–4041.

[89] Drew G Bresnahan and Yang Li. “Classification of driver head motions using a mm-wave FMCW radar and deep convolutional neural network”. In: *IEEE Access* 9 (2021), pp. 100472–100479.

[90] Hong Nhung Nguyen et al. “One-shot learning-based driver’s head movement identification using a millimetre-wave radar sensor”. In: *IET Radar, Sonar & Navigation* 16.5 (2022), pp. 825–836.

[91] Yuqing Bu et al. “Multidomain Fusion Method for Human Head Movement Recognition”. In: *IEEE Transactions on Instrumentation and Measurement* 72 (2023), pp. 1–8.

[92] Shunqiao Sun, Athina P Petropulu, and H Vincent Poor. “MIMO radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges”. In: *IEEE Signal Processing Magazine* 37.4 (2020), pp. 98–117.

[93] Kai Chen et al. “HeadSee: Device-free head gesture recognition with commodity RFID”. In: *Peer-to-Peer Networking and Applications* 15.3 (2022), pp. 1357–1369.

[94] Xuanke He et al. “RFID based non-contact human activity detection exploiting cross polarization”. In: *IEEE Access* 8 (2020), pp. 46585–46595.

[95] Guilherme Figueiredo, Brandon Hubbs, and Adarsh D Radadia. “Monitoring Head Orientation Using Passive RFID Tags”. In: *IEEE Journal of Radio Frequency Identification* (2023).

[96] Chao Yang, Xuyu Wang, and Shiwen Mao. “RFID-based driving fatigue detection”. In: *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE. 2019, pp. 1–6.

[97] Zihan Yan et al. “Emoglass: an end-to-end ai-enabled wearable platform for enhancing self-awareness of emotional health”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–19.

[98] Denys Matthies, Nastaran Saffaryazdi, and Mark Billinghurst. “Wearable Sensing of Facial Expressions and Head Gestures”. In: *NordiCHI’22 Workshop. https://doi. org/10.13140/RG*. Vol. 2. 26960.38408. 2022, p. 2.

[99] Dhruv Verma et al. “Expressear: Sensing fine-grained facial expressions with earables”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.3 (2021), pp. 1–28.

[100] Katsutoshi Masai et al. “Facial expression recognition in daily life by embedded photo reflective sensors on smart eyewear”. In: *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 2016, pp. 317–326.

[101] Denys JC Matthies, Bernhard A Strecker, and Bodo Urban. “Earfieldsensing: A novel in-ear electric field sensing to enrich wearable gesture input through facial expressions”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017, pp. 1911–1922.

[102] Niall O’Mahony et al. “Deep learning vs. traditional computer vision”. In: *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*. Springer. 2020, pp. 128–144.

[103] Yirui Wu et al. “Edge-ai-driven framework with efficient mobile network design for facial expression recognition”. In: *ACM Transactions on Embedded Computing Systems* 22.3 (2023), pp. 1–17.

[104] E Komagal and B Yogameena. “PTZ-Camera-Based Facial Expression Analysis using Faster R-CNN for Student Engagement Recognition”. In: *Computer Vision and Machine Intelligence Paradigms for SDGs: Select Proceedings of ICRTAC-CVMIP 2021*. Springer, 2023, pp. 1–14.

[105] Muhammad Hameed Siddiqi et al. “Depth camera-based facial expression recognition system using multilayer scheme”. In: *IETE Technical Review* 31.4 (2014), pp. 277–286.

[106] Myunghoon Suk and Balakrishnan Prabhakaran. “Real-time mobile facial expression recognition system-a case study”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2014, pp. 132–137.

[107] Lucas D Terissi, Juan Carlos Gómez, et al. “3D Head Pose and Facial Expression Tracking using a Single Camera.” In: *J. Univers. Comput. Sci.* 16.6 (2010), pp. 903–920.

[108] Mostafa Hamed Abdalla et al. “Design and Implementation of Digital IF Waveform Generation, Acquisition, and Receiver circuits for Radar Systems Applications”. In: *2020 12th International Conference on Electrical Engineering (ICEENG)*. 2020, pp. 217–222. DOI: 10.1109/ICEENG45378.2020.9171698.

[109] Syed Aziz Shah and Francesco Fioranelli. “RF Sensing Technologies for Assisted Daily Living in Healthcare: A Comprehensive Review”. In: *IEEE Aerospace and Electronic Systems Magazine* 34.11 (2019), pp. 26–44. DOI: 10.1109/MAES.2019.2933971.

[110] Xiaochao Dang et al. “Wireless Sensing Technology Combined with Facial Expression to Realize Multimodal Emotion Recognition”. In: *Sensors* 23.1 (2022), p. 338.

[111] Li Zhang et al. “Non-contact Dual-modality emotion recognition system by CW radar and RGB camera”. In: *IEEE Sensors Journal* 21.20 (2021), pp. 23198–23212.

[112] Carolina Gouveia et al. “Study on the usage feasibility of continuous-wave radar for emotion recognition”. In: *Biomedical Signal Processing and Control* 58 (2020), p. 101835.

[113] Xiaochao Dang, Zetong Chen, and Zhanjun Hao. “Emotion recognition method using millimetre wave radar based on deep learning”. In: *IET Radar, Sonar & Navigation* 16.11 (2022), pp. 1796–1808.

[114] Yanjiao Chen et al. “WiFace: facial expression recognition using Wi-Fi signals”. In: *IEEE Transactions on Mobile Computing* 21.1 (2020), pp. 378–391.

[115] Yu Gu et al. “Wife: Wifi and vision based intelligent facial-gesture emotion recognition”. In: *arXiv preprint arXiv:2004.09889* (2020).

[116] Yu Gu et al. "EmoSense: computational intelligence driven emotion sensing via wireless channel data". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 4.3 (2019), pp. 216–226.

[117] Ahsan Noor Khan et al. "Deep learning framework for subject-independent emotion detection using wireless signals". In: *Plos one* 16.2 (2021), e0242946.

[118] Leyuan Jia et al. "BeAware: Convolutional neural network (CNN) based user behavior understanding through WiFi channel state information". In: *Neurocomputing* 397 (2020), pp. 457–463.

[119] Noor Amalina Ramli, Anis Nurashikin Nordin, and Norsinnira Zainul Azlan. "Development of low cost screen-printed piezoresistive strain sensor for facial expressions recognition systems". In: *Microelectronic Engineering* 234 (2020), p. 111440.

[120] Weiye Xu et al. "Anti-Spoofing Facial Authentication Based on COTS RFID". In: *IEEE Transactions on Mobile Computing* (2023).

[121] Filippo Battaglia, Giancarlo Iannizzotto, and Lucia Lo Bello. "A person authentication system based on RFID tags and a cascade of face recognition algorithms". In: *IEEE Transactions on Circuits and Systems for Video Technology* 27.8 (2016), pp. 1676–1690.

[122] Chengwen Luo et al. "Rfaceid: Towards rfid-based facial recognition". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.4 (2021), pp. 1–21.

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

[124] 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).

[125] Koosha Sharifani and Mahyar Amini. "Machine Learning and Deep Learning: A Review of Methods and Applications". In: *World Information Technology and Engineering Journal* 10.07 (2023), pp. 3897–3904.

[126] 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.

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

[128] Ku Chhaya A Khanzode and Ravindra D Sarode. "Advantages and Disadvantages of Artificial Intelligence and Machine Learning: A Literature Review". In: *International Journal of Library & Information Science (IJLIS)* 9.1 (2020), p. 3.

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

[130] 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.

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

[132] Gaurav Menghani. “Efficient deep learning: A survey on making deep learning models smaller, faster, and better”. In: *ACM Computing Surveys* 55.12 (2023), pp. 1–37.

[133] Ce Zheng et al. “Deep learning-based human pose estimation: A survey”. In: *ACM Computing Surveys* 56.1 (2023), pp. 1–37.

[134] Mingle Xu et al. “A comprehensive survey of image augmentation techniques for deep learning”. In: *Pattern Recognition* (2023), p. 109347.

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

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

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

[138] 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.

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

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

[141] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[142] 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.

[143] Ibrahim Alnujaim et al. “Hand Gesture Recognition Using Input Impedance Variation of Two Antennas with Transfer Learning”. In: *IEEE Sensors Journal* 18.10 (2018), pp. 4129–4135. DOI: `10.1109/JSEN.2018.2820000`.

[144] Shahin Amiriparian et al. "“Are You Playing a Shooter Again?!” Deep Representation Learning for Audio-Based Video Game Genre Recognition". In: *IEEE Transactions on Games* 12.2 (2020), pp. 145–154. DOI: 10.1109/TG.2019.2894532.

[145] Dustin P Fairchild et al. "Through-the-wall micro-Doppler signatures". In: *Chen, VC, Tahmoush, D., Miceli, WJ (Eds.)* (2014).

[146] Yong Wu et al. "Convolution Neural Network based Transfer Learning for Classification of Flowers". In: *2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP)*. 2018, pp. 562–566. DOI: 10.1109/SIPROCESS.2018.8600536.

[147] *MS Windows NT Kernel Description*. http://https://machinelearningmastery.com/transfer-learning-for-deep-learning/.htm. Accessed: 2019-09-16.

[148] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[149] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[150] Christian Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[151] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[152] KV Seshagiri Rao, Pavel V Nikitin, and Sander F Lam. "Impedance matching concepts in RFID transponder design". In: *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*. IEEE. 2005, pp. 39–42.

[153] MS Yeoman and MA O'neill. "Impedance matching of tag antenna to maximize RFID read ranges & design optimization". In: *2014 COMSOL Conference, Cambridge, UK*. 2014.

[154] Daniel Dobkin. *The RF in RFID: UHF RFID in practice*. Newnes, 2012.

[155] Lubna et al. "IoT-Enabled Vacant Parking Slot Detection System Using Inkjet-Printed RFID Tags". In: *IEEE Sensors Journal* 23.7 (2023), pp. 7828–7835.

[156] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.

[157] Leif E Peterson. "K-nearest neighbor". In: *Scholarpedia* 4.2 (2009), p. 1883.

[158] Shunjie Han, Cao Qubo, and Han Meng. "Parameter selection in SVM with RBF kernel function". In: *World Automation Congress 2012*. IEEE. 2012, pp. 1–4.

[159] Gary F Simons and Charles D Fennig. "Ethnologue: Languages of Honduras". In: (2017).

[160] Jordan Fenlon and Erin Wilkinson. “Sign languages in the world”. In: *Sociolinguistics and Deaf communities* (2015), pp. 5–28.

[161] Dustin P Fairchild et al. “Through-the-wall micro-Doppler signatures”. In: *Chen, VC, Tahmoush, D., Miceli, WJ (Eds.)* (2014).

[162] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[163] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[164] Kashif Ahmad and Nicola Conci. “How deep features have improved event recognition in multimedia: A survey”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15.2 (2019), pp. 1–27.

[165] Forrest N Iandola et al. “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size”. In: *arXiv preprint arXiv:1602.07360* (2016).

[166] Aravind Krishnaswamy Rangarajan and Raja Purushothaman. “Disease classification in eggplant using pre-trained VGG16 and MSVM”. In: *Scientific reports* 10.1 (2020), pp. 1–11.

[167] Bo Xiao et al. “Head motion modeling for human behavior analysis in dyadic interaction”. In: *IEEE transactions on multimedia* 17.7 (2015), pp. 1107–1119.

[168] Chunlin Zhao et al. “Multivariate autoregressive models and kernel learning algorithms for classifying driving mental fatigue based on electroencephalographic”. In: *Expert Systems with Applications* 38.3 (2011), pp. 1859–1865.

[169] Gianluca Borghini et al. “Assessment of mental fatigue during car driving by using high resolution EEG activity and neurophysiologic indices”. In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2012, pp. 6442–6445.

[170] Monagi H Alkinani, Wazir Zada Khan, and Quratulain Arshad. “Detecting human driver inattentive and aggressive driving behavior using deep learning: Recent advances, requirements and open challenges”. In: *Ieee Access* 8 (2020), pp. 105008–105030.

[171] Xuesong Wang and Chuan Xu. “Driver drowsiness detection based on non-intrusive metrics considering individual specifics”. In: *Accident Analysis & Prevention* 95 (2016), pp. 350–357.

[172] Grégoire S Larue, Andry Rakotonirainy, and Anthony N Pettitt. “Predicting reduced driver alertness on monotonous highways”. In: *IEEE Pervasive Computing* 14.2 (2015), pp. 78–85.

[173] Sinan Kaplan et al. “Driver behavior analysis for safe driving: A survey”. In: *IEEE Transactions on Intelligent Transportation Systems* 16.6 (2015), pp. 3017–3032.

[174] Shahzeb Ansari et al. “Driver mental fatigue detection based on head posture using new modified reLU-BiLSTM deep neural network”. In: *IEEE Transactions on Intelligent Transportation Systems* (2021).

[175] Partha Chakraborty, Mohammad Abu Yousuf, and Saifur Rahman. “Predicting level of visual focus of human’s attention using machine learning approaches”. In: *Proceedings of international conference on trends in computational and cognitive engineering*. Springer. 2021, pp. 683–694.

[176] T Kujani and V Dhilip Kumar. “Head movements for behavior recognition from real time video based on deep learning ConvNet transfer learning”. In: *Journal of Ambient Intelligence and Humanized Computing* (2021), pp. 1–15.

[177] A Enis Cetin et al. “Review of signal processing applications of Pyroelectric Infrared (PIR) sensors with a focus on respiration rate and heart rate detection”. In: *Digital Signal Processing* 119 (2021), p. 103247.

[178] Sarah Masud Preum et al. “A review of cognitive assistants for healthcare: Trends, prospects, and future directions”. In: *ACM Computing Surveys (CSUR)* 53.6 (2021), pp. 1–37.

[179] Henry Dsouza et al. “Flexible, self-powered sensors for estimating human head kinematics relevant to concussions”. In: *Scientific reports* 12.1 (2022), pp. 1–8.

[180] Sharifa Alghowinem et al. “Head pose and movement analysis as an indicator of depression”. In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE. 2013, pp. 283–288.

[181] Chien-Hsu Chen, I-Jui Lee, and Ling-Yi Lin. “Augmented reality-based video-modeling storybook of nonverbal facial cues for children with autism spectrum disorder to improve their perceptions and judgments of facial expressions and emotions”. In: *Computers in Human Behavior* 55 (2016), pp. 477–485.

[182] Anil Audumbar Pise et al. “Methods for Facial Expression Recognition with Applications in Challenging Situations”. In: *Computational Intelligence and Neuroscience* 2022 (2022).

[183] Klaus R Scherer, Rainer Banse, and Harald G Wallbott. “Emotion inferences from vocal expression correlate across languages and cultures”. In: *Journal of Cross-cultural psychology* 32.1 (2001), pp. 76–92.

[184] Lisa Feldman Barrett et al. “Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements”. In: *Psychological science in the public interest* 20.1 (2019), pp. 1–68.

[185] Wataru Sato and Sakiko Yoshikawa. “Spontaneous facial mimicry in response to dynamic facial expressions”. In: *Cognition* 104.1 (2007), pp. 1–18.

[186] Chiara Filippini et al. “Thermal infrared imaging-based affective computing and its application to facilitate human robot interaction: A review”. In: *Applied Sciences* 10.8 (2020), p. 2924.

[187] M. I. Skolnik. *Radar Handbook*. Third Edition. McGraw-Hill Education, 2008.

[188] E. Levanon N. Mozeson. *Radar Signals*. Wiley, 2004.

[189] *1. Supervised learning — scikit-learn.org*. `https://scikit-learn.org/stable/supervised_learning.html#supervised-learning`. [Accessed 25-02-2024].

[190] Shuyu Shi et al. “Accurate location tracking from CSI-based passive device-free probabilistic fingerprinting”. In: *IEEE Transactions on Vehicular Technology* 67.6 (2018), pp. 5217–5230.

[191] Jiangang Hao and Tin Kam Ho. “Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language”. In: *Journal of Educational and Behavioral Statistics* 44.3 (2019), pp. 348–361.