

Efficient Multimodal Human Activity Recognition (HAR) solutions using Video, Audio, and Sensors

by

Tair Imankulov

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science

at the

NAZARBAYEV UNIVERSITY

June 2025

© Nazarbayev University 2025. All rights reserved.

Author
Department of Computer Science
Apr 11, 2025

Certified by
Adnan Yazici
Chair and Associate Professor
Thesis Supervisor

Accepted by
Vassilios D.Tourassis
Dean, School of Engineering and Digital Sciences

Efficient Multimodal Human Activity Recognition (HAR) solutions using Video, Audio, and Sensors

by

Tair Imankulov

Submitted to the Department of Computer Science
on Apr 11, 2025, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science

Abstract

Research on Human Activity Recognition (HAR) in smart home settings has become essential, opening up possibilities for automation, security, and medical applications. To improve recognition accuracy and real-time feasibility, this study integrates video, audio, and sensor data into a robust multimodal HAR framework using the YouHome ADL dataset. To successfully integrate various data modalities, the methodology uses advanced feature extraction, privacy-preserving preprocessing, and a late fusion approaches.

To ensure effective representation learning for various modalities, the HAR system uses a YOLOv8n+MobileNetV2+BiLSTM for video, a BiLSTM+GRU+TCN model for ambient sensor data, CNN+GRU for inertial sensor data, and a BiLSTM for audio-based features classification. Class imbalances were also addressed by applying class merging and dataset balancing using SMOTE, which greatly increased classification accuracy and decreased bias toward overrepresented activities.

The final multimodal HAR model performed significantly better than baseline techniques, according to experimental evaluations. On the top of that, several ablation studies evaluated the trade-offs between privacy, computational costs, and performance. Although real-time viability on devices with limited resources was not addressed, the results confirm the possible real-time execution of the proposed systems. These findings demonstrate how crucial fusion approach, model optimization, and data preprocessing are to build an efficient, multimodal HAR systems.

Thesis Supervisor: Adnan Yazici

Title: Chair and Associate Professor

Acknowledgments

This thesis would not have been possible without the guidance, support, and encouragement of many individuals and institutions.

First and foremost, I want to sincerely thank Adnan Yazici, my advisor, for his invaluable guidance and unwavering support during this project. His knowledge and perceptive discussions have greatly influenced the course of this thesis.

I also would like to express my gratitude to Nazarbayev University for offering the tools and a necessary research environment that made this study possible.

I would especially like to thank my partner, family, and friends for their understanding and constant support during this journey. Their encouragement has been a constant source of inspiration, particularly in difficult times.

Contents

1	Introduction	13
2	Related works	17
2.1	Video-based Approaches	17
2.2	Audio-based Approaches	18
2.3	Sensor-based Approaches	19
2.4	Multimodal Approaches	20
2.5	Real-time/Edge-based approaches	21
3	Methodology	25
3.1	Datasets	25
3.1.1	YouHome ADL dataset	25
3.1.2	HWU-USP dataset	27
3.1.3	UCF-101	29
3.2	Preprocessing Data	30
3.2.1	Video processing	30
3.2.2	Audio processing	32
3.2.3	Ambient sensor data processing	32
3.2.4	Inertial sensor data processing	32
3.2.5	Dataset balancing	33
3.3	Proposed Method	33
3.3.1	Architecture and processing	33
3.3.2	Fusion algorithms	35

4	Results	37
4.1	Selecting YOLO model	37
4.2	Baseline accuracies on YouHome dataset	38
4.3	Improved Model Performances	39
4.3.1	Impact of Dataset Balancing	39
4.3.2	Impact of Merging similar classes	41
4.3.3	Fusion results	43
4.4	Evaluation on other datasets	44
4.4.1	HWU-USP	44
4.4.2	UCF-101	45
4.5	Comparative Evaluation with Prior Work	46
4.6	Ablation study	48
4.6.1	Lighter RGB model	48
4.6.2	Heavier RGB model	49
4.6.3	Performance comparisons with prior work	49
4.6.4	Audio Privacy Protection	53
4.7	Edge Device Compatibility Evaluation	53
5	Discussion	55
5.1	Evaluation of Individual Modalities	55
5.2	Fusion Evaluation	56
5.3	Ablation Study Analysis	57
5.4	Limitations	57
5.5	Real-Time HAR Applications	58
6	Conclusion	61

List of Figures

3-1	Sample images and home layouts taken from [14] (orange circle marks edge node positions)	26
3-2	Sample data recordings from HWU-USP dataset taken from [23], including RGB (a) and depth images (b), IMU (c), and binary ambient sensor data (d)	28
3-3	Sample videos from UCF-101 dataset taken from [25]	30
3-4	The diagram of the proposed framework	34
4-1	YOLOv8s performance	37
4-2	YOLOv8n performance	37
4-3	Baseline confusion matrix	38
4-4	Class Distribution Before and After SMOTE	39
4-5	Confusion matrix after applying SMOTE	40
4-6	Merged classes distribution before and and After SMOTE	42
4-7	Accuracy and Loss graphs of the video model	42
4-8	Accuracy and Loss graphs of the audio model	43
4-9	Accuracy and Loss graphs of the sensor model	43

List of Tables

2.1	Summary of Related Works in HAR	24
3.1	Activity distribution in dataset from [14]	27
4.1	Fusion results on YouHome ADL dataset	44
4.2	Individual and Fused results on HWU-USP Dataset	45
4.3	Individual and Fused results on UCF-101 Dataset	45
4.4	Performance comparison with prior work on YouHome ADL Dataset .	46
4.5	Performance comparison with prior work on HWU-USP Dataset . . .	47
4.6	Performance comparison with prior work on UCF-101 Dataset	48
4.7	Results of the proposed lighter RGB architecture	49
4.8	Results of the proposed heavier RGB architecture	49
4.9	Performance comparison with prior work on YouHome ADL Dataset (all methods)	50
4.10	Performance comparison with prior work on HWU-USP Dataset (all methods)	51
4.11	Performance comparison with prior work on UCF-101 Dataset (all methods)	52

Chapter 1

Introduction

Smart home settings have developed into intelligent systems that can monitor and react to human behaviors in real-time. One of the key challenges in developing these environments is human activity recognition (HAR), which involves detecting and classifying human activities based on different modalities, including video, audio, and sensors. HAR applications include home automation, security surveillance, health monitoring, personalized help, and eldercare [20]. HAR systems may achieve high recognition accuracy, robustness, and flexibility in a variety of settings by combining various modalities [1].

Wearable sensors, video-based deep learning models, and ambient sensor networks are examples of single-modality techniques that have been the primary foundation of traditional HAR systems. However, each method has its limitations:

- Although video-based HAR offers comprehensive visual information, it is affected by privacy issues, lighting conditions, and occlusions. [20].
- Audio-based HAR can identify certain activities through sound characteristics but it has trouble in loud settings, due to excessive noise affecting the classification performance [11].
- In order to track user motions, wearable sensor-based HAR uses physiological sensors, gyroscopes, and accelerometers. It is quite good at catching motion,

however, issues like user discomfort, battery limitations, and limited activity coverage affect the method’s robustness [13].

- Although ambient sensors are less invasive and power-efficient, they lack the granularity required for complicated activity recognition. [23].

Multimodal HAR, which combines complementary modalities and uses sensor fusion techniques to increase accuracy, robustness, and contextual awareness, has emerged as a possible solution to the aforementioned problems. By utilizing the advantages of each modality, multimodal HAR combines many data sources to enhance the HAR performance. Due to the complementary information provided by different sensor types, multimodal approaches have been found to perform better than unimodal ones [1]. Combining multiple modalities allows HAR systems to infer contextual relationships between actions, reducing false positives [13]. Additionally, certain modalities might not be available in real-world applications because of environmental factors (e.g., low illumination for video, noise interference for audio) [11]. Fusion techniques assist in making up for these lost modalities.

Despite such benefits, there still exists challenges in Multimodal HAR, including computational complexity, privacy issues, viewpoint variation, lighting conditions, and real-time HAR. A key objective of HAR is increasing computational efficiency since it can greatly speed up model deployment and enable real-time action recognition in video streams. But maintaining high accuracy while guaranteeing model efficiency is a tricky balance that is still the subject of current research [1]. Additionally, the issue of privacy still exists in multimodal environments, where either a camera or microphone can capture confidential information.

An issue with viewpoint variations comes from a difficulty to identify human activity from arbitrary angles, relying on a particular viewpoint. Similarly, varying illuminance in different scenes can be a major issue, making it difficult to create flexible and reliable HAR systems that can continue to function consistently in real-world situations when lighting fluctuates [1].

To tackle some of these challenges, researchers are increasingly focusing on devel-

oping lightweight deep learning models for HAR that can efficiently operate on embedded devices [20]. Tiny Machine Learning (TinyML) has emerged as a promising field, enabling on-device sensor data analysis with extremely low power consumption. TinyML techniques enhance the efficiency of deep learning models on low-power edge devices through methods such as pruning, quantization, and distilling pre-trained networks to enable efficient inference. While recent studies have explored TinyML for HAR across various research directions, there is currently no framework that efficiently integrates video, audio, and ambient sensor modalities on edge devices within the TinyML paradigm.

This study focuses on developing a computationally efficient, multimodal HAR pipeline, employing video, audio, and ambient sensors. Current existing models are challenging to implement for real-world deployment due to their high latency, power consumption, and inconsistent fusion output. By applying TinyML techniques in combination with lightweight models, the proposed approach will enhance real-time human activity detection in smart homes while preserving computing efficiency. Additionally, by using multimodal, multi-viewpoint dataset like UICI YouHome ADL [21], the approach will solve the issue with viewpoint variation, illumination, and occlusion, since the dataset captures multiple camera viewpoints as well as variable lighting and occlusions inside the home environment. Moreover, the proposed method deals with privacy issues by applying an effective privacy-protecting preprocessing of the video and audio data. Finally, the pipeline was evaluated on 2 other datasets, the Heriot-Watt University/University of Sao Paulo (HWU-USP) activities dataset, and UCF-101 to ensure the robustness of the architecture. The primary contributions of this work include:

1. This study implements an efficient HAR framework, enabling real-time activity recognition.
2. The study utilizes a multimodal approach including video, audio, and sensor (ambient/inertial) modalities, making the framework more robust and improving classification accuracy.

3. The framework was trained and evaluated on the YouHome ADL dataset, which contains multiple camera viewpoints, making the activity recognition less angle-reliable and additionally dealing with occlusions. This dataset also captures varying lighting conditions, dealing with illuminance issues.
4. The proposed approach was further validated on other publicly available datasets to verify the robustness of the architecture.
5. The huge imbalance in the dataset was solved using several balancing techniques.

This thesis’s remaining sections are organized as follows:

An overview of current HAR systems, with an emphasis on unimodal and multimodal techniques, and an examination of their advantages and disadvantages are covered in Chapter 2. Chapter 3 provides a thorough description of the utilized datasets, and covers preprocessing and processing procedures. The classification models, fusion methods, and dataset balancing techniques are also described in this part. Chapter 4 provides an examination of the baseline and improved results, evaluation on other datasets, and ablation studies. Then, chapter 5 will discuss the obtained outputs, restrictions, and possible directions for development. Lastly, Chapter 6 will conclude a study by outlining key contributions of the proposed multimodal HAR approach.

Chapter 2

Related works

2.1 Video-based Approaches

The potential of video-based HAR in surveillance, healthcare, and human-computer interaction has led to extensive research in this field. Deep learning has transformed the discipline by making it possible to reliably classify video sequences and automatically extract features. Recent developments have concentrated on enhancing feature representation and fusion methods to increase the effectiveness and precision of HAR models. [24] addressed the issues of computational complexity and real-time inference in group exercise situations by introducing a lightweight multiperson activity recognition system for robot-assisted healthcare applications. On benchmark datasets such as UESTC RGB-D and NTU RGB+D, their model achieves state-of-the-art classification performance by utilizing a one-dimensional multilayer long short-term memory (MLSTM) network. In order to address the dearth of publicly accessible datasets concentrating on physical activities among the elderly, they have assembled a new Routine Exercise Dataset (RED).

Similarly, a human-centric attention model with deep multiscale feature fusion (HCAF+DMFF+SMA) was presented by [9] for HAR in smart healthcare systems. Their method improves the spatial and temporal learning capacities of video-based HAR models by combining an attention-based fusion mechanism with a MobileNetV3-based feature extraction. The model outperformed state-of-the-art techniques in tests

on general activity datasets (HMDB51, UCF101, UCF50) and fall detection datasets (Multiple Camera Fall Dataset, UR Fall Detection). Notably, their residual BiLSTM system mitigated the vanishing gradient issue that frequently arises in deep sequential models, improving long-range temporal dependency learning.

Both works highlight the ongoing evolution of video-based HAR towards more efficient, real-time, and robust multimodal recognition systems. However, they exhibit key limitations that constrain their applicability in real-world smart home environments. [24] focus on multiperson activity recognition but do not evaluate their method on low-power edge devices, raising concerns about real-time feasibility. Additionally, their approach relies on deep learning models (MLSTM) with significant computational demands. Similarly, [9] introduces a deep fusion model (HCAF+DMFF+SMA) for HAR, but its high computational complexity and reliance on vision-based features make it challenging for real-time deployment.

2.2 Audio-based Approaches

Acoustic signals are used in audio-based HAR techniques to categorize human activities according to sound patterns. Even though sound-based HAR can identify emergencies, abnormal behavior, and speech-based health indicators, it has not been used as much as vision-based methods. Several applications of audio-based HAR in healthcare and smart home settings have been investigated recently. A deep learning-based sound recognition model for occupant behavior monitoring and emergency event detection in single-person houses was proposed by [12]. Along with emergency situations like shouting, puking, glass breaking, and explosions, their method categorizes routine activities like cleaning, showering, and listening to music. With an average precision of 78.0% and recall of 90.8%, the CNN and LSTM architecture-built model proved to be successful in differentiating between everyday sounds and emergencies.

In a different study, [19] used transfer learning to present a lightweight audio-based HAR architecture. Their method uses a fully connected classifier to handle high-level auditory embeddings that are extracted using YAMNet, a pre-trained con-

volutional neural network. Home tasks including preparing coffee, frying eggs, using the microwave, and cleaning dishes are categorized by the system. This approach allows real-time HAR with high classification accuracy while preserving the computing economy, making it ideal for portable and low-power devices.

Although these studies demonstrate the increasing importance of audio-based HAR in healthcare and smart home applications, audio-based HAR techniques alone have notable drawbacks. According to [12], overlapping sounds and background noise can lower recognition accuracy. Furthermore, the authors in [19] observed that background water sounds make it difficult for audio-only classification to classify activities that have similar auditory characteristics, including hand washing, dish-washing, and tooth brushing. Since audio data can inadvertently collect sensitive information, [5] underlined the importance of real-time, on-device processing in guaranteeing user security.

2.3 Sensor-based Approaches

A well-known method for classifying human activities is sensor-based HAR, which makes use of ambient sensors like motion detectors, temperature sensors, and smart home appliances as well as wearable sensors like heart rate monitors, gyroscopes, and accelerometers. The capacity of these techniques to function in privacy-sensitive settings without depending on video or audio data makes them especially helpful in the fields of healthcare, smart homes, and human-computer interaction.

In healthcare applications, wearable sensor-based HAR (WSHAR) has been thoroughly investigated, especially for fall detection and elderly monitoring. According to [18], multimodal wearable sensors are becoming increasingly important for tracking activities connected to health and identifying human actions. These systems frequently collect motion data using inertial measurement units (IMUs) found in smartphones, smartwatches, and other wearable technology. Machine learning and deep learning models are then used to process the data. Although wearable sensors offer high-frequency motion data, the positioning of the sensors greatly affects their

accuracy, and the recognition performance varies depending on whether the device is placed on the ankle, waist, or wrist. Furthermore, certain studies suggest that using numerous sensors on various body areas enhances recognition accuracy; however, this comes at the expense of deployment expenses and user discomfort.

Conversely, HAR based on ambient sensors uses environmental sensors to identify activity without the need for direct sensor attachment. In their investigation of HAR, the authors in [3] showed how well smart home sensors—such as motion detectors, temperature sensors, door sensors, and appliance usage trackers—can identify Activities of Daily Living (ADLs) in smart home environments. However, there are significant obstacles and restrictions with ambient sensor-based HAR. Ambient sensors rely on potentially ambiguous surrounding signals, in contrast to wearable sensors that directly record body movement. A motion sensor in the kitchen, for instance, may signal when someone is cooking, cleaning, or just passing by. These systems also have trouble recognizing temporal patterns because of sensor location changes across different house layouts and data sparsity from event-triggered recording. According to [7], sensor fusion—the combination of many wearable and ambient sensors—improves recognition accuracy while preserving user comfort and privacy.

2.4 Multimodal Approaches

To increase recognition accuracy and robustness, **Multimodal HAR** combines many sensor modalities, including wearable, ambient, video, and audio sensors. Multimodal HAR can overcome the drawbacks of unimodal techniques, including ambiguities in sensor-based HAR, background noise in audio-based HAR, and occlusions in video-based HAR, by utilizing complementing data sources.

The benefits of multimodal HAR in healthcare and smart home applications have been shown by recent studies. A multimodal monitoring system for ADL identification was presented by [13], combining ambient sensors, companion robots, and wearable sensors. Their technology achieves over 90% identification accuracy in smart home environments by fusing motion data, sound events, video, and position monitoring

using a Dynamic Bayesian Network (DBN). Energy-efficient sensor activation, which minimizes needless data collecting and prolongs battery life on wearable devices, is a significant advance in their work.

To illustrate how decision-level fusion may greatly enhance detection performance, [11] investigated audio-video fusion for fall detection. Their solution uses Dempster-Shafer (D-S) fusion to improve decision reliability and integrates MobileNetV2 for audio-based classification with YOLOv7-Pose for video-based skeletal tracking. Their tests demonstrated the complementing nature of multimodal data by demonstrating that combining audio and video improved fall detection sensitivity from 81.67% (video-only) to 97.50% (multimodal fusion).

Through the development of a deep learning-based fusion model, [10] significantly extended multimodal HAR for smart healthcare applications. Using self-attention mechanisms in order to improve important activity features while eliminating irrelevant information, their method combines Convolutional Neural Networks (CNNs) for video processing with ConvLSTM for time-series sensor data. Their technique outperformed conventional unimodal methods with an accuracy of 97.61% when tested on the UP-Fall dataset.

Multimodal HAR still faces a number of obstacles in spite of these advancements. Since video, audio, and sensor data operate at various sampling rates and necessitate sophisticated time-alignment techniques, synchronization between modalities is a major problem. Furthermore, multimodal data fusion raises computational complexity, which makes real-time execution difficult. Concerns about privacy also surface, especially when integrating audio and video data in residential settings.

2.5 Real-time/Edge-based approaches

Due to the requirement for real-time, low-latency, and privacy-preserving HAR systems in smart home settings, edge-based HAR has drawn more and more attention. On low-power, embedded devices like Raspberry Pi, ESP32, and STM32 microcontrollers, edge HAR allows a direct activity classification, lowering computational over-

head and data transmission costs in contrast to cloud-based systems.

[4] introduced a video-based deep learning framework for action detection on embedded platforms. Their approach achieves real-time performance on Visual Processing Units (VPUs) and Raspberry Pi-like devices by combining an LSTM for spatiotemporal activity recognition with a MobileNetV2-SSD-based people detection system. According to their findings, cutting computing costs while maintaining state-of-the-art accuracy is possible when deep learning architectures are optimized for edge deployment.

Another research has been done on ambient sensor-based HAR for embedded, real-time activity recognition in smart homes. Using continuous environmental sensor data, authors from [6] developed an ensemble learning framework for HAR that was effectively implemented on a Raspberry Pi. To increase identification accuracy while guaranteeing low power consumption and quick inference, their approach combines Temporal Convolutional Networks (TCN), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU). Using Raspberry Pi, ESP32, and STM32 microcontrollers, [16] further developed a real-time HAR algorithm tailored for embedded home automation systems. Their research emphasized the necessity of designing algorithms with hardware awareness by illustrating the trade-offs between processing time, classification accuracy, and power efficiency.

A computationally efficient real-time HAR approach, combining a novel Deep Skip-connection GRU (DS-GRU) network for temporal sequence modeling, LiteFlowNet for optical flow-based motion feature extraction, and a lightweight CNN for person detection was proposed by [27]. Utilizing pyramidal convolutional features and the MOSSE tracker, the system performs robust tracking and activity categorization with low latency and a simplified model. Using five benchmark datasets (HMDB51, UCF-101, UCF-50, Hollywood2, and YouTube Actions), the suggested approach outperforms a number of state-of-the-art models while keeping the model size small (48 MB) and allowing inference in less than 1 second for a 30-frame movie. The DS-GRU shows its appropriateness for embedded and edge-based surveillance applications by achieving comparable accuracy to deep LSTM networks with $1.4\times$

faster inference.

These methods have been further expanded by multimodal HAR on-edge devices, which integrate many sensory modalities for reliable, real-time recognition. [2] presented a multimodal HAR framework that processes RGB, depth, and inertial sensor data using an enhanced CNN-GRU classifier. Their approach uses a genetic algorithm for the best feature fusion, improving multimodal representations and removing duplicate data to improve recognition. The suggested CNN-GRU model facilitates real-time multimodal HAR on low-power embedded devices by effectively capturing spatial and temporal relationships, improving action categorization. The efficacy of multimodal fusion on edge platforms was proved by evaluations on the Berkeley-MHAD, NTU-RGB+D, and UTD-MHAD datasets, which showed state-of-the-art accuracy.

This thesis builds upon these works by developing a multimodal HAR system optimized for real-time utilization and integrating video, audio, and ambient sensor data to improve recognition robustness. To the best of my knowledge, there are currently no existing methods for handling all 3 activity recognition modalities (video, audio, sensors), which are also lightweight and can be used for real-time processing. By optimizing deep learning architectures for edge inference, this work addresses the key trade-offs in latency, accuracy, and power efficiency for real-time smart home HAR applications.

Table 2.1: Summary of Related Works in HAR

Literature	Architecture	Accuracy	Real-time	Lightweight	Modality		
					Video	Audio	Sensors
[24]	Lightweight CNN + MLSTM	Avg. F1: 98.53%		✓	✓		
[9]	HCAF+DMFF +SMA, MobileNetV3, ResBiLSTM	Fall detection: 100%, general HAR: 90%			✓		
[12]	CNN + LSTM	Precision: 78.0%, Recall: 90.8%				✓	
[19]	YAMNet + FC Classifier	>80%	✓	✓		✓	
[3]	Sensor Rule-based ML	Aruba: 96%, Milan: 90%, Cairo: 87%					✓
[13]	Dynamic Bayesian Network	>90%			✓	✓	✓
[11]	YOLOv7-Pose + MobileNetV2 + D-S Fusion	Avg, Acc: 98%			✓	✓	
[10]	CNN + ConvLSTM + Self-Attention	97.61%			✓		✓
[4]	LSTM + MobileNetV2-SSD	Avg. Acc: 99% (on several datasets)	✓	✓	✓		
[6]	Ensemble of TCN + LSTM + GRU	Avg. Acc: 96.52%		✓			✓
[27]	DS-GRU + LiteFlowNet + CNN	UCF: 95%, YouTube Actions: 97%, HMDB: 72%		✓	✓		
[2]	CNN-GRU + Genetic Fusion	Avg. Acc: 97% (on several datasets)		✓	✓		✓
Proposed Method	MobileNet +BiLSTM (video), BiLSTM (audio), LSTM+GRU +TCN (sensor)	Real-time, high accuracy	✓	✓	✓	✓	✓

Chapter 3

Methodology

3.1 Datasets

Experiments were conducted on 3 publicly available datasets - YouHome ADL dataset, HWU-USP, and UCF101 datasets. Two of the datasets focus on multimodal HAR, containing various data modalities ranging from environmental sensors to RGB videos, while the UCF101 is mainly for video-based tasks. The main approach will be based on the YouHome dataset, and a similar pipeline will be evaluated on the remaining datasets.

3.1.1 YouHome ADL dataset

The UIUC YouHome ADL Dataset [21], which consists of 2,289 instances of 31 regular daily tasks carried out by 20 individuals, is an invaluable resource for HAR applications. From simple domestic activities like cooking, using a computer, and watching TV to more intricate activity sequences, it records a wide range of behaviors. The dataset includes multiple camera viewpoints and data collected from video, audio, motion, temperature, humidity, and light intensity sensors, making it highly comprehensive for multimodal HAR research. Moreover, it includes both single-user and multi-user activities, as well as simple and compound events, which involve sequential or simultaneous actions.

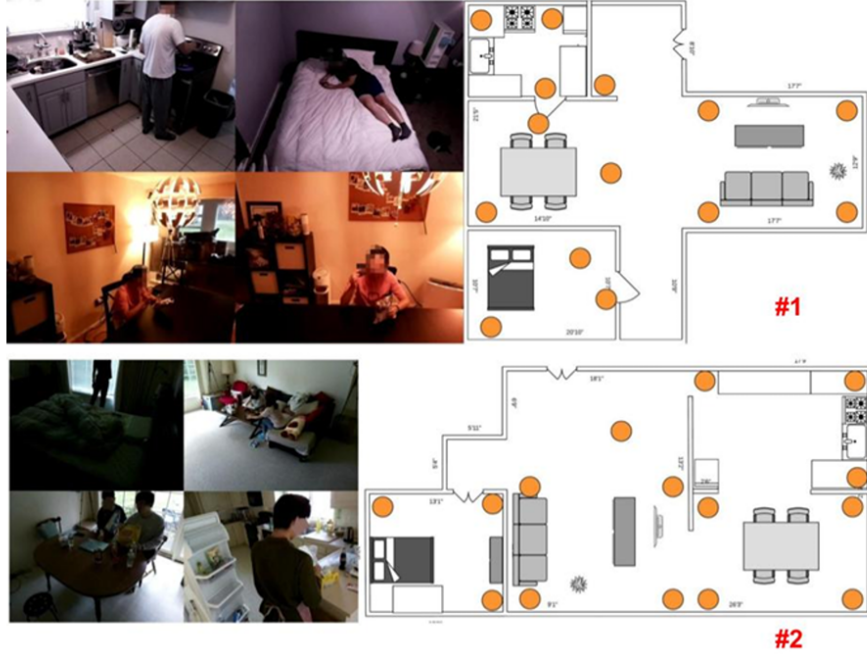


Figure 3-1: Sample images and home layouts taken from [14] (orange circle marks edge node positions)

Figure 3-1 shows the dataset’s diverse and realistic environments, including changes in occlusions, lighting, and human poses. It is, therefore, more challenging and indicative of real-world circumstances than those gathered in controlled settings, improving the robustness of HAR models. The dataset’s capacity for generalization is further enhanced by the fact that activities are documented in several locations, avoiding a straight mapping between an activity and a particular setting. This is illustrated in Figure 3-1 as the orange circle marks - edge nodes, which recorded all of the data modalities in different locations.

The incorporation of radar sensor data is a distinctive feature that improves privacy while still efficiently recording motion patterns. Nonetheless, there are certain biases in the dataset, especially with regard to how the samples are distributed among the activities. As can be seen in Table 3.1, there are much fewer examples of common behaviors like "entering" and "leaving" (6 samples each) than of acts like "drinking from a bottle" and "drinking from a cup" with more than 150 samples each.

This gap might cause models to struggle with underrepresented activities while

Table 3.1: Activity distribution in dataset from [14]

Activity		Activity		Activity		Activity	
Cook Cut	28	Enter	6	Use coffee machine	14	Use shelf	42
Cook use microwave	28	Exercise	56	Use computer	151	Use sink	28
Cook use oven	28	Getup	42	Use dish washer	20	Use switch	152
Cook use stove	28	Lay on bed	42	Use game controller	91	Use tablet	152
Drink from bottle	152	Leave	6	Use kettle	28	Use vacuum	42
Drink from cup	154	Nap	56	Use mop	42	Watch TV	56
Eat snack	148	Play board game	119	Use phone	153	Write	153
Eat utensil	106	Read	152			Total	2289

excelling at commonly occurring ones. It also has limited multi-person activity data, lacking audio information for these scenarios, restricting models that depend on auditory cues for classification. Despite these challenges, The YouHome dataset is an extensive and challenging benchmark for creating reliable HAR models, particularly for multimodal and real-world smart home applications.

3.1.2 HWU-USP dataset

Heriot-Watt University/University of Sao Paulo (HWU-USP) activities dataset [23] is a complex dataset, gathered at Heriot-Watt University’s Robotic Assisted Living Testbed (RALT). It captures a broad variety of everyday activities—from simple ones, like reading, to more complicated ones with long-term dependencies, like preparing a sandwich. This dataset simultaneously combines video, IMU, and ambient sensor data, as displayed in Figure 3-2. RGB-D videos were captured from a fixed angle with the help of a TIAGo robot, and participants wore smartwatches and waist-mounted IMU devices. As for the ambient sensors, they included binary switches on cupboard doors, a fridge, a drawer, a PIR motion sensor, and a power meter for a kettle.

Over the course of two weeks, sixteen healthy adult volunteers participated in the data acquisition, each of whom completed nine daily tasks. Anonymity was guaranteed for participants, as the faces in the output videos were blurred, and no audio was recorded. The captured activities included reading a newspaper, using a phone or laptop, preparing tea or a sandwich, setting the table, cleaning the kitchen, and others. The average duration of each activity was 48 seconds, which is longer than

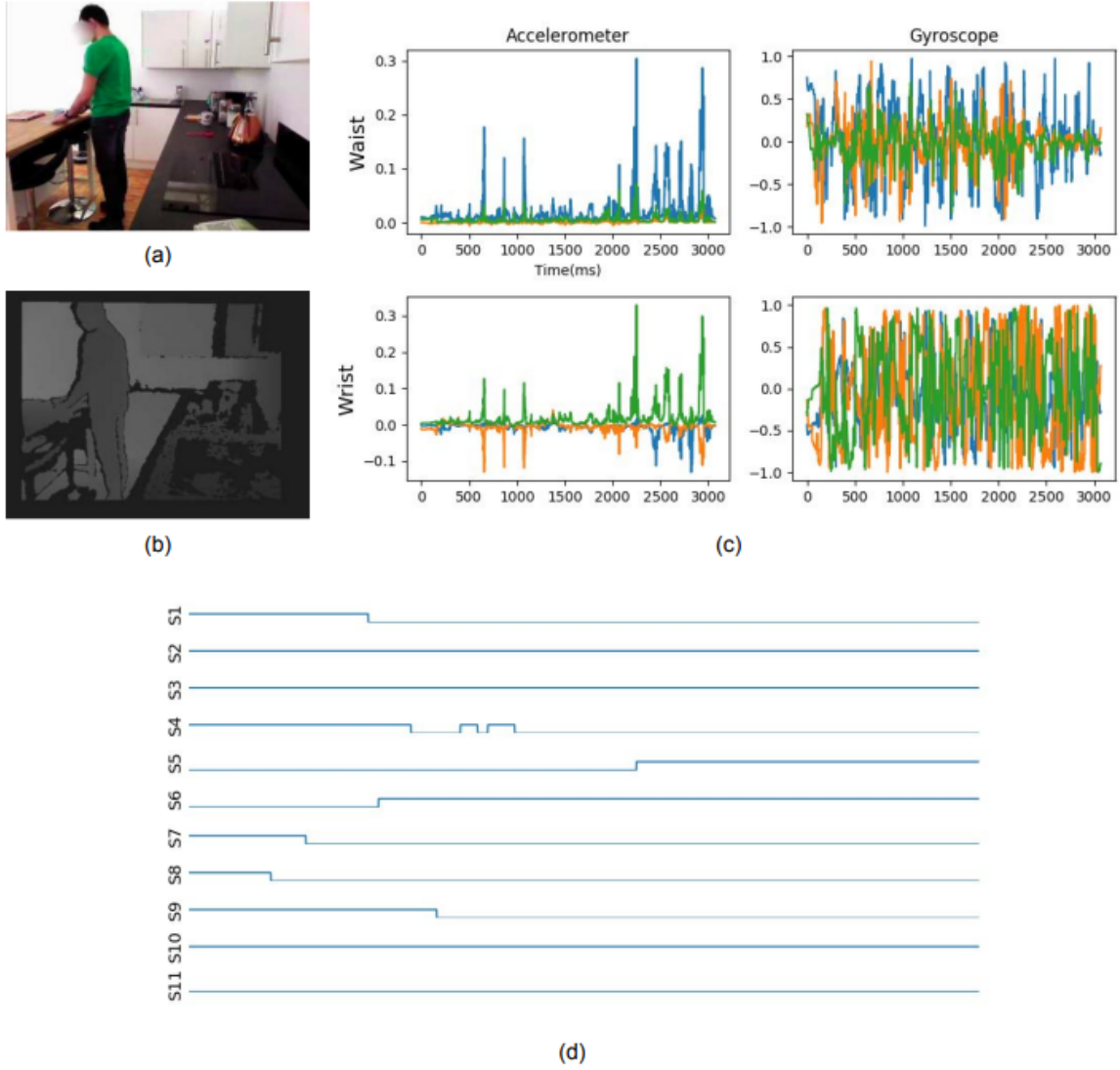


Figure 3-2: Sample data recordings from HWU-USP dataset taken from [23], including RGB (a) and depth images (b), IMU (c), and binary ambient sensor data (d)

typical activity datasets, and the recordings lasted roughly 20 minutes per person.

There are 144 activity recordings in the dataset, with a combined duration of approximately 116 minutes. The activities were created to mimic real-world tasks that require using a variety of kitchen tools and appliances. Making a sandwich, for instance, required several steps, including gathering ingredients from various places and putting them together.

Although the samples in this dataset are very limited (only 16 samples per activity), there exist many recent works that utilized this dataset for their HAR systems along with other well-known large-scale datasets like UTD-MHAD and UP-Fall detection. Moreover, the matching modalities with YouHome dataset makes HWU-USP a great dataset for evaluation of the proposed architecture.

3.1.3 UCF-101

UCF-101 is an extensive action recognition dataset that was gathered from YouTube and includes 13,320 realistic videos in 101 action categories. It greatly expands the quantity and variety of actions as an extension of the prior UCF-50 dataset. UCF-101 is one of the most challenging benchmarks in the field because, in contrast to many staged datasets, it poses real-world difficulties such as changes in camera motion, object appearance, viewpoint, background clutter, and lighting [25]. Each action in the dataset is organised into 25 groups, each containing 4–7 videos that frequently share contextual similarities like background or point of view. All actions can be divided into the following categories: Sports, Body-Motion Only, Playing Musical Instruments, Human-Object Interaction, and Human-Human Interaction. Figure 3-3 illustrates the sample recording inside the dataset.

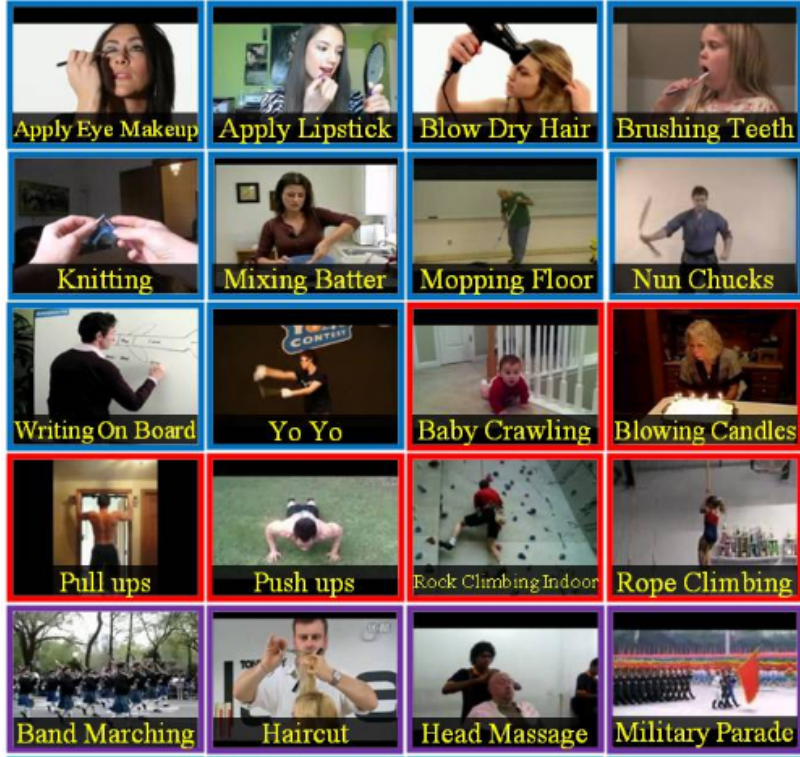


Figure 3-3: Sample videos from UCF-101 dataset taken from [25]

3.2 Preprocessing Data

3.2.1 Video processing

The preprocessing step was mostly taken from [21]. In their approach, the authors considered both privacy protection and computational efficiency while maintaining high classification accuracy.

First, frames from the videos were obtained considering frame skipping, in order to reduce the computational complexity. Next, Sobel filtering was used on video frames to preserve recognition accuracy while protecting user privacy. When raw images are transformed into contour representations, sensitive details are obscured yet important structural elements needed for activity recognition are preserved. Then, the model accuracy and real-time processing restrictions were balanced by resizing video frames to 192x192 pixels in order to maximize computing efficiency. This further improves privacy protection because [21] discovered that lower-resolution frames might still

provide enough information for classification tasks while obscuring identifiable characteristics of a person. In the case of this thesis, the frames were further resized to 224x224 pixels to match the shape of a feature-extracting model - MobileNetV2.

Next, the authors detected and cropped human figures from frames using a YOLOv5s object detection algorithm, which reduced background interference and focused on the relevant regions. Even though the YOLOv5s is comparably lightweight and has a great detection rate, I thought about changing it to a more recent version - YOLOv8s, but then decided to change it to a recent and lighter version - YOLOv8n. In [21] they retrained the YOLO model with the YouHome dataset and achieved the highest detection accuracy compared with the original YOLO model and also the YOLO model that was retrained on the COCO Human dataset. Similarly, the retraining of both YOLOv8s and YOLOv8n were done in this thesis and their performances were compared.

The additional preprocessing step includes computing Optical Flow from the extracted Sobel-filtered frames in order to capture temporal dependencies between each frame. According to [23], optical flow is the dense motion information that is calculated from successive video frames and captures the magnitude and direction of movement inside the scene. The optical flow is used to simulate temporal dynamics in the context of HAR, concentrating on how pixels change over time as a result of the motion of the body or an object.

This study utilizes the Farneback optical flow method [15], which is a dense motion estimating technique that uses a second-order polynomial to represent each pixel neighborhood in an image. It determines the motion's direction and speed for each pixel by monitoring the way these polynomials change across successive frames. By combining the original technique with Sobel filtering, we can distinguish real moving objects from noise more effectively, increasing the detection's robustness in challenging environments. The merged and cropped frames are then fed into a lightweight CNN for feature extraction.

3.2.2 Audio processing

Mel-spectrograms, which preserve both temporal and spectral properties and offer a time-frequency representation of sound, were extracted from the audio data. This method involves converting raw audio signals into log-scaled mel spectrograms, which effectively capture rich acoustic information over short time intervals [11]. Then the extracted mel-spectrograms are fed to CNN similar to VGG taken from [22], for feature extraction.

3.2.3 Ambient sensor data processing

Sensor readings from temperature, humidity, motion, and light sensors were normalized to ensure uniform feature scaling. A low-pass filter was applied to smooth variations and remove sensor noise, improving signal quality. Additionally, to handle missing sensor values, linear interpolation was applied, ensuring continuous data representation and avoiding gaps in model input sequences. Then, temporal statistical features such as mean, standard deviation, and variance were extracted to represent variations in environmental conditions using a dynamic sliding window approach. It helps us to analyze trends over time by grouping readings into overlapping windows.

3.2.4 Inertial sensor data processing

It was decided to also consider inertial data along with the ambient sensor data for greater robustness on HWU-USP dataset. Raw tri-axial accelerometer and gyroscope measurements are processed through a low-pass filter in the inertial-sensor branch to eliminate high-frequency noise and sensor drift. After noise reduction, the Gammatone Cepstral Coefficients (GCC) were extracted from inertal data. GCC is similar to Mel-frequency Cepstral Coefficients (MFCCs) but in GCC the triangular filter and log operations are replaced with the Gammatone filter and cubic operations [2]. These GCC-derived features are then input into a lightweight CNN+GRU classifier, where the GRU units describe the temporal development of the inertial signal and the convolutional layers learn spatial-structural patterns.

3.2.5 Dataset balancing

As was mentioned in section 3.1.1, the YouHome ADL dataset is highly imbalanced. In order to deal with it Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance all classes. In particular, [8] mentioned the following key concepts of the SMOTE:

i) SMOTE exclusively chooses k -nearest neighbors from the same class for every instance in the minority class X .

ii) SMOTE creates a new synthetic sample x_{new} after selecting one sample x at random from the k -nearest neighbors.

SMOTE was applied to each data modality, since different modalities have different features, thus, separate synthetic samples need to be generated for each modality separately.

Using SMOTE considerably increased recognition accuracy compared to the imbalanced baseline, however, to further enhance performance, it was decided to merge similar instances of activities in the dataset before balancing with SMOTE. For instance, merging activities like "drinking from a cup" and "drinking from a bottle" into one activity "drinking" would reduce redundancy while preserving essential distinctions. After merging several classes the remaining imbalance in the dataset was eliminated by applying SMOTE. These improvements ensured robustness across various activity circumstances and helped to create a more generalizable HAR model.

3.3 Proposed Method

3.3.1 Architecture and processing

The proposed HAR system follows a multi-stage pipeline:

1. Video processing includes frame resizing, privacy-preserving filtering, optical flow mapping, human recognition, and feature extraction.
2. Normalizing environmental sensor values, handling noises, extracting mel-spectrograms from audio data, and then extracting features from each modality.

- Several fusion algorithms were used to integrate individual modality predictions in a late fusion strategy.

Figure 3-4 shows the overall structure of the proposed framework. It can be seen that each modality is trained separately, and their output predictions will be fused using decision-level fusion algorithms.

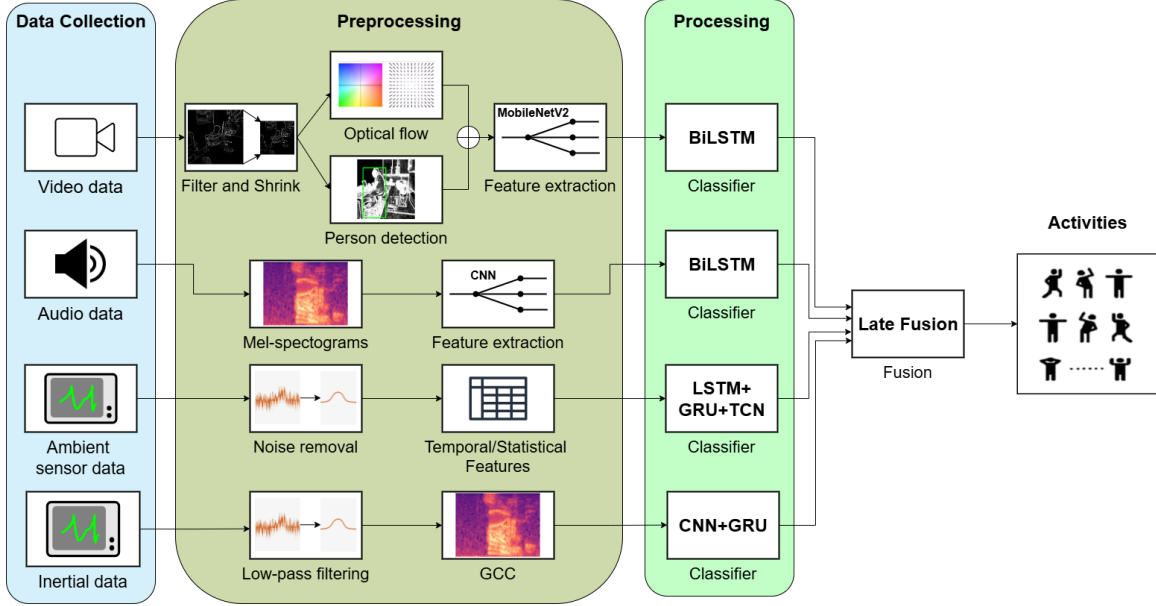


Figure 3-4: The diagram of the proposed framework

In order to achieve the best possible balance between accuracy and efficiency, lightweight deep learning models were used. Utilizing MobileNetV2’s effectiveness in edge computing settings, it was used for feature extraction of the videos. Then the extracted features from the videos will be passed to Bidirectional LSTM, which captures temporal information from both directions (past and future). Similarly, mel-spectrograms from audio data were analyzed using BiLSTM, while time-series data from environmental sensors was handled using a combination of LSTM, GRU, and TCN-based model inspired by [6]. The GCC data from inertial sensors was processed using CNN+GRU based model, similar to [2].

As for the hyperparameters, categorical cross-entropy loss was used to train the models, while grid search and cross-validation were used to improve the hyperparameter tuning. To optimize performance, key hyperparameters were systematically

changed, including learning rate, batch size, number of layers, and dropout rate. The best learning rate was 0.001, dropout range was 0.2-0.5, and batch size for different modalities was different, ranging from 32 to 128.

In order to avoid overfitting, the models were trained for 50-100 epochs using an early stopping mechanism. Accuracy, F1-score, and confusion matrices were used to measure performance and determine classification reliability. Furthermore, the architecture was validated on other datasets (HWU-USP and UCF-101) to verify the feasibility and ensure the robustness of the framework.

3.3.2 Fusion algorithms

Average fusion, weighted fusion, and majority voting were the three late fusion strategies that were used in this approach to aggregate the predictions from separate modalities.

In **Average fusion**, each modality’s predicted class probabilities are uniformly averaged, meaning that every modality has the same weight. In formal terms, the fused prediction is calculated as follows, where p_m represents the predicted probability distribution from modality m :

$$\hat{p} = \frac{1}{M} \sum_{m=1}^M p_m,$$

where M is the number of modalities. The final class label is chosen as $\arg \max(\hat{p})$.

Conversely, each modality in **Weighted fusion** is given a different weight, depending on past performance or heuristic knowledge. In the current approach, the vision modality has a higher weight (0.5) and the audio and sensor modalities lower weights (0.25 each), according to their individual accuracies:

$$\hat{p} = \sum_{m=1}^M w_m \cdot p_m,$$

where w_m is the weight for modality m , and $\sum w_m = 1$. This method emphasizes modalities that provide more discriminative power.

The last considered decision-level fusion approach is **Majority voting (Max-**

Vote). The class label that is most commonly predicted across modalities is chosen using this fusion technique. Every modality votes for the class it predicts, which is determined by the $\arg \max$ of its probabilities. The label that receives the most votes is the final prediction result of the fusion:

$$\hat{y} = \text{mode} \left(\{\arg \max(p_m)\}_{m=1}^M \right).$$

The class with the smallest index is chosen in the case of a tie.

Chapter 4

Results

4.1 Selecting YOLO model

The ability to identify human beings in video frames was assessed using two versions of the YOLO detection model to find a more suitable one. Both models were retrained on Sobel-filtered and resized images from the YouHome ADL dataset. The authors of the dataset also provided YOLO labels for their pictures, which made the fine-tuning process easier. As can be seen in Figures 4-1 and 4-2, the performance (in Frames Per Second) of the smaller and quicker version, YOLOv8n, has almost double the performance of YOLOv8s while having a small drop in detection accuracy. Therefore, YOLOv8n was chosen for the final implementation because it demonstrated reduced computing complexity, which makes it more appropriate for edge devices.

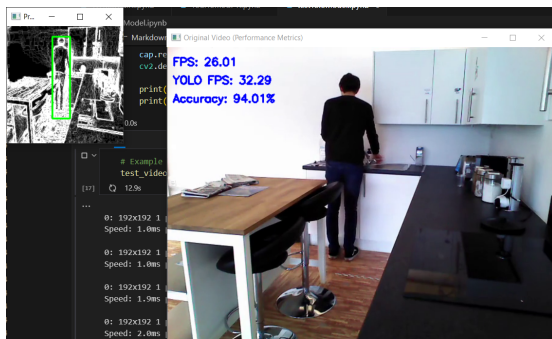


Figure 4-1: YOLOv8s performance

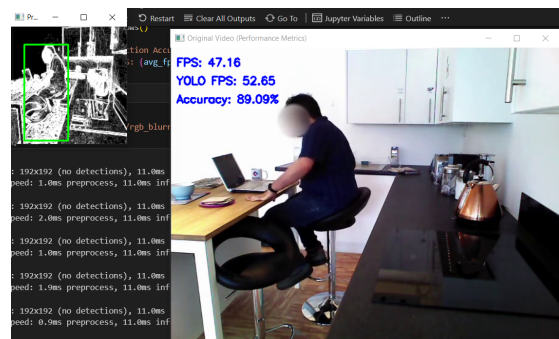


Figure 4-2: YOLOv8n performance

4.2 Baseline accuracies on YouHome dataset

To establish a baseline for the video-based model, an initial approach was to use a feature extractor and classifier from [21]. They used the ResNet18 backbone as a feature extraction model and they modified the classifier to identify the 31 activities. Utilizing the same pipeline as [21], this thesis managed to get 94% training accuracy, and 48% testing accuracy. Figure 4-3 illustrates a confusion matrix of the baseline performance.

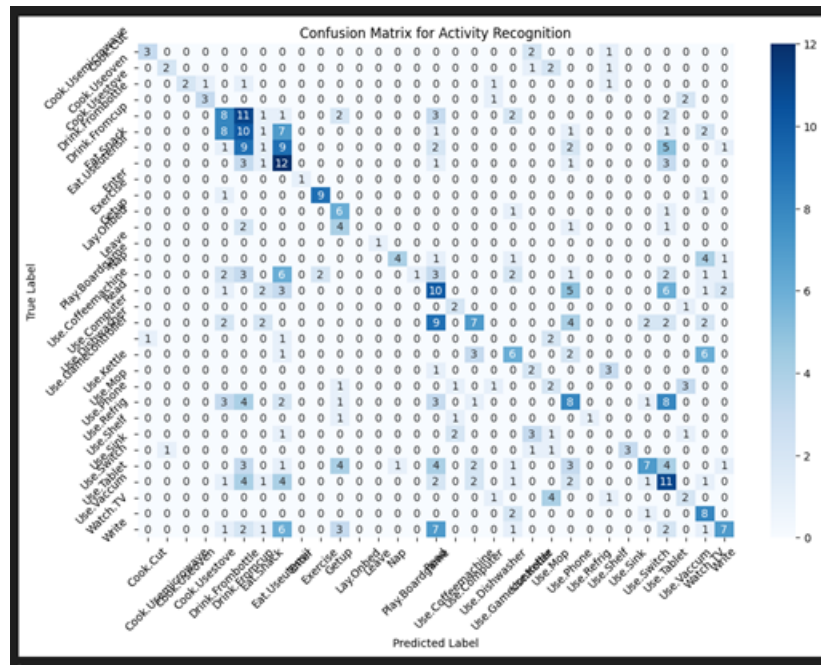


Figure 4-3: Baseline confusion matrix

Since [21] utilized only the video part, it was decided to use a Random Forest Classifier (RFC) for the extracted features from audio and sensor data. As was expected, the performance of the imbalanced YouHome ADL dataset came out low, resulting in around 20% and 40% of validation accuracy for audio and sensor modalities, respectively.

This served as a reference point for evaluating improvements introduced by the proposed deep learning models. Further models significantly outperformed the baseline, confirming the advantages of the end-to-end deep learning pipeline.

4.3 Improved Model Performances

4.3.1 Impact of Dataset Balancing

The high imbalance of the YouHome ADL dataset possibly affected the recognition accuracy due to the biases to the majority classes. Thus, balancing before processing it could lead to an increase in performance. After applying all the necessary pre-processing steps, it was time to apply SMOTE, mentioned in section 3.3.3. Since SMOTE works with extracted features, the Inflated 3D Networks (I3D) model was used as an initial feature extractor for the dataset videos. I3D learns spatiotemporal information directly from videos, which makes it a great tool in tasks like activity recognition that relies on both spatial and temporal data. Figure 4-4 illustrates class distribution in the dataset before and after applying SMOTE.

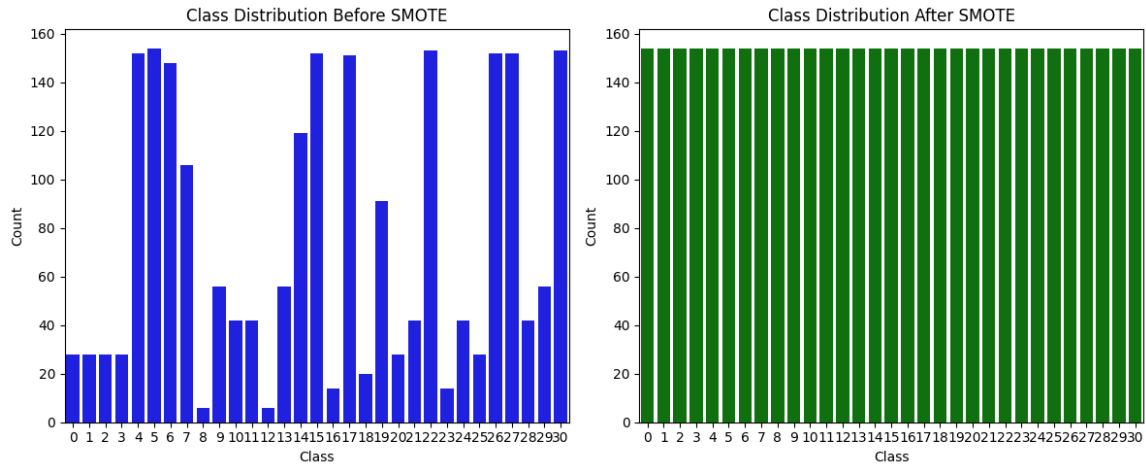


Figure 4-4: Class Distribution Before and After SMOTE

After balancing the dataset, the initial resnet18 model was changed to another classifier - the BiLSTM model, which also works perfectly with temporal data. Applying the hyperparameters from section 3.3.1, the training was started and the results showed an increase in validation accuracy up to 68%. The figure 4-5 below displays a confusion matrix after testing the balanced dataset.

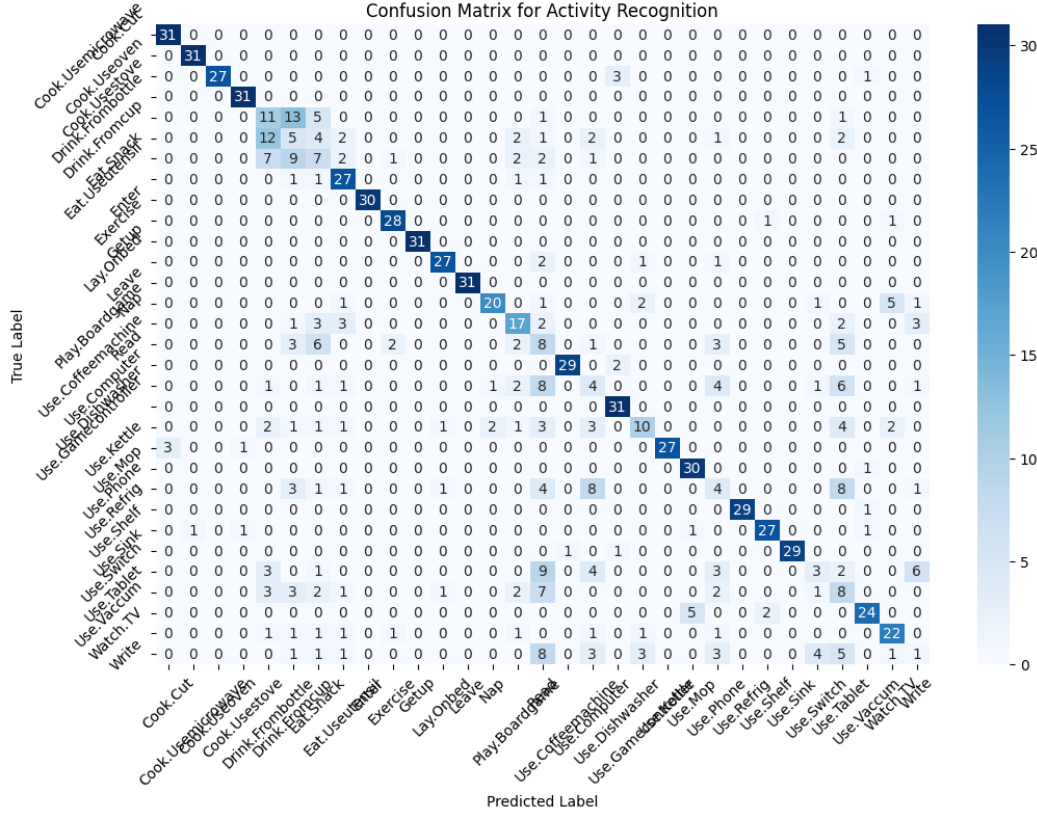


Figure 4-5: Confusion matrix after applying SMOTE

Since the main goal of the thesis is to deploy the real-time HAR system, it is necessary to consider the weights of the models. Even though I3D is perfect for extracting features from the videos, it would not be suitable for real-world deployment, due to its computational weight. Hence, it was decided to swap this feature extractor to the lighter one - MobileNetV2, which is suitable for image processing and computationally cheaper.

After switching to the pre-trained MobileNetV2 as the feature extraction model, the validation accuracy dropped significantly to 34%, compared to 68% achieved by the I3D model. The possible reason for that could be a MobileNetV2 being a 2D CNN and, therefore, does not capturing temporal dependencies across frames, unlike I3D which is a spatiotemporal model designed for video input. To avoid using heavy 3D CNN, an Optical Flow algorithm was introduced to encode motion information between consecutive frames. After that, the optical flow maps were combined with the

static frames, enabling MobileNetV2 to better approximate the temporal modeling capabilities of I3D. In the end, this resulted in a notable increase in recognition accuracy, nearly matching the more computationally demanding I3D model's performance (68% accuracy).

For classification of the sound features, similar BiLSTM model was used instead of RFC. After extracting mel-spectograms from the audio, their features were passed to SMOTE to balance the dataset. This resulted in increase in testing accuracy from 20% to 67%.

As for the sensor data, the classification model was changed to the combination of LSTM, GRU, and TCN layers. After feeding the extracted spatiotemporal features of the ambient sensor data to this model, the resulting accuracy was 72%, which is 32% greater than the initial sensor baseline.

4.3.2 Impact of Merging similar classes

To further enhance the performance of the models, merging similar instances of activities was executed. Classes like "Cook.Cut", "Cook.Useoven", and "Cook.Usestove" were combined into one class "Cooking". Separately, these activities contained 28 samples each, as can be seen in the table 3.1, but after merging, the new "Cooking" class contains 112 samples. While this approach increases the generalizability of the model and reduces misclassifications between semantically similar actions, there are activities that are still underrepresented. For instance, combining the lowest sample classes - "Enter" and "Leave", would result only in 12 samples combined. Since there are no such other similar activity, we need to either leave it as it is, creating large amount of synthetic samples (up to 300 with new combined classes), or drop both activities completely. To not overload the model with synthetic data, it was decided to drop these activities completely.

After merging similar classes, SMOTE was applied to the new set of classes to balance the obtained dataset. Figure 4-6 illustrates the activity distribution of merged classes before applying SMOTE, and after.

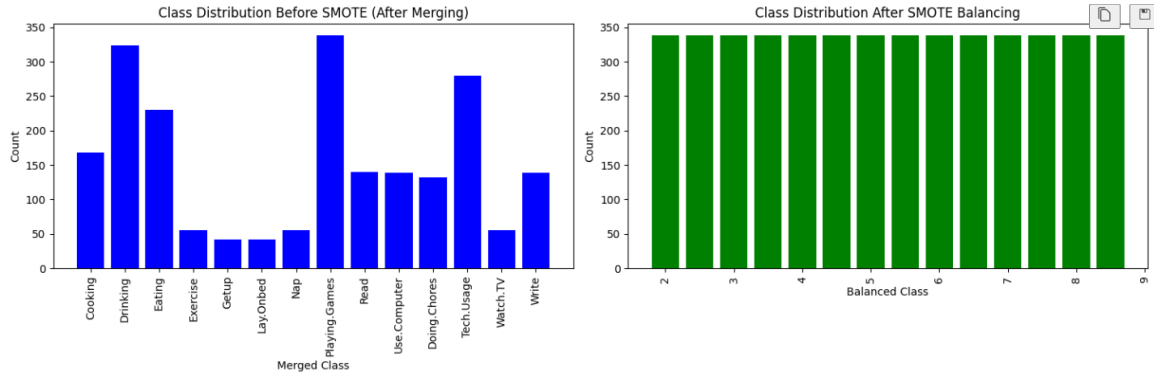


Figure 4-6: Merged classes distribution before and and After SMOTE

Next, training was done on this dataset, and the final recognition accuracy showed a significant improvement, comparing to the one before merged activities. For video modality, the new validation accuracy became 83% (increase in 16%), as was shown in the Figure 4-7. Audio model reached new accuracy of around 70%. Figure 4-8 displays loss and accuracy trends over the epochs for the audio classification. As for the sensors, the classes merging resulted did not improve on the classification accuracy, leaving it at around 70% (Figure 4-9).

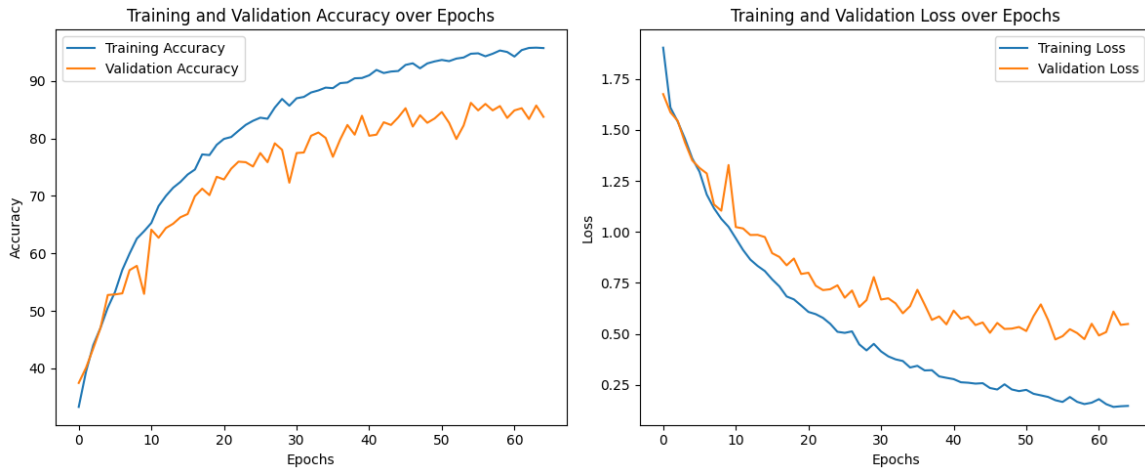


Figure 4-7: Accuracy and Loss graphs of the video model

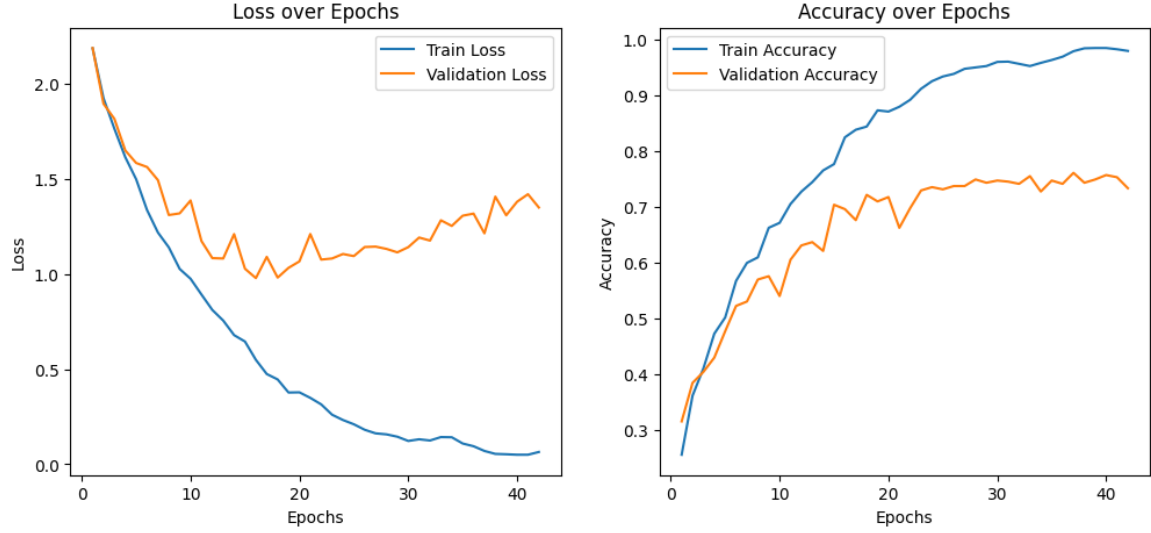


Figure 4-8: Accuracy and Loss graphs of the audio model

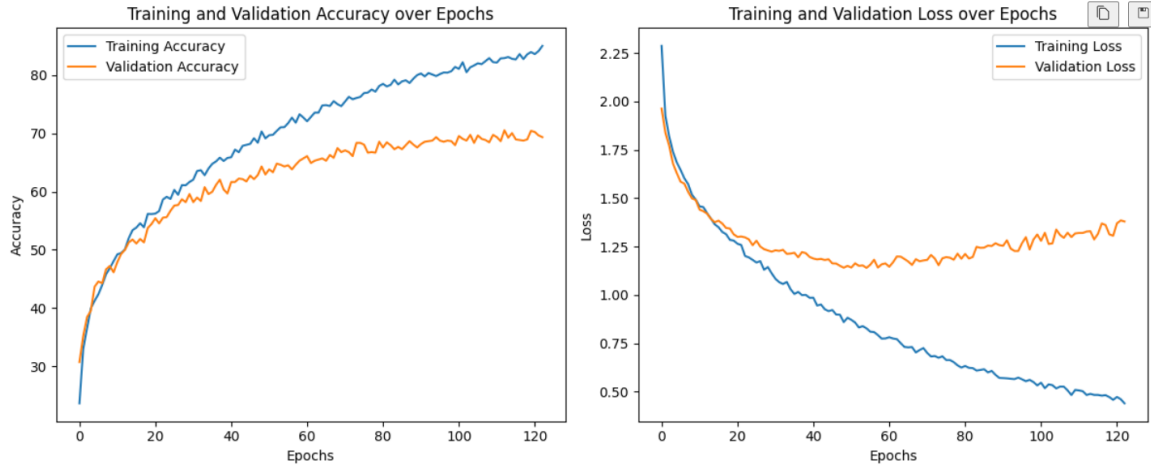


Figure 4-9: Accuracy and Loss graphs of the sensor model

4.3.3 Fusion results

As was stated in section 3.3.2, several decision-level fusion algorithms were used to evaluate the models' combined performance. For each fusion method, we calculated the **accuracy**, **F1-score**, **precision**, **recall**, and **inference time**. These metrics provide a comprehensive view of model performance and computational efficiency. Table 4.1 summarizes the results.

Table 4.1: Fusion results on YouHome ADL dataset

Fusion Method	Accuracy	F1-score	Precision	Recall
Average Fusion	95.6%	95.6%	95.8%	95.6%
Weighted Fusion	93.2%	93.1%	93.3%	93.2%
Max-Vote	88.8%	88.8%	90.4%	88.8%

4.4 Evaluation on other datasets

4.4.1 HWU-USP

An additional assessment was carried out on the HWU-USP dataset in order to further evaluate the generalization capabilities of the suggested multimodal HAR architecture. This dataset offers a solid standard for confirming the model’s cross-domain adaptability because it contains similar modalities (video, ambient sensors) as the YouHome ADL dataset, but different in terms of environmental setup and activity distribution. To ensure consistency, separate training, and testing were carried out with the same preprocessing and model configurations. This assessment aims to identify any potential limitations when applied to other data sources and ascertain whether the architecture functions dependably outside of the original dataset.

Due to the fact that the HWU-USP dataset contained video, IMU, and ambient sensor modalities missing the audio part, the fused evaluation was done only for available modalities that matched the main pipeline. After passing the features to separate models, the predicted scores were fused using decision-level fusion techniques described in section 3.3.2. Table 4.2 below displays the results obtained by individual models, as well as the fused output.

Table 4.2: Individual and Fused results on HWU-USP Dataset

Model/Method	Accuracy	F1-score	Precision	Recall
Vision model	50%	41.8%	37%	50%
Ambient model	58.3%	55.6%	54.6%	58.3%
Inertial model	44%	37%	33.3%	44.4%
Average Fusion	50%	41.8%	37%	50%
Weighted Fusion	80.6%	76.1%	75.2%	80.6%
Max-Vote Fusion	55.6%	50.2%	52.3%	55.6%

As can be noticed from the table, individual modalities did not perform well for dataset, however, the late fusion significantly increased the resulting accuracy. The highest accuracy that the models achieved was by using weighted fusion, estimating around 80% validation accuracy.

4.4.2 UCF-101

Similarly the challenging UCF-101 dataset was evaluated using the current framework. This dataset is primarily used for video-based HAR, but since the videos inside contained also sound, it was decided to try to apply the proposed pipeline for videos and audios inside the dataset. As can be seen in the table below, the architecture demonstrated that the vision branch alone achieved a solid 75% classification accuracy, while the audio branch only managed 50%.

Table 4.3: Individual and Fused results on UCF-101 Dataset

Model/Method	Accuracy	F1-score	Precision	Recall
Vision model	75.4%	75.4%	76.8%	75.5%
Audio model	52.9%	52.8%	55.2%	52.9%
Average Fusion	72.6%	72.4%	73.6%	72.6%
Weighted Fusion	75.9%	75.9%	77.4%	75.9%
Max-Vote Fusion	52%	50.3%	57.8%	52%

After using late-fusion techniques to integrate these complementary modalities, we can see a slight improvement in accuracy of a weighted fusion, which increased performance to 76%.

4.5 Comparative Evaluation with Prior Work

In this section, the comparison of the obtained results with the other research on the according datasets will be provided. The tables will list testing times if provided, accuracy, and estimated FLOPs and total number of parameters. The estimation was done with the help of AI tools.

Table 4.4: Performance comparison with prior work on YouHome ADL Dataset

Literature	Architecture	Testing Time	Accuracy	FLOPs (est)	Parameters (est)
Junhao et al. (2022)	YOLOv5s (edge), ResNet18 (cloud)	–	98.39% (event) 99.84% (user ID)	2-3 GFLOPs	20 M
Mingdian Liu (2023)	Pretrain + MLP	2.169 s (video) 2.640 s (vid+aud)	93.4% (video) 95.7% (vid+aud)	–	–
Proposed method	YOLOv8n + MobileNetV2 + BiLSTM (video), BiLSTM (audio), LSTM+GRU+TCN (sensors)	10 min (1.05s/sample)	95.6% (fused)	1 GFLOPs	5 M

Table 4.4 illustrates the performance results of different papers on the YouHome dataset. There is not much work done on this dataset, however, the existing methods reached high recognition accuracies, proving the viability of the dataset. By looking at the results, it is clear that the proposed approach reached the state-of-art accuracy

on the YouHome dataset utilizing all three dataset modalities (video, audio, and sensors).

As for the HWU-USP evaluation, the proposed framework did not reach the SOTA accuracy, but despite that, outputs high recognition results while having less computational costs. Table 4.5 provides an extensive comparison of the current approach with the prior works on the HWU-USP dataset.

Table 4.5: Performance comparison with prior work on HWU-USP Dataset

Literature	Architecture	Testing Time	Accuracy	FLOPs (est)	Parameters (est)
Abro et al. (2025)	CNN + Adaptive Neurofuzzy Inference System (ANFIS)	–	89%	4GFLOPs	25M
Alazeb et al. (2024)	LSTM-RNN	–	97.5%	–	–
Abro & Jalal (2024)	ANN	–	83.3%	–	–
Batool et al. (2024)	CNN-GRU	–	97.99%	0.8-1GFLOPs	–
Proposed method	YOLOv8n + MobileNet + BiLSTM (video), BiLSTM (audio), LSTM+GRU+ TCN (sensors)	1 min 37s (2.7s/sample)	80.5%	1GLOPs	5M

Across the UCF-101 benchmark, we can observe a clear trade-off between recognition accuracy and model complexity. The proposed approach obtained 76% accuracy, staying slightly behind from the SOTA results, but on the other hand, offering lower latency and resource demands.

Table 4.6: Performance comparison with prior work on UCF-101 Dataset

Literature	Architecture	Testing Time	Accuracy	FLOPs (est)	Parameters (est)
Hoang et al. (2023)	Shuffle-3DCNN	–	89.55%	15 GFLOPs	4.7M
Gupta et al. (2023)	3DCNLSTM (3D CNN + MLSTM + YOLOv6 + OpenPose)	–	97.65%	50-100 GFLOPs	30-50M
Murugan et al. (2024)	R(2+1)D CNN (ResNet-based)	–	82%	20-30 GLOPs	10-20M
Batool et al. (2024)	Bi-LSTM-GRU	–	97.62%	–	–
Proposed method	YOLOv8n + MobileNet + BiLSTM (video), BiLSTM (audio), LSTM+GRU+TCN (sensors)	9 mins 12s (0.21s/sample)	76%	1GLOPs	5M

4.6 Ablation study

4.6.1 Lighter RGB model

The ablation study in the Table 4.7 illustrates the results of the current architecture without most of the preprocessing steps from the initial video model, passing pure RGB frames to the MobileNetV2. Since most of the video-based HAR approaches examine RGB frames, it was necessary to do the research on the trade-offs between performance and privacy. The evaluation was also done in Leave-One-Subject-Out (LOSO) fashion, assessing the model on unseen person. As can be seen from the table, YouHome dataset had decreased accuracy, while other datasets had significant accuracy boost.

Table 4.7: Results of the proposed lighter RGB architecture

Dataset	Individual accuracies	Fused accuracies	Training time	Testing time
YouHome ADL	Vision: 40% Audio: 35% Amb.Sensor: 31%	Avg: 39% Weighted: 41.8% Max_vote: 33.6%	5min 45s	52s
HWU-USP	Vision: 89% Amb.Sensor: 55.6% Inert.Sensor: 44%	Avg: 89% Weighted: 89% Max_vote: 66.67%	6min 29.5s	28.2s
UCF-101	Vision: 86.6% Audio: 50%	Avg: 82.6% Weighted: 87% Max_vote: 61%	7min 56s	2min

4.6.2 Heavier RGB model

Another ablation study displayed in the Table 4.8 demonstrates the approach with computationally heavier model - ConvNext. Facebook AI Research introduced ConvNeXt, a pure convolutional network architecture that updates the traditional ResNet design to reduce the performance gap using vision transformers. Utilizing this model assess the trade-offs between performance and precision, showing the possibly of the architecture to achieve SOTA results, while having greater computational costs

Table 4.8: Results of the proposed heavier RGB architecture

Dataset	Individual accuracies	Fused accuracies	Training time	Testing time
YouHome ADL	Vision: 81.8% Audio: 35% Amb.Sensor: 31%	Avg: 81.8% Weighted: 84% Max_vote: 55%	6min 29s	1min
HWU-USP	Vision: 88.9% Amb.Sensor: 55.6% Inert.Sensor: 44%	Avg: 88.9% Weighted: 89% Max_vote: 66.67%	7min	1min
UCF-101	Vision: 93% Audio: 50%	Avg: 93.8% Weighted: 94% Max_vote: 75%	9min 40s	2min 10s

4.6.3 Performance comparisons with prior work

Table 4.9: Performance comparison with prior work on YouHome ADL Dataset (all methods)

Literature	Architecture	Testing Time	Accuracy	FLOPs (est)	Parameters (est)
Junhao et al. (2022)	YOLOv5s (edge), ResNet18 (cloud)	–	98.39% (event) 99.84% (user ID)	2-3 GFLOPs	20 M
Mingdian Liu (2023)	Pretrain + MLP	2.169 s (video) 2.640 s (vid+aud)	93.4% (video) 95.7% (vid+aud)	–	–
Proposed method	YOLOv8n + MobileNetV2 + BiLSTM (video), BiLSTM (audio), LSTM+GRU+TCN (sensors)	10 min (1.05s/sample)	95.6%	1 GFLOPs	5 M
Proposed method (Lighter RGB)	MobileNetV2 + BiLSTM (video), BiLSTM (audio), LSTM+GRU+TCN (sensors)	52 s (0.09s/sample)	41.8%	0.5 GFLOPs	3.667 M
Proposed method (Heavier RGB)	ConvNext + BiLSTM (video), BiLSTM (audio), LSTM+GRU+TCN (sensors)	1 min (0.1s/sample)	84%	5-6 GFLOPs	30 M

Table 4.10: Performance comparison with prior work on HWU-USP Dataset (all methods)

Literature	Architecture	Testing Time	Accuracy	FLOPs (est)	Parameters (est)
Abro et al. (2025)	CNN + Adaptive Neurofuzzy Inference System (ANFIS)	–	89%	4GFLOPs	25M
Alazeb et al. (2024)	LSTM-RNN	–	97.5%	–	–
Abro & Jalal (2024)	ANN	–	83.3%	–	–
Batool et al. (2024)	CNN-GRU	–	97.99%	0.8-1GFLOPs	–
Proposed method	YOLOv8n + MobileNet + BiLSTM (video), BiLSTM (audio), LSTM+GRU+TCN (sensors)	1 min 37s (2.7s/sample)	80.5%	1GLOPs	5M
Proposed method (Lighter RGB)	MobileNet + BiLSTM (video), BiLSTM (audio), LSTM+GRU+TCN (sensors)	28.2s (1.76s/sample)	88%	0.5 GFLOPs	3.667 M
Proposed method (Heavier RGB)	ConvNext + BiLSTM (video), BiLSTM (audio), LSTM+GRU+TCN (sensors)	1 min (3.75s/sample)	89%	5-6 GFLOPs	30 M

Table 4.11: Performance comparison with prior work on UCF-101 Dataset (all methods)

Literature	Architecture	Testing Time	Accuracy	FLOPs (est)	Parameters (est)
Hoang et al. (2023)	Shuffle-3DCNN	–	89.55%	15 GFLOPs	4.7M
Gupta et al. (2023)	3DCNLSTM (3D CNN + MLSTM + YOLOv6 + OpenPose)	–	97.65%	50-100 GFLOPs	30-50M
Murugan et al. (2024)	R(2+1)D CNN (ResNet-based)	–	82%	20-30 GLOPs	10-20M
Batool et al. (2024)	Bi-LSTM-GRU	–	97.62%	–	–
Proposed method	YOLOv8n + MobileNet + BiLSTM (video), BiLSTM (audio), LSTM+GRU+TCN (sensors)	9 mins 12s (0.21s/sample)	76%	1GLOPs	5M
Proposed method (Lighter RGB)	MobileNet + BiLSTM (video), BiLSTM (audio), LSTM+GRU+TCN (sensors)	2 mins (0.13s/sample)	87%	0.5 GFLOPs	3.667 M
Proposed method (Heavier RGB)	ConvNext + BiLSTM (video), BiLSTM (audio), LSTM+GRU+TCN (sensors)	2 min 10s (0.14s/sample)	94%	5-6 GFLOPs	30 M

4.6.4 Audio Privacy Protection

If we recall the Figure 3-4, the main architecture only considers privacy protection of the videos, whereas audio inputs could also contain sensitive information. Thus, to enforce audio privacy, spectrogram masking was applied, where each mel-spectrogram was processed with two random frequency masks (up to 15 mel bins wide) and two time masks (up to 30 frames long) before feature extraction. After retraining the current audio-CNN under identical hyperparameters to the unmasked baseline, the results showed a considerable decrease in recognition accuracy to around 10%. This ablation highlights the trade-off between stronger audio anonymization and performance in the main pipeline.

4.7 Edge Device Compatibility Evaluation

Since there were no real-time executions on edge devices to test the efficiency of the models properly, it was important to analyze the storage capacity and theoretical inference times on the edge platforms (ESP32, Raspberry Pi 4, etc.) to determine the real-world deployability of the proposed multimodal HAR pipelines.

Storage capacity. After calculating the model sizes of each proposed approach, the obtained results ranged from 30MB to 170MB, depending on the heaviness of the video model. Thus, utilizing *Raspberry Pi 4* (microSD up to 128GB) would pose no storage constraints, easily accommodating all models between 30MB and 170 MB. However, if we consider other devices like *ESP32* (4MB flash, 520kB SRAM) or *Arduino* (1MB flash and 256kB SRAM), the trained models would exceed the onboard storage without external memory.

Theoretical Inference Metrics. Looking at the performance tables above (Tables 4.9 - 4.11) we can derive per-inference workloads which range from 0.5 GFLOPs (3.667 M parameters) to 5–6 GFLOPs (30 M parameters). Assuming a Raspberry Pi 4 CPU

peak of roughly 20 GFLOPS, the smallest model would take

$$\frac{0.5}{20} \times 10^3 \approx 25 \text{ ms}$$

per inference, and the largest about

$$\frac{6}{20} \times 10^3 \approx 300 \text{ ms},$$

which lies within real-time bounds for Raspberry Pi 4. By contrast, an ESP32 (≈ 0.4 GFLOPS) would need

$$\frac{0.5}{0.4} \times 10^3 \approx 1250 \text{ ms} \quad \text{to} \quad \frac{6}{0.4} \times 10^3 \approx 15000 \text{ ms}$$

(1.25–15 s), clearly prohibitive for HAR.

Therefore, among the edge platforms, deployment on Raspberry Pi 4 would be ideal for feature-rich and efficient multimodal HAR.

Chapter 5

Discussion

5.1 Evaluation of Individual Modalities

In every modality, the baseline models that used conventional feature extraction and classification pipelines (Random Forest and ResNet18) did not perform well initially. However, after dataset balancing and model upgrades, the suggested end-to-end deep learning architecture demonstrated steady improvement.

Class imbalance inside the YouHome dataset was a major issue for recognition results due to the trained models being biased towards the majority classes. This problem was greatly lessened by class merging and SMOTE, improving generalizability of the models and lowering overfitting to specific behaviour definitions.

After dataset balancing, each modality displayed an increase in classification accuracy. The video stream increased its accuracy to 68%, compared to the initial baseline of 48%. Although the utilized I3D network yielded higher accuracy, its complexity made it unsuitable for real-time use. The integration of temporal modeling using MobileNetV2 feature extractor and Optical Flow mapping, instead of heavy 3D CNNs, decreased the computational costs compared to the baseline models (I3D, resnet18). MobileNetV2 + BiLSTM achieved equivalent performance while lowering the FLOPs by an order of magnitude, indicating a great trade-off between accuracy and speed.

The audio model also showed classification enhancements after balancing the

dataset. By using SMOTE, and moving from an ML model (RFC) to a DL model (BiLSTM), the audio classification accuracy increased from 20% to 67%. This outcome shows that deep temporal models are able to extract more significant information from mel-spectrograms than traditional ML classifiers.

As for the ambient sensor data, the hybrid model composed of LSTM, GRU, and TCN layers outperformed the Random Forest baseline by 32%, reaching an accuracy of 72%. This validates the use of temporal convolutional features in modeling structured ambient sensor patterns.

5.2 Fusion Evaluation

The predictions from the separate modalities were combined using decision-level fusion techniques. The techniques included weighted fusion, max-vote, and average fusion. The highest overall accuracies obtained with the proposed method for YouHome, HWU-USP, and UCF-101 datasets were 95.6%, 80.5%, and 76%, respectively. This outcome shows how important it is to combine complementary data from several modalities while preserving real-time responsiveness.

Additionally, even when separate modalities performed poorly, an appropriate late fusion algorithm enabled to output of strong recognition performance. On the HWU-USP dataset, for example, the video and sensor modalities both performed poorly on their own (achieving accuracies of 50% and 58.3%, respectively). However, fusion increased the overall accuracy to 80.5%, confirming the generalizability of the approach.

Furthermore, because fusion occurs at the decision level, the system remains robust even if one modality is unavailable. It will still function properly by relying on the remaining modalities' predictions without retraining or architectural changes.

5.3 Ablation Study Analysis

Two ablation studies were conducted in order to explore the trade-off between model complexity and generalization. Replacing the optical-flow-augmented video stream with raw RGB frames reduced YouHome fused accuracy down to 41.8% (average fusion), while boosting performance on HWU-USP and UCF-101 by 3–5% under LOSO evaluation. This could mean that it relied too much on the synthetically generated data, struggling with unseen person evaluation. Conversely, swapping MobileNetV2 for ConvNeXt pushed YouHome vision accuracy up to 81.8%, but increased FLOPs by $5\times$ and inference time by $2\times$. Additionally, while HWU-USP dataset did not had much improvement, UCF-101 considerably boosted its accuracy up to 93%. This validates the possibility of the architecture to reach SOTA performance, while having increased resource demands.

Another ablation study examined the effects of adding speech privacy protection. By masking the obtained mel-spectrograms from the sound, it is possible to filter out sensitive speech content from the audios. The obtained outcome was similar to comparing rgb frames and sobel-filtered frames, which resulted in considerable accuracy drop. In the case of the audio, it was around 10% decrease.

5.4 Limitations

Even though the suggested multimodal HAR framework has shown encouraging results and is comparably lightweight, there are still a number of drawbacks that need to be recognized and taken into account in further research.

Despite promising cross-domain results, all of the experiments were conducted on desktop hardware. It was therefore not possible to assess the real performance on physical edge hardware, particularly in terms of power consumption and latency under concurrent workloads. Without more testing under real-world circumstances, this gap limits the ability to declare with confidence that the proposed system is, in fact, efficient and ready for deployment. Furthermore, even if the calculated number

of FLOPs and parameters are still moderate, they may present increased latency or memory limitations when transferred to embedded systems. Quantization, real-time benchmarking on edge platforms, and hardware-specific optimization will all be part of future efforts.

5.5 Real-Time HAR Applications

Deploying HAR in real-time unlocks a variety of practical use cases including Smart Home Monitoring and Industrial Worker Safety.

- **Smart Home Monitoring:** By continuously monitoring sensor streams and segmenting events as they happen, real-time HAR in smart homes can help vulnerable people maintain their autonomy in their own homes, identify deviations or disruptions in routines that could indicate new health or safety issues, and send out immediate alerts in critical situations—like falls or prolonged inactivity—so that help can be provided quickly. [17] Moreover, in smart home environments, real-time HAR enhances quality of life and resource efficiency by providing more precise control over comfort and energy management (e.g., modifying lighting or HVAC in response to identified activities).
- **Industrial Worker Safety:** Real-time HAR can provide workers on the manufacturing line with immediate, data-driven protection and support. By continuously monitoring fine-grained motion signals through inertial sensors or video streams through cameras, it can quickly identify unsafe behaviours (such as improper lifting, falls, or encroachment on hazardous zones) and initiate automated interventions or alerts to prevent accidents [26]. In order to prevent musculoskeletal injuries, this continuous monitoring also makes it possible to conduct proactive ergonomic assessments, identifying patterns of repeated strain and guiding changes to workstation configurations or job rotations. Beyond safety, these systems provide real-time analytics to production planning, simplifying resource allocation, optimizing workflows, and directing focused training

to support skill development, all of which increase operational efficiency and worker well-being.

Chapter 6

Conclusion

With the use of the YouHome ADL dataset, this study demonstrated a strong multimodal human activity recognition (HAR) framework that incorporates video, audio, and sensor modalities to improve classification accuracy and real-time usability. A comprehensive HAR system tailored for real-time contexts was created using a methodology that included privacy-preserving preprocessing, efficient feature extraction, and late fusion techniques.

The outcomes showed the efficacy of deep learning models customized for various data modalities, lighting conditions, and viewpoint variations, with the final multimodal pipeline outperforming baseline models by a large margin. Furthermore, after dealing with dataset imbalance using class merging and SMOTE, the classification performance was greatly improved, also enhancing the generalizability of the models.

Overall, this study emphasizes how crucial fusion methods, model selection, and data preprocessing are for efficient multimodal HAR systems. The results confirm the possibility for real-time activity recognition on resource-constrained devices, such as Raspberry Pi 4, by utilizing lightweight deep learning models. Although this possibility was not tested in the scope of this thesis, future research could be done on the optimization and evaluation of this method on edge-based devices.

Bibliography

- [1] Ali K. AlShami, Ryan Rabinowitz, Khang Lam, Yousra Shleibik, Melkamu Mersha, Terrance Boulton, and Jugal Kalita. Smart-vision: Survey of modern action recognition techniques in vision. *arXiv preprint arXiv:2501.13066v1*, 2025.
- [2] MOUAZMA BATTOOL, MONEERAH ALOTAIBI, SULTAN REFA ALOTAIBI, DINA ABDULAZIZ ALHAMMADI, MUHAMMAD ASIF JAMAL, AHMAD JALAL, and BUMSHIK LEE. Multimodal human action recognition framework using an improved cnngru classifier. *IEEE Access*, 2024.
- [3] Damien Bouchabou, Sao Mai Nguyen, Christophe Lohr, Benoit LeDuc, and Ioannis Kanellos. Using language model to bootstrap human activity recognition ambient sensors based in smart homes. *electronics*, 2021.
- [4] Antonio Carlos Cob-Parro, Cristina Losada-Gutiérrez, Marta Marrón-Romera, Alfredo Gardel-Vicente, and Ignacio Bravo-Muñoz. A new framework for deep learning video based human action recognition on the edge. *Elsevier*, 2024.
- [5] STEFANIA CRISTINA, RODRIGO PÉREZ-RODRÍGUEZ, VLADIMIR DESPOTOVIC, and ANDSLAVISA ALEKSIC. Audio- and video-based human activity recognition systems in healthcare. *IEEE Access*, 2024.
- [6] Hatice Vildan DUDUKCU, Murat TASKIRAN, Zehra Gulru CAM TASKIRAN, and Nihan KAHRAMAN. Human activity recognition with ensemble learning. In *2024 26th International Conference on Digital Signal Processing and its Applications (DSPA)*, 2024.
- [7] Bruna Maria Vittoria Guerra, Emanuele Torti, Elisa Marenzi, Micaela Schmid, Stefano Ramat, Francesco Leporati, and Giovanni Danese. Ambient assisted living for frail people through human activity recognition: state-of-the-art, challenges and future directions. *frontiers*, 2023.
- [8] Rebeen Ali Hamad, Wai Lok Woo, Bo Wei, and Longzhi Yang. Self-supervised learning for activity recognition based on datasets with imbalanced classes. *IEEE TRANSACTIONS ONE MERGING TOPICS IN COMPUTATIONAL INTELLIGENCE*, 2024.
- [9] Altaf Hussain, Samee Ullah Khan, Imad Rida, Noman Khan, and Sung Wook Baik. Human centric attention with deep multiscale feature fusion framework for activity recognition in internet of medical things. *Elsevier*, 2023.

- [10] Md. Milon Islam, Sheikh Nooruddin, and Fakhri Karray. Multimodal human activity recognition for smart healthcare applications. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2022.
- [11] Shiqin Jiao, Guoqi Li, Guiyang Zhang, Jiahao Zhou, and Jihong Li. Multi-modal fall detection for solitary individuals based on audio-video decision fusion processing. *Heliyon*, 2024.
- [12] Jinwoo Kim, Kyungjun Min, Minhyuk Jung, and Seokho Chi. Occupant behavior monitoring and emergency event detection in single-person households using deep learning-based sound recognition. *Building and Environment*, 2020.
- [13] Fei Liang, Zhidong Su, and Weihua Sheng. Multimodal monitoring of activities of daily living for elderly care. *IEEE SENSORS*, 2024.
- [14] Mingdian Liu. Fine-tuning foundation models for activity of daily living analysis. Master’s thesis, Iowa State University, 2023.
- [15] Zhihao Ma, Tianyu Wang, Sheng Xu, Xinxing Mu, Qiang Wang, and Qingqiang Guo. Moving object detection based on farneback optical flow. In *Proceedings of the 42nd Chinese Control Conference*, 2023.
- [16] Houda Najeh, Christophe Lohr, and Benoit Leduc. Real-time human activity recognition on embedded equipment: A comparative study. *Applied sciences*, 2024.
- [17] Houda Najeh, Christophe Lohr, and Benoit Leduc. Real-time human activity recognition on embedded equipment: A comparative study. *Applied Sciences*, 2024.
- [18] Jianyuan Ni, Hao Tang, Syed Tousiful Haque, Yan Yan, and Anne H.H. Ngu. A survey on multimodal wearable sensor-based human action recognition. *LATEX CLASS FILES*, 2024.
- [19] Marco Nicolini, Federico Simonetta, and Stavros Ntalampiras. Lightweight audio-based human activity classification using transfer learning. In *ICPRAM2023- 12th International Conference on Pattern Recognition Applications and Methods*. World Academy of Science, Engineering and Technology, 2023.
- [20] Sheikh Nooruddina, Md. Milon Islama, Fakhri Karraya, and Ghulam Muhammad. A multi-resolution fusion approach for human activity recognition from video data in tiny edge devices. *Elsevier*, 2023.
- [21] Junhao Pan, Zehua Yuan, Xiaofan Zhang, and Deming Chen. Youhome system and dataset: Making your home know you better. In *2022 IEEE International Symposium on Smart Electronic Systems (iSES)*. IEEE, 2022.

- [22] Thomas Pfitzinger and Hendrik Wöhrle. Embedded real-time human activity recognition on an esp32-s3 microcontroller using ambient audio data. In *Proceedings of the 12th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*. IEEE, 2023.
- [23] Caetano Mazzoni Ranieri, Scott MacLeod, Mauro Dragone, Patricia Amancio Vargas, and Roseli Aparecida Francelin Romero. Activity recognition for ambient assisted living with videos, inertial units and ambient sensors. *sensors*, 2021.
- [24] Syed HammadHussain Shah, Anniken Susanne T. Karlsen, Mads Solberg, and Ibrahim A. Hameed. Anefficient and lightweight multiperson activity recognition framework for robot-assisted healthcare applications. *Elsevier*, 2023.
- [25] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. Technical Report CRCV-TR-12-01, Center for Research in Computer Vision, University of Central Florida, 2012.
- [26] Sungho Suh, Vitor Fortes Rey, and Paul Lukowicz. *Wearable Sensor-Based Human Activity Recognition for Worker Safety in Manufacturing Line*, pages 303–317. Springer Nature Switzerland, 2024.
- [27] Amin Ullah, Khan Muhammad, Weiping Ding, Vasile Palade, Ijaz Ul Haq, and Sung Wook Baik. Efficient activity recognition using lightweight cnn and ds-gru network for surveillance applications. *Applied Soft Computing*, 2021.