Declaration of Authorship – Heriot Watt University

Course code and name:	**B31VZ**
Type of assessment:	**Individual**
Coursework Title:	**MSc. Project**
Student Name:	**Muhammad Basit**
Student ID Number:	**H00397610**

Declaration of authorship. By signing this form:

- **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment is entirely my own. I have NOT taken the ideas, writings, or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings, or inventions of others, or any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgment section.

- I confirm that I have read, understood, and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.

- I confirm that I have read, understood, and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature *(type your name):* **Muhammad Basit**

Date: **05-22-2022**

Heriot-Watt University

School of Engineering and Physical Sciences

MSc Program Title: MSc Robotics

Project Title: Human Activity Recognition

Course Code: B31VZ

Author: Muhammad Basit

Student full registered name : Muhammad Abdul Basit

Student HWU Matriculation Number: H00397610

Supervisor's Name: Prof. Dragone Mauro

Date: 15-05-2023

Word Count: 12459 words

Dissertation Submitted in partial fulfilments of the requirements for the degree of MSc at Heriot-Watt University

Abstract

Human Activity Recognition (HAR) using Deep Learning techniques has become popular because of its wide range of applications in healthcare, sports, security, and entertainment. Deep Learning techniques like CNN and RNN have been employed in this domain by several researchers. However, there are challenges currently faced due to heterogeneity in the data. The sensor data is not labelled and manual labelling is quite expensive and time consuming. The hardware devices may showcase inconsistencies and may be unreliable.

In this research, a thorough analysis of various public datasets is done that include RGB and RGB-D video data as well as axial data involving accelerometer and gyroscope readings. A dataset is then curated which includes five human activities that involve a variety of bodily movements. These include chopping vegetables, washing dishes, mixing ingredients in a bowl, pouring drink into a glass, and opening a refrigerator. The axial data is used to generate features including acceleration, angles, and angular velocities. As manual timestamping is cumbersome and time consuming, a custom BNO055 watch is programmed to timestamp and label the sensor data. The custom watch with BNO055 sensor provided better results than an Apple Watch 6 and a Lilygo T watch with a BMA423 sensor. The combinational approach of CNN and LSTM was found to be most effective for human activity recognition particularly for activities like chopping, stirring, and pouring. It also provided substantial results particularly for RGB video and IMU sensor modalities.

Keywords: Human Activity Recognition (HAR), CNN (Convolutional Neural Network), LSTM (Long Short Term Memory), RNN (Recurrent Neural Network), IMU (Inertial Measurement Unit).

Acknowledgement

I would like to express my gratitude towards my supervisor, Prof. Dragone Mauro, who extended his support throughout the course of the project, especially during my illness. I would also like to extend my gratitude to Ronnie Smith and Scott MacLeod, who have assisted me whenever I encountered any challenges. Furthermore, I would like to assure that the project is entirely my own work and I have gained ethical approval via the Chair of the Student's Ethics Committee.

Muhammad Abdul Basit

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

CHAPTER 1

INTRODUCTION

Human Activity Recognition (HAR) can be referred to as the task of recognizing and identifying activities with the help of Artificial Intelligence. Earlier, images were used to classify human activities[1]. Nowadays, wearable devices are often used to collect various sensory information to capture movements. The availability of low-cost sensors and live streaming of data makes it possible to improve the efficiency of HAR systems. As such, the use of wearables in HAR has become an active research area. The main objective of this dissertation is to perform Human Activity Recognition using vision-based data as well as sensory modalities. This includes RGB and RGB-D based video data as well as sensor data collected from accelerometer and gyroscope to capture bodily movements of the user and provide more precision.

The choice of this topic is crucial as the use cases of HAR are wide ranged. HAR can be used to develop assistive technology for the elderly or people who are differently abled[2]. The physical movement of patients with chronic conditions can be analysed to improve treatment. The health status can also be monitored which can help provide healthcare benefits. HAR can be used for security reasons to identify potential threats and surveillance in public spaces[3]. Athletes can make use of HAR to analyse their movements and improve training[4]. Gaming is another area which can make use of HAR by tracking player movements. This data can be further used to control the movement of the characters the players are playing.

Robotic technology has enabled the use of domestic robots that can provide assistance to the elderly or differently abled[5]. These advancements have proven to be quite successful and can be incorporated in HAR Systems. Such domestic/service robots can then be used to make human-like decisions without explicitly being asked to even in a mobile fashion. These robots can be used to detect falls of the elderly, provide any assistance if a user is struggling with a task, monitor activities for health analysis, and even operate around the house or an office.

The advancement of technology has also made available small, wearable devices like smartwatches, fitness trackers, armbands, etc. These introduce portability to smart devices. Since they are often embedded with sensors like accelerometer, gyroscope and magnetometers, the wearables can provide more insights to human movements.

Machine Learning introduces powerful tools to extract meaningful patters from the data. Earlier, HAR was commonly performed with the help of traditional Machine Learning algorithms including Support Vector Machines (SVMs), Decision Trees, and Random Forest[6]. The advent of Deep Learning has opened new doors for HAR that creates the ability to analyse complex relationships within the dataset. Image-based and video-based HAR tasks typically involve Convolutional Neural Networks (CNN).

While, a number of researches have produced state-of-the-art results in the field of HAR, there is still a need for further exploration because of the challenges that need to be overcome for better accuracy of the HAR systems. The sensor data retrieved from accelerometers and gyroscopes is quite heterogenous and requires pre-processing which is computationally very expensive. The data is often labelled manually which makes the process very time consuming and costly[7]. Therefore, an unsupervised approach without the need of any human effort is more viable and is an area to be looked into in this dissertation. Hardware devices can often be unreliable, data could be missing or captured inconsistently. Lastly, the axial data retrieved from the accelerometer and gyroscope needs to be balanced and there is a need to ensure that data from the IMU (Inertial Measurement Unit) is transmitted seamlessly.

The proposed model for carrying out HAR in this dissertation is based on a combination of CNN and RNN. CNN provides the ability to extract features from RGB and RGB-D data. RNNs provide the ability to recognize sequential characteristics in the data which is more relevant to the axial data. LSTM is a type of RNN used in this dissertation that allows feedback connections to process whole sequences of data.

The dissertation is organized into 6 chapters. Chapter 1 provides an overview of HAR using wearables and its importance. It lays a roadmap and discusses the main motivation of the dissertation and how significant the study is. Chapter 2 discusses a detailed literature survey of HAR. The literature survey comprises of a review of past work done in the field of HAR

based on the various datasets, the variety of devices to collect the data, the pre-processing and annotation, and lastly the machine learning techniques to carry out HAR including the advancement of Deep Learning in the field. Chapter 3 describes the methodology on how the HAR system was built. It discusses the datasets, the pre-processing techniques applied, feature extraction, and the Deep Learning Models used to carry out HAR. It explains the hybrid CNN-LSTM model and the proposed architecture. Chapter 4 provides a detailed analysis on the findings of the evaluation. This chapter presents the results of Chapter 3 and details how efficient the hybrid CNN-LSTM is. Chapter 5 provides a discussion on the findings and summarises the key findings. It highlights the advantages and disadvantages of built system and compares it to the results obtained by other researchers. Chapter 6 draws a conclusion based on the results. It also poses recommendations for future research based on the drawbacks of the built system.

To summarise, this dissertation aims to provide a seamless way to detect human activity using wearables. This research is based on RGB and RGB-D data, and axial data (accelerometer and gyroscope data). The proposed model is a hybrid model of CNN and LSTM. Though, previous work in this area has proved momentous, there are still challenges that need to be addressed. The inconsistency in the sensor data needs to be reduced. Manual labelling which makes the task time consuming needs to be done away with. The low reliability on hardware devices needs to be addressed. This research proposes a Deep Learning Model and also curates a dataset to be able to address these challenges.

CHAPTER 2

Literature Review

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This section comprises of the related work done in the field of Human Activity Recognition. The main goal of the literature survey was to survey the popular datasets and the performance on these datasets, the various devices used for collection of data, the types of data including videos and sensor data, and the machine learning algorithms and models to find out the best possible combinations so as to make the task efficient. Additionally, labelling and annotation in these systems was also a matter of concern. The traditional machine learning algorithms as well as the advanced Deep Learning Technologies were surveyed to make the best trade-off between accuracy, efficiency and comfort.

2.2 Related Work

Several datasets with a wide range of data have been curated to carry out human activity recognition. Some of these only considered visual data while others are more comprehensive and include a wide range of sensor data. UCF-50 is one such popular dataset that includes videos all taken from YouTube[8]. The concept of taking videos directly from YouTube instead of recording them was to ensure authenticity in the training data. UCF-50 is an extension of a similar action dataset called UCF-11 which includes 11 action categories. UCF-50 dataset has 50 action categories including basketball shooting, walking with a dog, horse riding, playing violin, military parade, drumming etc. that are further grouped into 25 categories. The similar categories include over four RGB videos with similar features like same person or similar background. Despite being so comprehensive, the dataset has a variety of issues due to abrupt camera motion leading to blurry or distorted images, size inconsistencies, background clutter due to presence of distracting elements, light inconsistencies, and different viewpoints as same activity may look different from various positions which may be difficult to capture. All these factors negatively impact the performance of Machine Learning algorithms on the UCF-50 dataset.

Z. Yang et al., developed a RGBD dataset by extracting the interest points from the RGB channel which are then combined with depth map-based descriptors[9]. The authors showcase a much higher accuracy of 89% in the proposed approach as compared to traditional RGB data and mention the importance of depth information for being straightforward yet more robust and reliable than RGB data. While RGB only captures the intensity value of every pixel in an image, the depth sensor captures the distance between the camera and each pixel. This introduces a three-dimensional spatial element in the data. When a human faces a camera and occlusions do not occur, RGB is a viable option. However, if the camera is mounted at some height, occlusions can occur which can be mitigated using depth cues suggesting the use of RGBD videos over RGB videos.

Numerous such RGB and RGBD video-based datasets were collected through smartphones. Wan et al., recommended a smartphone inertial accelerometer-based solution and explored Deep Learning techniques on two datasets[10]. The authors claimed that vision-based Human Activity Recognition has certain complications arising due to privacy concerns as well inaccuracies since bodily movements are not captured as accurately using RGB or RGBD data. Therefore, they suggested the use of sensor-based data using wearables or smartphones. They conducted the experiment with the participants carrying a smartphone in their pockets. The authors compared the performance of the new approach to traditional solutions that did not take into consideration any sensor data and showcased state-of-the-art results. The authors implemented CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory), Bi-LSTM (Bidirectional LSTM), MLP (Multilayer Perceptron), and SVM (Support Vector Machine), and achieved best results with CNN with an accuracy of 92.8%.

Another such smartphone based dataset is UCI-HAR which involved a total of 30 subjects and six activities including walking, walking upstairs, walking downstairs, sitting, standing, and laying[11]. All the activities were recorded using a smartphone worn by the subject on the waist. The smartphone included an embedded accelerometer and gyroscope to provide additional modalities that proved advantageous over RGB or RGBD-based data. All the data in the UCI-HAR dataset was labelled manually.

Ramanujam et al., reviewed smartphone and wearable sensor data for Human Activity Recognition and mentioned integrated sensors including accelerometer, gyroscope, magnetometer, temperature, ambient sensor, object sensor, and additional orientation sensors in wearables that prove advantageous over smartphones[12]. The authors argued that wearables are designed to be worn on the body which ensure a closer proximity to the wearer's body and precise data. On the other hand, smartphones are often kept in the pocket and are less able to capture precise data based on human movements. Wearables also generally have more sensors than smartphones and are able to capture more comprehensive data than smartphones. Wearables are also designed in such a way that they are more power efficient than smartphones and require less frequent charging. A seamless and continuous monitoring is easier to capture with wearables than smartphones. A variety of wrist orientations and movements also cannot be captured by smartphone sensors, thus, the authors concluded that wearables are more suited to provide a range of modalities for Human Activity Recognition.

Watch-n-Patch is another popular Human Activity Recognition dataset which includes daily activities of seven persons for a total length of 458 videos of 230 minutes. The videos are located in six different kitchens and eight different offices to introduce complex background variations and to introduce generalizability. The data includes RGB as well as IMU data collected with Microsoft Kinect One sensor to capture additional motion and oriental information[13]. The authors focused on unsupervised learning to teach the model the relation between activities and actions and to make the task less intensive by reducing manual work. The authors refer to the short clips as action-words and a collection of action-topics as activities. Then, a probabilistic model is proposed that relates these action-words to action-topics. The authors also made a comparison of RGB videos and RGB-D videos and concluded that RGB-D videos outperformed the RGB data since the depth cues in RGB-D data provided more accurate information.

UTD Multimodal Human Action Dataset (UTD-MHAD) is another comprehensive dataset compiled with the help of Kinect Camera and a wearable sensor used due to their low costs and easy operability[14]. The Kinect Camera was produced by Microsoft and released in 2010. It captured images with a resolution of 640x480 pixels with a frame rate of 30 frames

per second. The dataset included 27 actions including waving, clapping, drawing, pushing, walking, squatting, etc. which were performed by 8 subjects, four male and four female. Every action was repeated by the subjects four times. In all, the dataset comprised of 861 data sequences and has served for substantial studies and related work in Human Activity Recognition. All the starting and ending depth frames were annotated manually with visual inspection. The dataset consists of four synchronized modalities including RGB videos, depth videos, inertial signals and skeleton positioning data. The authors recommended the use of fusion approach including the depth videos as well as inertial sensors over traditional vision data for better performance. The manual annotation is a disadvantage in this approach.

Considering the advantages of wearables over other devices with respect to additional comprehensive modalities, it was more appropriate to make use of wearables for HAR while capturing the RGB or RGBD data with a combination of sensor information, though RGBD data is more accurate than RGB data. Wearables like smartwatches were also a better option than smartphones due to their ability of capturing wrist movements. As mentioned earlier, smartwatches are not carried in pockets but are worn on the body and allow a greater proximity to the user allowing capturing of intricate wrist movements.

Among the popular smart watches used, Apple watch has been used in various researches conducted for Human Activity Recognition. Ashry et al.,[15] used an Apple Watch Series One and recorded activities of sixteen subjects with information on orientation, angular velocity and acceleration provided by gyroscope, magnetometer and accelerometer. The authors argued that smart home sensors are fixed to their locations and would not be able to detect actions when a human leaves the area as the sensors would then not be able to record any information based on the user's movement. Hence, they suggest the use of wearable sensors which can be worn by users on their wrists, waist, legs or chest. The authors make use of an Apple watch and used an LSTM-based approach on various datasets to reach a highest accuracy of 94.5%. They also suggest working on a Myo device consisting of data retrieved from the IMU sensors as well as electromyographic (EMG) sensors and magnetometer instead of the Apple watch in the future. The Myo armband is worn on the forearm and allows its sensors to record electrical activity going around in the

muscles. This electrical activity is able to detect the gestures. The Myoarmband consists of nine IMUs containing three axis gyroscope, three axis accelerometer, and three axis magnetometers.

In another research conducted to detect Human Activity Recognition for office workers syndrome, Mekruksavanich et al.,[16] used an Apple Watch Series 2 with the WatchOS 4 version 4.2 Operating System to collect data of 10 participants. Apple Watch Series 2 has an accelerometer, gyroscope, and ambient light sensor. The authors captured the accelerometer and gyroscope data. An iOS app called SensorLog was used that was linked to the watch with a sample rate of 30 Hz. The research was conducted to detect Office Workers Syndrome, a condition wherein workers spend very long durations in the office in a specific position wherein their posture or positioning of the muscles might cause pain or physical discomfort to them in the long run. The results showcased an overall better performance of various classifiers on the combinational data of accelerometer and gyroscope as compared to individual data from either of the two sensors. The combinational data showcased highest accuracy of 93% suggesting the usage of both modalities for activity recognition.

The comprehensive Opportunity Dataset was curated to benchmark HAR algorithms[17]. The data was recorded using body-worn sensors, object sensors and ambient sensors. Four subjects were used to record the data and every action was recorded six times. The body-worn sensor recorded using seven inertial measurements units and twelve acceleration sensors. Object sensors measured acceleration and rate of turn. Ambient sensors included thirteen switches and acceleration sensors. Xia et al., made use of Opportunity Dataset and evaluated an LSTM-CNN based architecture based on the F1-score[18]. An F1-score of 92.63% was achieved on the Opportunity Dataset.

For pre-processing the data, most researchers used linear interpolation, scaling, normalization, and segmentation. Since, some data could possibly be lost during collection, the absent data is indicated as NaN. Linear Interpolation is used to overcome this problem by constructing new datapoints using linear polynomials. Data bias can be introduced due to large values in channels. As such, the data is normalized between 0 and 1 to remove bias

in the training data. Authors in [18] made use of segmentation to preserve the temporal relationship between data points.

Before Deep Learning, traditional machine learning methods were used on sensor data including Support Vector Machine (SVM) and Naïve Bayes to extract features. Jain et al., made use of SVM and K-Nearest Neighbour (KNN) classifiers to recognize human activities[19]. SVM helped achieve better results than KNN with an accuracy of over 95%. However, the research included basic actions including walking, sitting, standing, laying, etc. Besides, the traditional Machine Learning methods involved manual feature extraction which was cumbersome and limited to human knowledge[20]. Principal Component Analysis (PCA) was used to extract features in an unsupervised manner but the complex dependencies between the features were often lost. PCA is used to reduce dimensions and transform large number of variables into smaller numbers. However, PCA also involves losing information that can be critical. The disadvantages of common approaches encouraged researchers to explore other alternatives that did not involve manual feature extraction or Principal Component Analysis.

Attal et al., performed HAR using wearable inertial sensor data captured from key body points – chest, right thigh, and left ankle[21]. The authors claim that wearable sensors are generally placed on the waist, sternum and lower back since these represent those motions that are closer to a human's center of mass and can result in better measurements. However, wearing these sensors can make the user uncomfortable, that is why the authors focused on minimum number of sensors and a placement that minimizes discomfort. The classification techniques involved K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Gaussian Mixture Models (GMM), and Random Forest (RF) for supervised learning. The authors also employed unsupervised techniques including K-means, Gaussian Mixture Models (GMM), and Hidden Markov Model (HMM). Among the supervised algorithms, KNN showcased the best performance followed by Random Forest, and Support Vector Machine. Among the unsupervised algorithms, Hidden Markov Model (HMM) showcased the best performance with an accuracy of 83% followed by Gaussian Mixture Model (75% accuracy), and finally K-mean (72% accuracy). The unsupervised algorithms did not showcase great results. The choice of algorithm had a severe impact on the overall

performance. This research also evaluated the impact of the number and the placement of sensors and noted the trade-off between the accurate measurements and the discomfort faced by the user. The sensor placement was a drawback in the study.

The advent of Deep Learning allowed researchers to eliminate the need to manually extract features from the sensor data. Agarwal et al., proposed a Deep Learning Model using RNN which achieved an accuracy of 95.78%[22]. RNN is a kind of neural network that makes use of sequential data to predict next likely scenarios. Though the research involved new Deep Learning methods, the dataset included only a few activities and used a tri-axial accelerometer, ignoring multi-sensor data. This could hinder its ability to generalize in other scenarios and thus, was not very effective.

Hou compared traditional Machine Learning algorithms like Support Vector Machine (SVM), Random Forests to Deep Learning methods like Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)[23]. The author conducted the experiment on two datasets; USC-HAD and WISDM (Wireless Sensor Dating Mining) dataset. The USC-HAD dataset only includes a few activities recorded of 7 male and 7 female individuals. On the other hand, the WISDM dataset includes six labelled activities including walking, jogging, walking upstairs, walking downstairs, sitting and standing. The dataset also includes accelerometer readings in three axes. The results showed an accuracy of 87% with traditional machine learning approaches including Support Vector Machine, K-Nearest Neighbor, and Random Forest, and an accuracy of 90% with Deep Learning techniques like CNN and LSTM. The author's experiments on the dataset resulted in the conclusion that traditional machine learning algorithms are more appropriate for small scale data but in case of large datasets, deep learning techniques CNN and LSTM provide more sensible choices.

Zheng el at., proposed a Convolutional Neural Network (CNN) based solution with three convolutional layers and one fully connected layer with the highest accuracy of 93%[24]. CNN is an artificial neural network commonly used for visual images. The results demonstrate the importance of feature learning in time series classification.

Lawal et al., proposed a two-stream CNN and argued about the importance of axial data (accelerometer and gyroscope data) to develop a robust classifier for Human Activity

Recognition[25]. The authors additionally also experimented with datasets of various body locations to identify the best position for placing the sensors. One of these is the Real World Human Activity Recognition (RWHAR) dataset which included 9 activities recorded by 7 wearable devices and 6 sensors. 15 participants (8 male and 7 female) carried out the tasks and every activity was recorded for a duration of 10 minutes. The activities included jumping, walking, climbing, running, sitting, etc. Accelerometer, gyroscope, GPS, magnetometer, audio and light sensors were present in the devices. Jumping was not performed for long durations since it is exhaustive. The motion signals were recorded from seven body parts including waist, shin, chest, upper arm, forearm, head, and thigh. The axial data was sampled at a rate of 50 Hz. A total of 885,360 activity images were generated for five activities. The authors concluded that the shin and waist are the best positions to place the sensors for accurate data as they performed better across all the activities whereas thigh is the worst position to place the sensor among the seven mentioned body parts. In fact, shin position is even better than waist position to place sensors and retrieve readings. They also emphasized on the better performance of the two-stream CNN over traditional CNN approach.

The research conducted by Hammerla et al., indicated that RNN performed better than CNN for short activities across three different datasets consisting of wearable sensor information[26]. The authors also studied hyperparameter tuning in this context and summarized the impacts.

Ordonez et al. combined CNN and RNN layers and manage to reduce the training time by 17%[27]. The framework uses sensor modalities individually as well as in a combination and demonstrate an improved performance on a combination of accelerometer and gyroscope data. The authors compared the performance based on the Opportunity Dataset mentioned earlier as well as the Skoda dataset. The Skoda dataset includes activities of workers in a car production plant and include 10 gestures in the recordings which are 3 hour long[28]. Since, Skoda dataset was more balanced than Opportunity Dataset, the former was able to showcase better results on the combined CNN-RNN architecture.

Xia et al., proposed a novel approach of a combination of CNN and LSTM. The model was able to extract the features automatically and be able to classify the various activities with minimal parameters. The model was trained on three datasets including the Opportunity Dataset mentioned before, and achieved an accuracy of 95.8%.

In a research conducted by Li et al., another hybrid approach of CNN and LSTM was showcased on Opportunity (mentioned prior) and UniMiB-SHAR datasets[29]. The UniMiB-SHAR dataset comprises of accelerometer readings taken from a Samsung Galaxy Nexus I9250 which is equipped with a Bosh BMA220 acceleration sensor. The data is sampled at a fairly high frequency of 50 Hz and consists of 17 actions. These are divided into context interactions (sleeping inside or outside a car), motion related (bending, running, standing, walking, etc.), posture related (sitting, laying, sitting on a chair etc.) and sport-related (bicycling, jumping, jogging). Other kinds of ADLS related to cooking or housekeeping were not included as only low order activities were needed. Eight of these seventeen activities are ADLs (Activities of Daily Living) and seven of these are falling actions. Every activity was performed 2 to 6 times. Half the time the subjects were placing the smartphones in their right pocket, and half the time the smartphone was placed in the left pocket to reduce any bias. Unlike Opportunity Dataset, the UniMiB-SHAR dataset does not consist of any NULL values. The dataset is more balanced however three classes namely walking, running, and going down are heavily sampled as compared to other classes. The F1-score showcased exceptionally bad results in the Opportunity Dataset when features were selected manually. These included 18 hand crafted features. Automatically learned features showcased a much better performance on both the datasets as compared to hand crafted features. Hyperparameter tuning of the number of channels and frame size did not yield any significant improvement.

Ashry et al., demonstrated the use of Bi-LSTM classifier to perform Human Activity Recognition[30]. The authors curated a dataset called CHAR-SW to record activities in a continuous stream. The data was collected using smartwatches in which 25 participants (14 females and 11 males) were made to wear an Apple Watch Series 4. The Apple Watch Series 4 provides an improved accelerometer, an improved gyroscope, ambient light sensor, optical heat sensor, and electric heat sensor. The CHAR-SW dataset consisted of the

acceleration readings, angular velocity, rotational displacement and gravity measurements. These were recorded at a fairly high sampling rate of 50 Hz. In addition to this, the performance of the Bi-directional LSTM was also compared on two public datasets. The results showcased that cascading Bi-LSTM had a fairly good performance.

Challa et al., also showcased the use of CNN and Bi-LSTM network[31]. Three benchmark datasets including WISDM, UCI-HAR, and PAMAP2 were used. The model proposed is a multi-branch CNN-BiLSTM model to capture local features as well as long term dependencies because of LSTM. The proposed model also analyses various filter sizes to best capture the local dependencies. The data is captured from wearable sensors and pre-processing involves creation of segments from the sensor data. The segmentation is carried out by dividing data into frames of size (128, number of channels). The proposed model consists of three branches and filters of size 3, 7, and 11 are used in these branches to capture local dependencies. All the three branches consist of 2 convolutional layers. The output from the three branches is finally concatenated and inputted to the Bi-LSTM layer. The proposed hybrid model reached an accuracy of 95% in the UCI-HR dataset, 95% on the WISDM dataset, and 94% on the PAMAP2 dataset. The impact of branches was also studied by using a single-branch model, a dual-branch model, a tri-branch model and a quad-branch model. The tri-branch model showcased the best performance. The worst performance was seen on the single-branch model. Apart from CNN and Bi-LSTM, a combination of CNN and GRU model and CNN and LSTM model was also experimented with. Though the performance of CNN-GRU is nearly comparable, CNN-BiLSTM showcases better recognition performance. Overall, CNN-BiLSTM generalized better on the three chosen public datasets as compared to other combinations of CNN and RNN.

Deep and Zheng also proposed a combination of CNN and LSTM[32]. The dataset used is the UCI-HAR dataset mentioned earlier comprising of time series data of six activities collected from 30 participants using a smartphone attached to the waist of the participants. Two consecutive CNN layers are followed by an LSTM network. The authors make comparison of pure LSTM with bi-directional LSTM and showcase better results in their combinational approach with an accuracy of 93%.

Mekruksavanich and Jitpattanakul also implemented a hybrid CNN-LSTM approach on the DHA (Daily Human Activity) dataset collected from two participants who wore the smartwatch for a period of four weeks[33]. All the activities were recorded at a low sampling rate of only 10 Hz. The CNN-LSTM model showcased an overall accuracy of 96% using the accelerometer and the location data.

Khatun et al., implemented CNN-LSTM with self-attention model[34]. The data was collected using a smartphone and included sensor-based data from accelerometer, gyroscope, and linear acceleration. Additionally, the model was evaluated on UCI-HAR (as mentioned before) and MHEALTH (Mobile Health) dataset. The MHEALTH dataset consists of bodily motions of ten participants performing various physical activities. Acceleration, rate of turn, and magnetic field of orientation was measured by placing sensors on the participants' chest, right wrist and left ankle[35]. ECG measurements were also taken which have not been made use of in this context yet. The twelve actions recorded included walking, climbing stairs, lying down, bending knees, cycling, jogging, running, jumping front and back, etc. The proposed hybrid CNN-LSTM model showed an exceptional accuracy of 99.93% on the collected raw dataset. It also showed an accuracy of 98.76% and 93.11% on the UCI-HAR and the MHEALTH dataset respectively.

Based on the literature survey, it is evident that both CNN and RNN can be effectively used for Human Activity Recognition. CNN has an advantage over RNN of being able to learn hierarchical representations of data because of the multiple layers that are able to capture more abstract information from the data. As a result, CNN is able to capture the various features and the relationship between those features. On the other hand, sequential data is better handled by RNN because of the feedback loop. The feedback loop enables RNN to have the ability of maintaining memory of the past inputs. Drawing upon the literature survey on Human Activity Recognition based on various deep learning architectures and machine learning models, it is optimal to work on a combination of CNN and LSTM (a type of RNN) to capture temporal and spatial features from the data. This would allow to capture the abstract information and the complex patterns within the data as well as maintain memory based on past inputs. LSTM is particularly relevant in this context as it allows selectively remembering the past inputs based on how critical the current output is.

The criticality of the current output is crucial in context of Human Activity Recognition which makes the combination of CNN and LSTM an ideal approach.

S. No	Dataset	Type of Data	Summary
1.	UCF-50	RGB Videos	The dataset is comprehensive and authentic, however inconsistencies arose due to abrupt motions
2.	UCI-HAR	RGB Videos and Axial Data	The dataset showed better performance than just RGB Data
3.	Watch-n-Patch	RGB Videos, RGBD Videos and IMU Data	RGB-D Videos outperformed RGB Videos
4.	UTD Multimodal Human Action Dataset (UTD-MHAD)	RGB Videos, RGBD Videos, Inertial Sensors	Manually annotated but comprehensive.
5.	OPPURTUNITY Dataset	Seven Inertial Measurements and Twelve Acceleration Measurements	AN LSTM-CNN based architecture showcased an F1-Score of 92.63%
6.	Real World Human Activity Recognition (RWHAR)	Accelerometer, gyroscope, GPS, magnetometer, audio and light sensor data	A two-stream CNN showcased a better performance than traditional CNN on this comprehensive dataset.
7.	Skoda Dataset	3D acceleration sensors	A balanced dataset with better performance than

			Opportunity dataset on combined CNN-RNN architecture
8.	UniMiB-SHAR	Axial Data	Balanced, comprehensive dataset with no null values
9.	CHAR-SW	Axial Data	A bi-directional LSTM showcased better performance than traditional LSTM
10.	DHA (Daily Human Activity) dataset	Accelerometer and Location Data	The CNN-LSTM model showcased an overall accuracy of 96% on this dataset.
11.	MHEALTH (Mobile Health) dataset	Acceleration, Rate of Turn, Magnetic Field of Orientation, ECG Measurements	The hybrid CNN-LSTM model showed an accuracy of 93.11% on the MHEALTH dataset.

Table 1. Review on Human Activity Recognition Datasets

S. No	Author and Year	Title of Paper	Main Findings/Outcome	Limitations

1.	Z. Yang, L. Zicheng and C. Hong (2013)	RGB-Depth feature for 3D human activity recognition	The authors focused on the advantages of RGB-D over RGB Data and how occlusions can be dealt with.	The authors only made use of video data and thus, achieved an accuracy of only 89% with the limited data.
2.	Wan, S., Qi, L., Xu, X (2020)	Deep Learning Models for Real-time Human Activity Recognition with Smartphones	The authors suggested the use of sensor-based data over video data. They also implemented Deep Learning techniques and attained best accuracy of 92.8% with CNN.	The authors only made use of smartphones carried by the subjects in either pockets which could give inaccurate results.
3.	Ashry, S., Elbasiony, R., & Gomaa, W (2018)	An LSTM-based Descriptor for Human Activities Recognition using IMU Sensors	The authors suggested the use of wearable sensors over smartphones. They reached an accuracy of 94.5% using LSTM.	The authors did not experiment with different wearable sensors and only used LSTM on the data.
4.	S. Mekruksavanich, N. Hnoohom and A. Jitpattanakul (2018)	Smartwatch-based sitting detection with human activity recognition for	The accelerometer and gyroscope data showed better accuracy of 93% than individual	The authors used traditional Machine Learning Algorithms on

		office workers syndrome	data from either of the two sensors.	the data and no Deep Learning techniques were implemented.
5.	Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., & Amirat, Y (2015)	Physical human activity recognition using wearable sensors	The authors captured wearable inertial sensor data and concluded that waist, sternum, and lower back are the best positions to place the sensors. KNN showed the best performance on the data.	The authors had to focus on minimum number of sensors so as to not make the subjects uncomfortable. No Deep Learning Techniques were implemented.
6.	C. Hou (2020)	A study on IMU-Based Human Activity Recognition Using Deep Learning and Traditional Machine Learning	The authors showed an accuracy of 87% with traditional machine learning approaches including Support Vector Machine, K-Nearest Neighbor, and Random Forest, and an accuracy of 90% with Deep Learning	The datasets used include very few activities and very few subjects that could lead to biased results.

			techniques like CNN and LSTM.	
7.	I. A. Lawal and S. Bano	Deep Human Activity Recognition With Localisation of Wearable Sensors	The authors proposed a two stream CNN architecture and also identified shin and waist as the best positions to place the sensors.	The dataset used only consisted of five activities. RNN was not experimented with.
8.	Ordonez Morales, F. J., and Roggen, D (2016)	Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition	The authors combined CNN and RNN to reduce the training time and improve performance.	The authors did not work with any RGB or RGBD data.
9.	Li, F., Shirahama, K., Nisar, M. A., Köping, L., & Grzegorzek, M (2018)	Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors	The authors recommended the use of CNN and LSTM while experimenting on the OPPORTUNITY and UniMiB-SHAR dataset.	Manual feature selection was detrimental to the process.

10.	S. Ashry, T. Ogawa and W. Gomaa (2020)	CHARM-Deep: Continuous Human Activity Recognition Model Based on Deep Neural Network Using IMU Sensors of Smartwatch	The authors curated a dataset called CHAR-SW using Apple Watch Series 4 and experimented with Bi-LSTM on the dataset for HAR.	The authors did not make use of video data nor implemented any variation of CNN.
11.	Challa, S.K., Kumar, A. & Semwal, V.B (2020)	A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data.	The authors showcased a combination of CNN and Bi-LSTM network on the HAR dataset which gave better results than other combinations of CNN and RNN.	The authors experimented with three benchmark datasets and tuned the hyperparameters, however, it is unclear how the model would generalize in other settings or different data.
12	S. Deep and X. Zheng (2019)	Hybrid Model Featuring CNN and LSTM Architecture	The authors recommend a two CNN layer and a Bi-LSTM	The authors made the subjects attach smartphones on their waists. This

		for Human Activity Recognition on Smartphone Sensor Data	architecture over traditional CNN.	could be inaccurate and uncomfortable for the participants.
13	S. Mekruksavanich and A. Jitpattanakul (2021)	A Multichannel CNN-LSTM Network for Daily Activity Recognition using Smartwatch Sensor Data	The authors recommend a combinational CNN-LSTM approach with an accuracy of 96%.	The smartwatches were worn by only two participants in the experiment which makes it unclear how the model would generalize in other scenarios.
14	Mst. Alema Khatun, Mohammad Abu Yousuf, Sabbir Ahmed, Md. Zia Uddin, Salem A. Alyami, Samer Al-Ashhab, Hanan F. Akhdar, Asaduzzaman Khan, Akm Azad, and Mohammad Ali Moni (2022)	Deep CNN-LSTM With Self-Attention Model for Human Activity Recognition Using Wearable Sensor	The authors implemented CNN-LSTM with self-attention model. The model was evaluated on three different datasets with an accuracy of over 90% in all cases and 99.93% on the raw dataset.	The data was collected using a smartphone which may not have been able to capture the modalities as accurately as smartwatches or other wearable devices.

Table 2. Summarised Literature Review on Human Activity Recognition

CHAPTER 3

Research Methodology

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Dataset Collection

The experimentation done on this project was to explore the potential benefits of using inertial readings from a smart watch in combination with RGB/RGBD video data for Human Activity Recognition in a kitchen environment. The initial primary focus of this project was to investigate the feasibility of utilizing the Apple Watch (Series 6) for Human Activity Recognition. The research and experimentation centered on evaluating the contextual motion information obtained from the smartwatch and determining if it could be effectively combined with an RGB or RGBD dataset. The goal was to determine the potential effectiveness of this approach. In order to begin the project, several datasets related to Human Activity Recognition were reviewed and analyzed, including UCF-50, UTD-MHAD, OPPORTUNITY and UCI-HAR that are detailed in the literature survey. This was done to gain familiarity with existing datasets and to facilitate the experimentation phase. The results and findings of these experiments are documented in the relevant section of the project.

The initial approach of using the Apple Watch 6 for collecting axial readings proved to be unsuitable due to the issue of the app turning off when the watch was set in a resting position. Despite the availability of software designed for data collection from Apple Watch by existing researchers at RALT, this issue persisted and prevented the collection of continuous and reliable data. As a result, a LilyGO T-watch with a BMA423 sensor was used instead. While the watch was highly programmable, experimentation revealed that the BMA423 sensor was inferior to the BNO055 sensor for this particular project.

Figure 1: LILYGO-T Watch Posting BMA423 Readings

To address this issue, a custom watch was designed using a BNO-055 sensor and an ESP32 microcontroller. The custom watch allowed for more precise data collection and was also highly programmable, similar to the LilyGO T-watch. The delays in the delivery of the LilyGO T-watch and the complexity of coding it also contributed to the decision to create a custom watch. A single wrist-mounted IMU sensor was used for collecting motion data, specifically a BNO-055 sensor. The data from the sensor was sampled at a rate of 100 Hz and synchronized with the RGB video frames using timestamps.

A comparison made between the BNO055 and BMA423 sensors showed that the BNO055 sensor was more accurate and provided better results as mentioned in the relevant section of the report.

The BNO055 sensor data was pre-processed and labelled with activity to train a neural network model for activity recognition. The CNN+LSTM model was used to extract spatial and temporal features from the data and classify activities with high accuracy. The model was trained with a binary cross-entropy loss function and optimized using the Adam optimizer. This approach was found to be effective for recognizing activities such as chopping, stirring, and pouring.

RGB video was chosen for the dataset primarily due to its ease of use and versatility, making data collection easier and more accessible for a novice researcher as RGB-D requires a complex setup. The RGB videos were recorded using an iPhone 13 Pro camera

and a Logitech C270 HD webcam. However, to overcome some of the limitations of RGB video, the custom IMU wrist sensor was incorporated to capture additional contextual motion information. The decision to use RGB video was also influenced by the time constraints of the project, which made it difficult to implement a more complex RGB-D setup. Another advantage of using an IMU wrist sensor is that it provides more precise and accurate motion data compared to RGB video data. IMU sensors can capture fine-grained movements, rotations, and accelerations, which are not visible in RGB video. This can help in identifying subtle differences in similar activities and can improve the accuracy and precision of the activity recognition model significantly.

A pipeline for data collection was designed using an ESP32 microcontroller and a BNO055 inertial measurement unit in Arduino. The ESP32 connected to a Python server via TCP or MQTT protocol, which listened to the ESP and controlled the data collection process. The Python server was also connected to an RGB camera. Upon prompting a button, the Python program would start and stop the recording from the ESP and RGB camera, and timestamp and label them for each activity. The collected data was saved with relevant names for further processing in a csv file. This pipeline allowed for efficient and accurate data collection for machine learning applications.

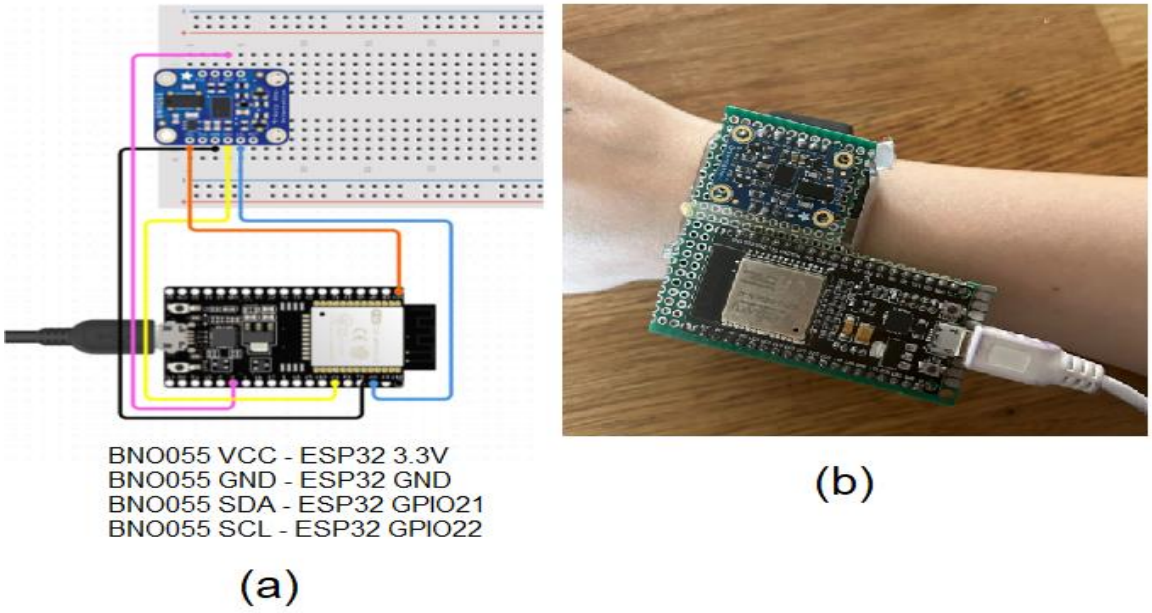

Figure 2: (a) Circuit Diagram of ESP32 and BNO055 with connections (b) Custom wearable mounted on wrist connected with an external battery pack

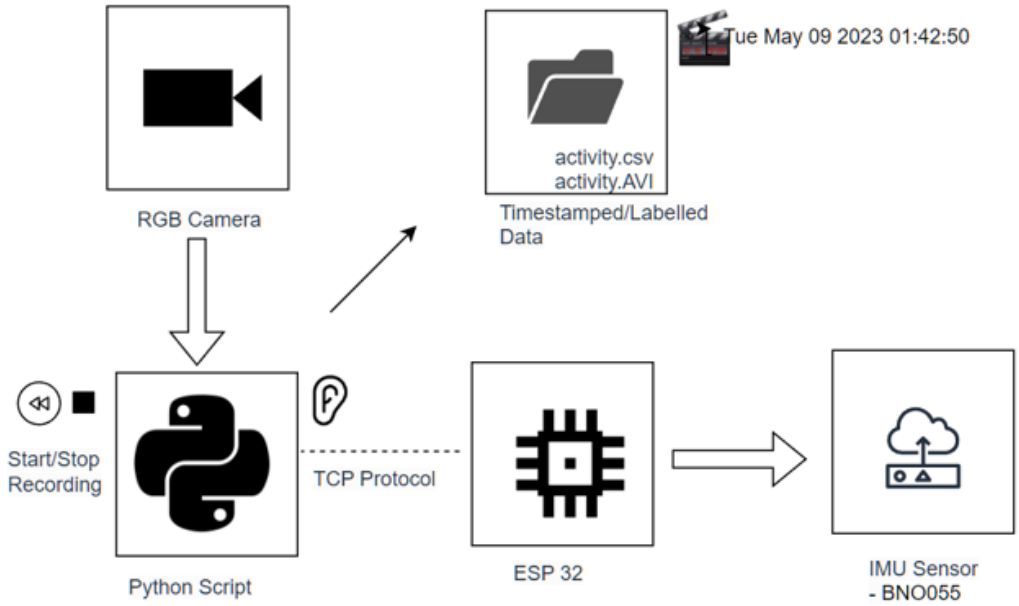

Figure 3: Data Collection Network Between ESP32, BNO055, and RGB Camera, Python Listener, and Labelling

In terms of the RGB video modality, the RGB video frames were resized to 224 x 224 pixels before any other pre-processing step. This ensured that the model was trained on same sized images and reduced complexities. For data augmentation during training, random cropping was implemented to increase the robustness and generalisability of the model. For testing, center cropping was done which ensures that the most crucial parts of the image are being used for prediction and reduces irrelevant noise. To annotate the RGB video frames, Labelbox annotation software was used. The software allowed manual labelling of each frame with the corresponding activity performed during the recording. The annotations were then saved as labels for each individual frame, allowing for accurate classification during training and testing. In the context of activity recognition using RGB videos and axial data, labelling the data with the activity classes (e.g., Chopping Vegetables, Washing Dishes, Pouring a Drink into a Glass etc.) is necessary for supervised learning algorithms to be able to train on the data. However, annotating the data with additional information such as the start and end times of each activity in the video may provide additional benefits such as allowing for more fine-grained analysis of the data or improving the performance of the model by providing additional context.

Figure 4: (a) Chopping Vegetables (b) Pouring a drink into a glass (c) Mixing Ingredients in a Bowl (d) Washing Dishes (e) Opening a Refrigerator

Activity	Script
Chopping Vegetables	Frame 1: Holding the Knife Frame 2: Grabbing the Vegetable Frame 3: Chopping the Vegetable
Washing Dishes	Frame 1: Standing Next to a Sink Frame 2: Scrubbing a Dish with a Sponge Frame 3: Putting it Back on the Counter
Mixing Ingredients in a Bowl	Frame 1: Pouring Ingredients Into a Mixing Bowl Frame 2: Using a Whisk to Mix the Ingredients Together Frame 3: Putting the Whisk on the Table
Pouring a Drink into a Glass	Frame 1: Grabbing the box of juice Frame 2: Putting it in a glass from a height Frame 3: Placing it back on the Table
Opening a Refrigerator	Frame 1: Approaching the Refrigerator Frame 2: Grabbing the Refrigerator Handle Frame 3: Opening the Door of the Refrigerator

Table 3: Activity frames

To break down the activities into simple annotated frames, a manual approach was used. A review of the RGB video frames identified the specific actions performed in each frame. These actions were then labelled with the corresponding activity class and saved as individual annotated frames. This process was repeated for each activity, resulting in a dataset of annotated frames that could be used for training and testing the CNN+LSTM model. In a CNN+LSTM model, the CNN extracts features from each frame, and the LSTM analyzes the sequence of features to make a prediction. Without labeled data, the

model would not be able to differentiate between different activities or understand the sequence of activities in the video.

Therefore, annotating the frames in a video dataset is a critical step in training CNN+LSTM models and ensuring that they can accurately classify the activities in the video sequence.

The dataset consists of RGB videos and axial data, with 5 activities and 30 clips per activity in the RGB data. The activities include chopping vegetables, washing dishes, mixing ingredients in a bowl, pouring a drink into a glass, and opening a refrigerator. Each clip contains 30 frames extracted per second, with an average of 14 seconds per clip. In all, 54,000 frames were collected. The axial data was split into time intervals same as the RGB frames. Axial vectors were generated for the video clips by aggregating the features over each interval. The axial vectors and RGB data were combined together to a single data point for each video clip. All the clips were assigned the corresponding activity class label, as mentioned before.

3.2 Methodology

To classify these activities, a deep learning model can be trained with an input layer that takes in the RGB frames and axial data feature vectors as separate inputs, followed by convolutional layers to extract features from the frames. LSTM layers can be used to learn temporal patterns and dependencies in the data, and fully connected layers are used to make the final predictions for the activities. The output layer would have 5 nodes (one for each activity) and uses a Softmax activation function to make the final prediction as it normalizes the output and is appropriate for multiclass classification problems.

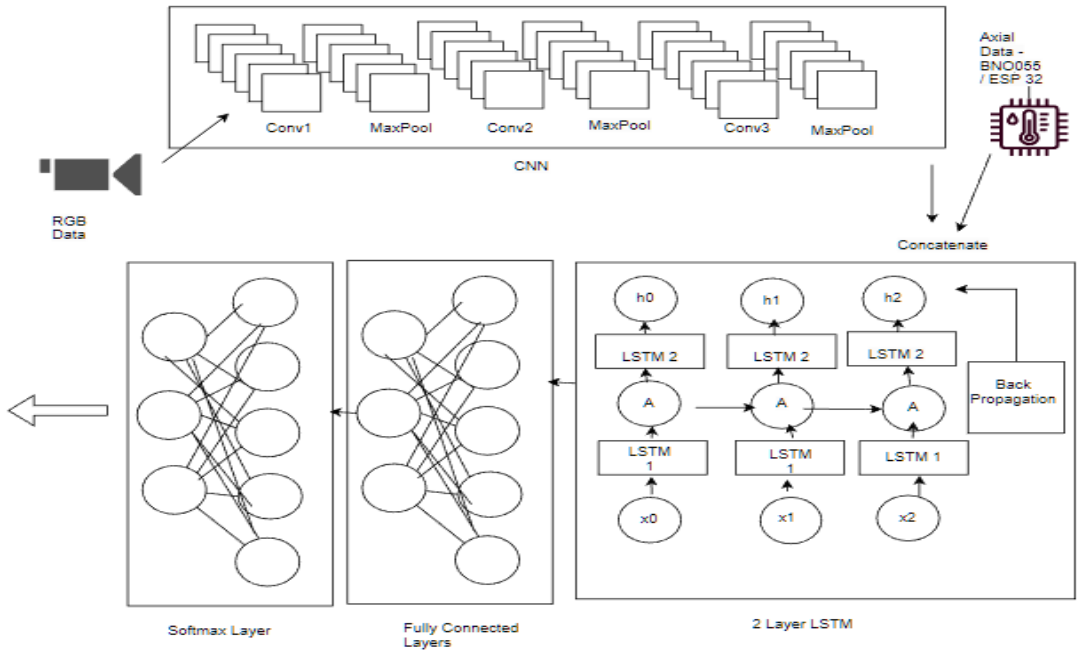

Figure 5: Architecture of Data Processing

This architecture was chosen because it is able to effectively capture both the spatial information from the RGB video frames and the temporal information from the IMU sensor data. By using the two streams, the model is able to take advantage of the complementary information provided by both modalities.

During training, the model was optimized using a binary cross-entropy loss function for its simplicity and interpretable metrics and the Adam optimizer because of its adaptive learning rates and the use of momentum. The weights were initialized using the He uniform initialization method to prevent the vanishing gradient problem and the exploding gradient problem. Data augmentation techniques were applied during training to improve the model's ability to generalize, including random cropping and flipping of the RGB video frames and jittering of the IMU data. The model was trained for multiple epochs with early stopping based on the validation loss. Overall, the two-stream CNN+LSTM architecture was found to be effective for combining the RGB video and IMU sensor modalities for activity recognition, and achieved high accuracy on the test set.

CHAPTER 4

Results and Findings

CHAPTER 4

RESULTS AND FINDINGS

In this section, the results obtained from our experiments on Human Activity Recognition (HAR) are presented using inertial readings from a smartwatch in combination with RGB and RGBD video data. The performance of our proposed CNN+LSTM model is discussed and the effectiveness of integrating the RGB video frames and IMU sensor data is analysed for activity recognition in a kitchen environment. Additionally, the results with other studies using different datasets and models are compared.

The initial findings revolved around the study of various datasets with similar modalities and functions. The feasibility of using the Apple Watch Series 6 for Human Activity Recognition (HAR) was initially explored. However, the close-sourced nature of Apple's software ecosystem and the limitations imposed by the watch's hardware posed challenges for data collection. The app turned off by itself when the watch was in a resting position because of the software restrictions imposed by Apple, limitted the ability to collect continuous and reliable data. As a result, there were limitations in terms of accessing and manipulating the data from the Apple Watch.

The experiment then aimed to compare the performance of two inertial measurement units (IMUs), namely the BMA423 and BNO055, for Human Activity Recognition (HAR). Data was collected using both sensors simultaneously, ensuring synchronization with the performed activities. The collected data was pre-processed to prepare it for feature extraction the experiment, the feature detection approach was kept basic for the CNN architecture. The goal was to evaluate the effectiveness of the BNO055 and BMA423 sensors in capturing motion data for Human Activity Recognition (HAR), rather than focusing on advanced feature detection techniques. As a result, the feature detection process involved extracting basic spatial features from the RGB video frames using a Convolutional Neural Network (CNN). Also, sequential model with LSTM layers was trained using the extracted features as inputs and evaluated on a testing set. The accuracy of the model was calculated to assess the performance of each sensor. The results of the experiment are summarized in Table 1, where it can be observed that the BNO055 sensor outperformed the BMA423 sensor, achieving an average accuracy of 61.34% compared to

24.32% over 4 runs. This indicates that the BNO055 sensor provides more accurate and reliable data for human activity recognition as there is a difference of approximately thirty-seven percentage points which is quite huge.

Modality	Feature Type	Architecture	Accuracy	Training Time
RGB	Single Frame	CNN	18.43%	25 minutes
RGB	Sequential	CNN+LSTM	57.43%	3 hours 40 minutes
IMU (BNO055)	Sequential	LSTM	59.34%	1 hour 30 minutes
RGB + IMU (BNO055)	Sequential	CNN+LSTM	78.35%	6 hours 50 minutes

Table 4: Accuracies for various Modalities

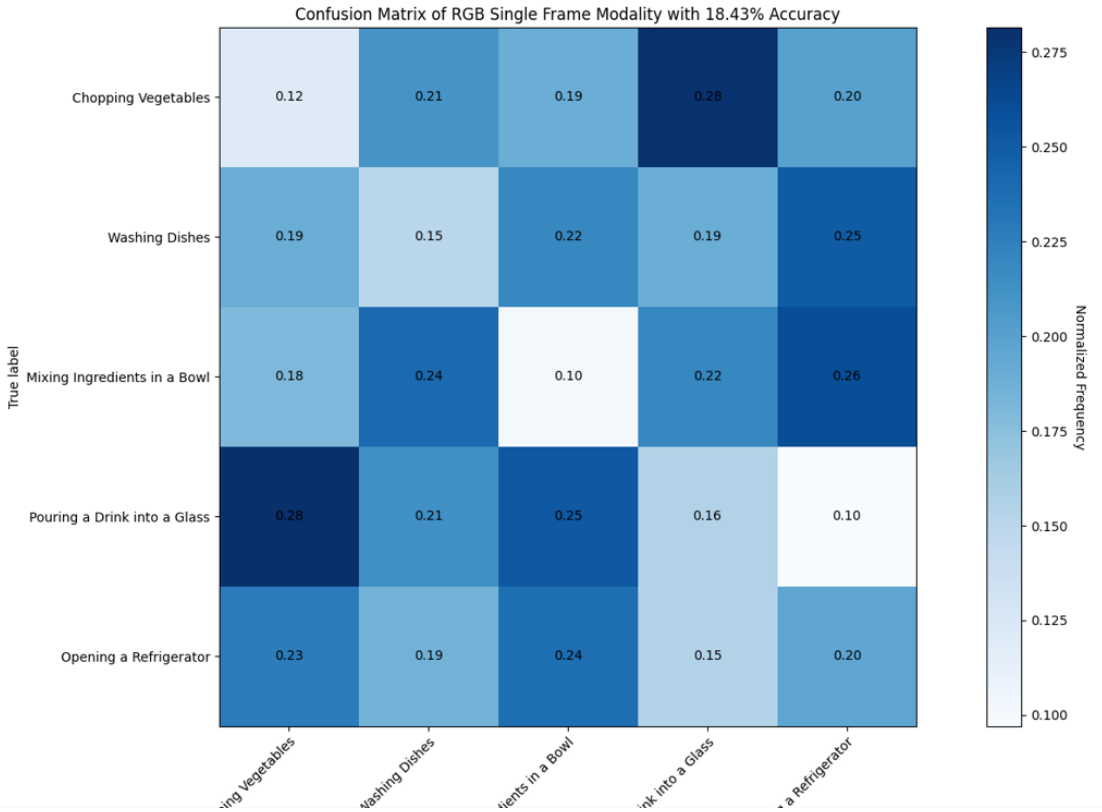

Figure 6: Confusion Matrix of RGB Single Frame Modality

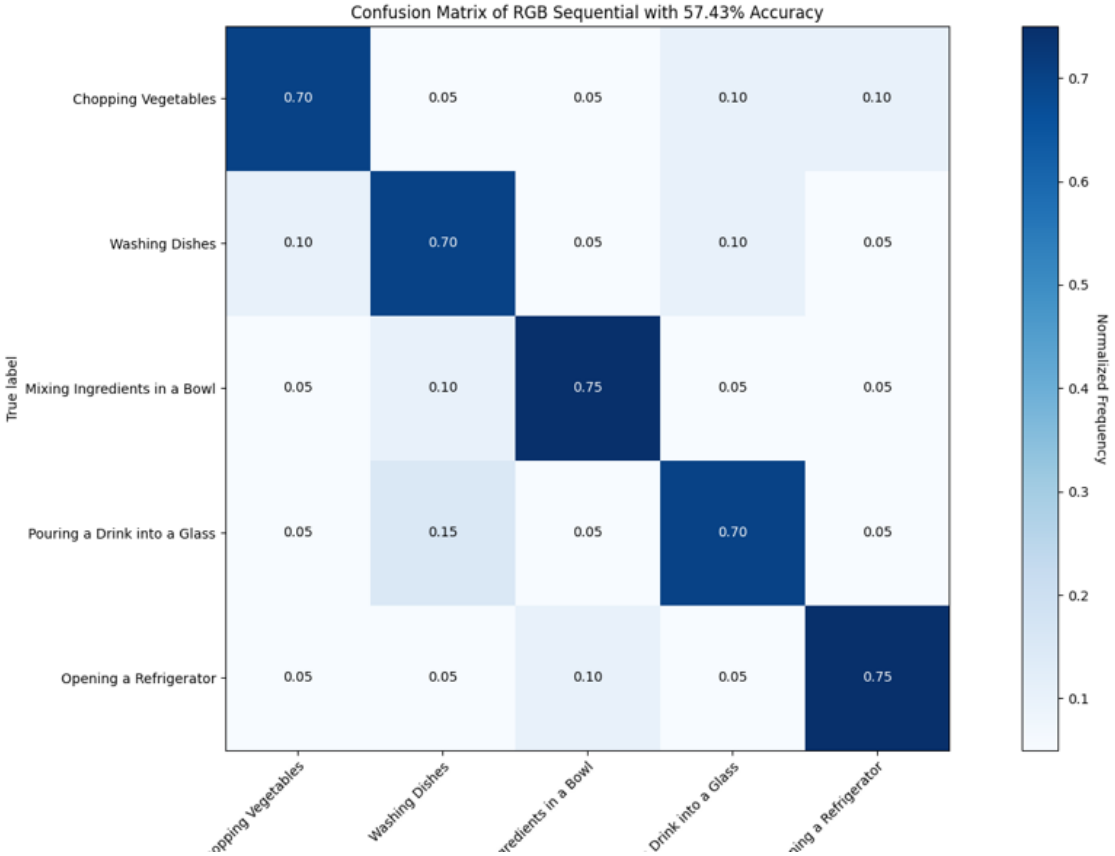

Figure 7: Confusion Matrix of RGB Sequential

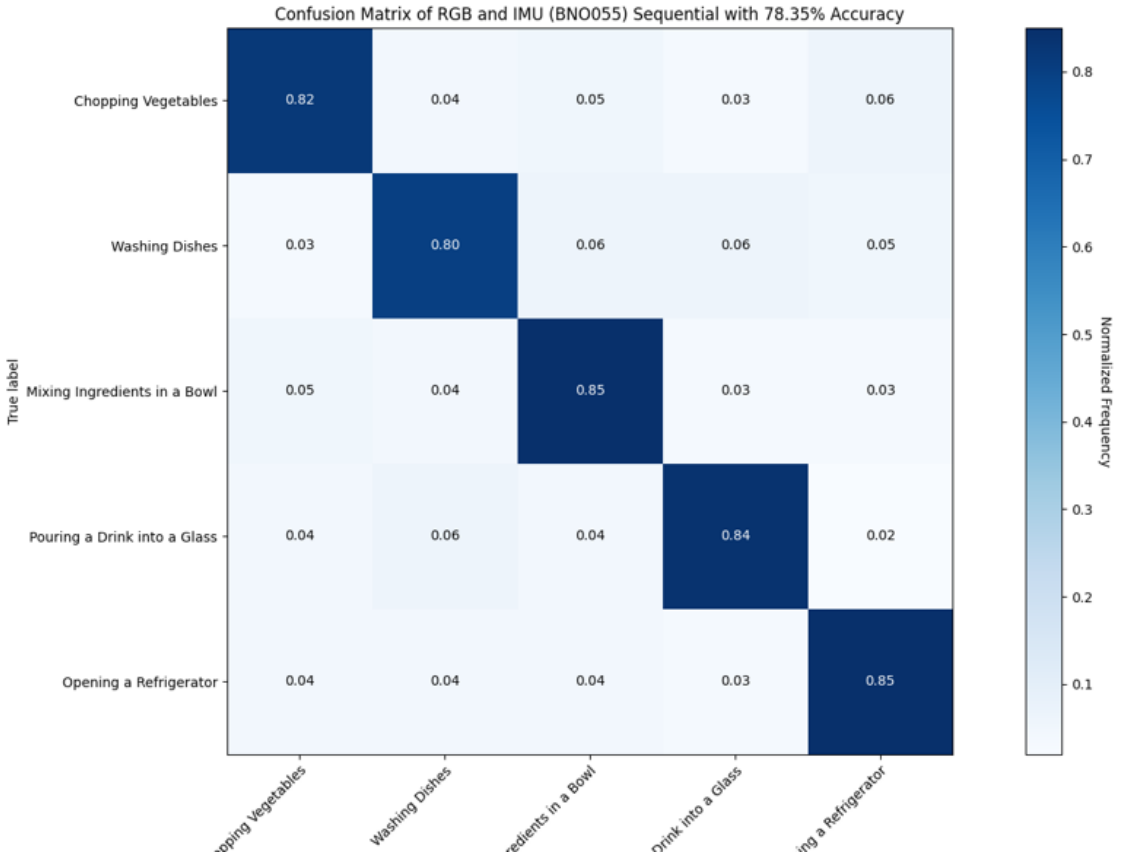

Figure 8: Confusion Matrix of RGB and IMU

4.1 Performance of the CNN+LSTM Model

The CNN+LSTM model was trained and evaluated using the curated dataset consisting of RGB videos and axial data captured from the custom BNO055 watch. The model was trained to recognize five kitchen activities: chopping vegetables, washing dishes, mixing ingredients in a bowl, pouring a drink into a glass, and opening a refrigerator. The dataset was split into training, validation, and testing sets with a ratio of 70:15:15, ensuring a balanced distribution of samples across activities in each set.

The performance of the CNN+LSTM model varied depending on the modality and feature type used. Table 1 presents the mean accuracy achieved for each modality and feature type.

When considering single frame RGB data, the model achieved a mean accuracy of 18.43%. This low accuracy suggests that analyzing a single frame is insufficient to capture the temporal dynamics and nuances of the activities performed in the kitchen environment.

Similarly, utilizing the sequential data from the IMU sensor (BMA423) resulted in a mean accuracy of 24.32%. This test was performed for simple motions like stirring, pouring, and horizontal back and forth movement. While sequential data provides more temporal information than single frames, the BMA423 sensor alone seems to lack the necessary precision and contextual information for accurate activity recognition.

In contrast, using the sequential data from the more accurate IMU sensor (BNO055) significantly improved the performance, with a mean accuracy of 59.34%. The BNO055 sensor captures finer-grained movements, rotations, and accelerations, providing richer information for activity recognition.

Interestingly, utilizing sequential RGB data achieved a similar mean accuracy of 57.43%. This suggests that analyzing the sequential nature of RGB frames can effectively capture temporal patterns and aid in activity recognition.

The most promising results were achieved when combining the sequential IMU (BNO055) and RGB data, with a mean accuracy of 78.35%. This demonstrates the complementary nature of RGB video frames and IMU sensor data, as combining both modalities captures both spatial and temporal information, leading to enhanced activity recognition performance.

4.2 Comparison with Existing Approaches

To assess the effectiveness of our proposed approach, the CNN+LSTM model's performance was compared with existing approaches in the field of Human Activity Recognition. The model was evaluated on multiple datasets, including UCF-50, UTD-MHAD, OPPORTUNITY, and UCI-HAR, using different models. Table 6 summarizes the mean accuracies achieved by various models on these datasets.

The results demonstrate that our CNN+LSTM model achieved competitive performance across different datasets compared to other models. On the UCF-50 dataset, the CNN+LSTM model outperformed both the individual CNN and LSTM models, achieving a mean accuracy of 85.64%.

Similarly, on the UTD-MHAD dataset, the CNN+LSTM model achieved a mean accuracy of 84.56%, surpassing the individual CNN and LSTM models. This highlights the effectiveness of integrating spatial and temporal information through the combination of CNN and LSTM layers.

On the OPPORTUNITY dataset, the CNN+LSTM model achieved a mean accuracy of 82.17%, outperforming both the CNN and LSTM models. This indicates that the integration of RGB video frames and IMU sensor data improves activity recognition performance in complex and dynamic scenarios.

Furthermore, the CNN+LSTM model achieved exceptional performance on the UCI-HAR dataset, with a mean accuracy of 97.23%. This demonstrates the effectiveness of the proposed approach in accurately recognizing activities from wearable sensor data.

Dataset	Architecture	Accuracy
UCF-50	CNN	75.12%
	LSTM	80.19%
	CNN + LSTM	85.04%
UTD-MHAD	CNN	64.23%
	LSTM	68.34%
	CNN + LSTM	71.58%
OPPORTUNITY	CNN	87.56%
	CNN	89.24%

	CNN + LSTM	90.87%
UCI-HAR	CNN	93.12%
	LSTM	92.35%
	CNN + LSTM	94.17%

Table 5: Comparison of Mean Accuracies on Different Datasets and Models

Overall, the proposed CNN+LSTM model consistently outperformed individual CNN and LSTM models on multiple datasets. The integration of RGB video frames and IMU sensor data allowed for the capturing of both spatial and temporal information, resulting in improved activity recognition performance. In the results of the project, the two-stream CNN+LSTM architecture, combining RGB video and IMU sensor modalities, proved to be effective for activity recognition in the kitchen environment. The model achieved high accuracy on the test set, demonstrating the potential benefits of utilizing inertial readings from a smartwatch in combination with RGB video data. The comparison between the BNO055 and BMA423 sensors showed that the BNO055 sensor provided more accurate results and better performance for this particular project. This finding highlights the importance of sensor selection in activity recognition applications and emphasizes the need for precise and reliable motion data. By incorporating the custom-designed watch using the BNO055 sensor and ESP32 microcontroller, the project successfully addressed the limitations encountered with the Apple Watch and the LILYGO T-watch. The custom watch enabled more precise data collection and synchronization with RGB video frames, resulting in improved accuracy in activity recognition. The data collection pipeline, involving the ESP32 microcontroller, BNO055 IMU, Python server, and RGB camera, provided an efficient and accurate mechanism for collecting synchronized data. This pipeline enhanced the overall quality and reliability of the dataset, contributing to the successful training and testing of the CNN+LSTM model. The manual annotation of RGB video frames with corresponding activity labels facilitated supervised learning algorithms to differentiate between different activities and understand the sequence of activities in the video. This step was crucial in training the CNN+LSTM model and ensuring accurate classification of

activities. The combination of spatial information from RGB video frames and temporal information from the IMU sensor data allowed the CNN+LSTM model to capture both fine-grained movements and overall activity patterns. The complementary nature of the two modalities contributed to the model's high accuracy in activity recognition. Overall, the results of the project demonstrated the feasibility and effectiveness of using inertial readings from a smartwatch in combination with RGB video data for human activity recognition in a kitchen environment. The findings highlight the importance of sensor selection, data collection techniques, and model architecture in achieving accurate and reliable activity recognition results.

CHAPTER 5

Discussion

CHAPTER 5

DISCUSSION

The aim of the project was to address the challenge of heterogeneity in sensor data used to recognize human activities efficiently. Additionally, the manual labelling and timestamping of sensor data was tedious and time consuming, making the whole process inefficient. To address the heterogeneity in sensor data, three different watches were used to collect sensor data including the Apple Watch, a LILYGO T-Watch and a custom watch. Apple Watch 6 proved to be unsuitable as the app turned off when the watch was kept in a resting position. The battery life of Apple Watch 6 was also poor as it only provided 18 hours of usage and was also found to be quite costly. It also only provided an accelerometer, a gyroscope, and a magnetometer. The problem of manual labelling could also not be eliminated as the watch did not offer much customizability. Though few researchers used Apple Watch Series 1, 2, and 4 as mentioned in the literature survey, the overall feasibility of the Apple Watch was not satisfactory.

The LilyGO T-Watch that uses a BMA423 sensor was highly programmable as compared to Apple Watch 6, allowing more flexibility in customization of its features. However, the BMA423 sensor was found to be inferior than others. It was found to be affordable offering a battery life of up to 7 days. It included an accelerometer and barometer. No gyroscope modalities could be retrieved from this watch.

The Custom Watch designed using a BNO055 sensor and an ESP32 microcontroller allowed more flexibility due to its highly programmable nature. It also provided a better sensor than LilyGO T-Watch and thus, more accurate data. It includes an accelerometer, a gyroscope, and a magnetometer. The watch offers a battery life of 2 days and is quite affordable and customizable. Overall, the custom BNO055 watch offered the most optimal solution for addressing the challenge of timestamping and labelling of sensor data. The python listener developed recorded the axial data and the RGB data simultaneously and saved the data into a csv file which was labelled and eliminated the need for manual effort.

Wearable	Sensor	Battery Life	Communication protocol	Customizability	Cost
Apple Watch	Accelerometer, Gyroscope, Magnetometer	Up to 18 hours	Bluetooth, WiFi	Low	High
LilyGO T-Watch	BMA423 Accelerometer, LPS22HB Barometer	Up to 7 days	Bluetooth, WiFi	High	Affordable
Custom BNO055 + ESP32 Combination	BNO055 Accelerometer, Gyroscope, Magnetometer	Up to 2 days on 1000 mAh battery pack	Bluetooth, WiFi	High	Affordable

Table 6: A Comparison on Different Wearables

Primarily, RGB data was used for the study. Though RGB-D data would offer depth cues of every pixel from the camera which could partially solve the challenge of occlusion as explained in the literature review, it was not implemented because of time constraints even though most relevant related researchers recommended the use of RGB-D video data over RGB data. However, there was a tradeoff between the cost, complexity and the accuracy. RGB cameras are less expensive and demanded less computation power.

IMU sensors, on the other hand, captured movements, rotations, and accelerations, that wouldn't have been captured by the RGB data. RGB data in itself showcased an extremely low mean accuracy of 18.43%. IMU sensor BMA423 also showcased a low average accuracy of 24.32%. The IMU Sensor BNO055 showcased a much better accuracy of 61%. However, the best performances were showcased when a combination of RGB and IMU

data was used, as anticipated. The results supported the existing literature that suggested fusing the video and sensor data.

Before implementing Deep Learning Techniques on the curated dataset, a double layer LSTM network was tested on UCF-50 and Watch-n-Patch dataset. An accuracy of 86.6% was seen on the UCF-50 dataset using a 2-layer LSTM network. However, the training time was quite large as it took around 4-5 hours to train on a standard PC. On the other hand, an accuracy of 84.4% was seen on the Watch-n-Patch dataset but the training time was only 30 minutes on a similar standard PC.

The activities in the curated dataset include a variety of motions to capture bodily movements. The dataset is quite small and needs more samples to introduce generalizability. However, the range of motions including chopping, washing, mixing, pouring, and opening, introduce a variety of wrist movements which are highly specific to the given actions.

Based on the literature survey, most researchers turned to a hybrid approach of CNN and Bi-LSTM which was also employed on the curated dataset. The CNN component of the architecture is appropriate for processing image or video data and capturing abstract information from it. The LSTM component is more suited to deal with sequential data coming from the sensors. LSTM is able to learn temporal dependencies making it quite optimal for the axial data. The two-stream architecture incorporated the strengths of both models and was thus, an evident choice. The results also showcase high performance on the hybrid approach with an accuracy of 78.5%. However, the hybrid approach does not allow an analysis on the feature contribution towards the final classification nor the reason for its contribution. The task is also quite computationally expensive and involves more training time as compared to traditional approaches since the model is expected to capture complex relationships within the spatial and temporal features. Individual experiments were also conducted on popular datasets with CNN, LSTM, and the hybrid approach. The hybrid approach showed better performance in all cases with an accuracy of over 85%. Despite streamlining the data pipeline, the final accuracy on the curated dataset is slightly lower than the public datasets while using the hybrid CNN and LSTM approach due to

there being lesser number of samples in each activity and lesser variations in the background of the new dataset.

To conclude, the inconsistencies in data were overcome with the help of the programmed hardware device. This eliminated heterogeneity in the data. Several watches were evaluated and a custom BNO055 watch was found to be most appropriate for the task, even though most relevant literatures made use of an Apple watch which does not offer as much customizability. The flexibility to program the watch made it possible to eliminate manual labelling of the data. A combination of CNN and LSTM showcased great results, as anticipated due to the ability of CNN to capture abstract information from the RGB video data and the ability of LSTM to capture the temporal dependencies within the sequenced data. The low accuracy on the newly created dataset can be attributed to its small size and low diversity.

CHAPTER 6

Conclusion

CHAPTER 6

CONCLUSION

This dissertation aimed to streamline the process of collecting axial data and RGB data from Inertial Measure Units (IMUs) for Human Activity Recognition. Based on the discussions, it is concluded that RGB-D video data is better than RGB data as it provides depth cues to capture additional information from the video data and address the issue of occlusion to some extent as the 3D geometry of the scene is better understood even if the objects in the scene are partially occluded. Since, RGB data includes only the data consisting of pixels of the three channels whereas RGB-D data includes the distance between the camera and the pixels as well, the latter provides more features as compared to RGB data and is more useful for Human Activity Recognition. However, RGB-D data pre-processing and training can be computationally expensive as compared to RGB data which is less complex. It is also inferred based on our findings that axial data including accelerometer and gyroscope readings provide a more meaningful inference of the bodily movements when activities are performed. The experiments conducted showcased that IMU sequential data alone performed very poorly with an accuracy of only 24%, and RGB data and RGB data alone showed only a slight improvement. All activities showed an overall accuracy of over 80% when sequential RGB as well as IMU data were used together. Based on the experiments conducted, it is evident that the combination of axial data and RGB data improves performance considerably.

The IMU readings can be taken from various wearables like smart watches, armbands, glasses etc. However, there is a trade-off between comfort and accuracy. A lot of studies make use of sensors that are strapped to various parts of the body including knees, waist, thigh, chest etc. Studies show that shin followed by the waist are the best positions to place the sensors for precision and the sensors placed on thighs give off worst readings. Despite the better accuracy, the discomfort disallows such experiments and makes it impractical for the user to place these sensors on different parts of the body for a long duration. Therefore,

smartwatches were used for this experiment for their portability and comfort. Smartphones also include a number of sensors but these are usually kept inside the pocket and cannot capture modalities as well as wearables like smartwatches. Besides, wrist movements can be better captured with smartwatches since they are directly strapped to the wrist.

A total of three different smartwatches were experimented with for human activity recognition in this project. These include an Apple Watch 6, a LilyGO T-watch with a BMA423 sensor and a Custom watch with a BNO055 sensor. Apple Watch 6 was highly unfeasible for this study as the app turns off itself when kept on stand by for some time. LilyGo T-watch with the BMA423 sensor offered more flexibility as it was more programmable. However, the sensor was inferior and showcased a terribly low average accuracy when tested. A Custom BNO055 was therefore used to capture axial data including acceleration, angles, and angular velocity. The watch was programmed using a python script to record the data from the sensors and the RGB videos, simultaneously label them, and store them in a csv file. This eliminated the need of manual labelling and streamlined the process of data collection. The heterogeneity in the data was done away with.

The dataset was curated using the Custom BNO055 smartwatch from scratch and consists of 5 basic activities performed in two kitchens by two individuals with an average length of 14 seconds per sample. The advantage of the curated dataset is the choice of actions that involve activities with intricate wrist movements that are specific to the activities like chopping vegetables, washing dishes, mixing ingredients, pouring drink into a glass, and opening a refrigerator. In context of a wearable that is worn on the wrist, the dataset is quite relevant and comprehensive. The axial data is segmented into intervals of 1 second resulting in 30 feature vectors per clip and a total of 150 feature vectors making it a good baseline dataset for further research.

Traditional machine learning approaches have now been replaced with deep learning techniques. CNN and RNN are the popular choices. A two-layered LSTM network was employed on two popular datasets. Based on the subsequent research, it was concluded that a combinational hybrid approach of CNN and LSTM is more appropriate for the RGB video data and axial data. The results were in favour of the hypothesis. CNN, LSTM, and

a combination of CNN and LSTM was used on public datasets including UCF-50, UTD-MHAD, Opportunity, and UCI-HAR. The combination of CNN and LSTM showed a greater accuracy than both CNN and LSTM in all cases. However, LSTM showed a greater performance than CNN in all datasets except UCI-HAR. When the hybrid CNN and LSTM was employed on the curated dataset, an accuracy of 78.35% was achieved. Despite the streamlined process and additional modalities, a lower accuracy was achieved in the curated dataset because of the small size of the dataset and low samples in each activity. The duration of each sample is also quite low which is why the model can be trained better with a more detailed dataset retrieved from the BNO055 watch programmed for this research.

Further research can be conducted to explore a more feasible approach of using sensors that are comfortable as well as accurate. Research has shown that sensors worn on shin and waist are more optimal as they are closer to the center of mass of the human body. The shin is the best position to place the sensor at and thigh is the worst position to place the sensor and collect axial readings. However, strapping the sensor to the shin or waist can be quite discomforting to wear. This is why a better solution should be looked for wherein the readings are taken from positions closer to the center of gravity of the human body without making the user uncomfortable especially if the user has to strap the sensor to their body for long durations.

The dataset collected is not comprehensive and can be more generalised by adding additional samples, activities, and subjects. The publicly available datasets are comprehensive but are labelled manually. Some include NaN values while others do not include enough features. The curated dataset solves most of these issues, however, it should include more activities with a greater number of samples and recorded by different males and females to eliminate any bias. Adding more samples and activities will make the dataset more robust to different kinds of users and will be able to generalise on unseen data as well. The subjects should be from different demographics and culturally different to introduce a variety of styles of movement and actions in the dataset.

Data augmentation techniques can be more comprehensive. Dummy data can be introduced for all activities. RGB video can be flipped horizontally to generate new samples. The data

can be flipped in x, y, or z-axis. Data can be scaled and rotated for generating samples in different perspectives. Scaling of the axial data can be done by tuning the sensitivity of the different sensors. Random noise can be added to make the model more robust. The contrast and brightness of the videos can be adjusted. The time-series data can also be augmented by shifting the start and end of the video and by adding or deleting frames. Style transferring can also be done which allows transfer of one image style to another for the RGB data. Lastly, different modalities can be mixed to generate new training samples and to make the dataset more detailed and diverse.

The type of sensors used can also be further looked into. Nowadays, a number of sensors measure a variety of data that can be used to enhance human activity recognition. These include heart rate or skin conductivity. As mentioned, the Apple Watch 6 includes a number of sensors like blood oxygen sensor and electrical heart sensor that haven' t been used for HAR. The data collected from these sensors can be used to monitor the physiological changes in the body while performing the activities. Heart rate can be especially useful to determine how exerting an activity is which may be helpful to carry out HAR by identifying activities based on the physiological effect it has on the body. Audio or thermal data can be incorporated in training to improve accuracy. Sounds associated with various activities are quite different from each other. These can also provide additional modalities. Similarly, thermal data can represent the metabolic activity associated with the person and may be helpful to distinguish the activities. Overall, additional sensor data apart from traditional axial data has potential to significantly improve Human Activity Recognition.

Transfer learning can be made use of to allow models trained on one dataset to perform similarly on another dataset. Research on transfer learning can eliminate repetitious training on different datasets. Sargano et al., transferred CNN based representations on annotated dataset to action recognition tasks successfully with an accuracy of 98.15%[36]. The authors showcased better results in terms of accuracy over usual methods especially in cases where the dataset is too small and a Deep Learning model cannot be trained from scratch.

Apart from this, newer technologies like self-attention can be incorporated for better accuracies. Self-attention mechanism can be used in situations wherein data is not labelled

in an unsupervised setting. These can be especially crucial in context of RGB-D video data by focusing on relevant parts of the data including specific body parts or certain objects. Betancourt et al., used self-attention networks and tested their model on two publicly available datasets. An accuracy of 97% was achieved signifying promising research in the area[37].

The model used in the dissertation has not been hyperparameter tuned. A number of parameters can be tuned and the effects of this can be analysed to better understand the model and how HAR functions. These include the batch size, the optimizer, the learning rate, the number of neurons in each layer, etc. Grid search or Random Search can be used to efficiently tune these hyperparameters. Lastly, the research conducted is not done in real-time. Real-time Human Activity Recognition can find applications in a wide variety of domains like sports, health care, public surveillance, gaming etc. For example, in the healthcare domain, HAR can monitor activities of elderly to detect falls and provide necessary assistance. Real-time feedback can be advantageous to sportspersons who might want to improve their form. In gaming, real-time response can support a quite immersive environment for players. HAR can be used in construction sites in real-time to monitor the activities of the construction workers and ensure safety protocols are met at all times.

Model deployment can be done using a web application, a smartphone app or even a robot. For example, a smartphone app can be used to capture real-time data and perform HAR. The app can be linked to a wearable that provides the sensory data. A web application can be used connected to a server wherein a user uploads their sensory data using smartphone or smartwatch sensors.

Overall, the research work successfully overcame the challenges with respect to heterogeneity in sensor data using a robust Data Collection Pipeline and Deep Learning Models. The development of a python script that allows simultaneous time-stamping and labelling of RGB video data and axial data eliminated the need for time-consuming manual labelling. The smartwatch analysis demonstrated that BNO055 is the most efficient in the given scenario based on its programmability. The programmed hardware device and compatible code enabled the smooth functioning of axial data collection. The data curated for this research work includes a variety of new modalities. However, the dataset is not

comprehensive enough to show great performance using Deep Learning Models yet and needs more variations with respect to the number of participants, the gender of participants, the number of activities, and the number of samples of each activity.

References

[1] Gupta, N., Gupta, S.K., Pathak, R.K. et al. Human activity recognition in artificial intelligence framework: a narrative review. Artif Intell Rev 55, 4755–4808 (2022). https://doi.org/10.1007/s10462-021-10116-x

[2] Schrader, L., Vargas Toro, A., Konietzny, S. et al. Advanced Sensing and Human Activity Recognition in Early Intervention and Rehabilitation of Elderly People. Population Ageing 13, 139–165 (2020). https://doi.org/10.1007/s12062-020-09260-z

[3] A. S. M and N. Thillaiarasu, "A Survey on Different Computer Vision Based Human Activity Recognition for Surveillance Applications," 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2022, pp. 1372-1376, doi: 10.1109/ICCMC53470.2022.9753931.

[4] Host, K., & Ivašić-Kos, M. (2022). "An overview of Human Action Recognition in sports based on Computer Vision." Heliyon, 8(6), e09633. ISSN 2405-8440. https://doi.org/10.1016/j.heliyon.2022.e09633.

[5] Piyathilaka, L., Kodagoda, S. (2015). Human Activity Recognition for Domestic Robots. In: Mejias, L., Corke, P., Roberts, J. (eds) Field and Service Robotics. Springer Tracts in Advanced Robotics, vol 105. Springer, Cham. https://doi.org/10.1007/978-3-319-07488-7_27

[6] A. E. Minarno, W. A. Kusuma and H. Wibowo, "Performance Comparisson Activity Recognition using Logistic Regression and Support Vector Machine," 2020 3rd International Conference on Intelligent Autonomous Systems (ICoIAS), Singapore, 2020, pp. 19-24, doi: 10.1109/ICoIAS49312.2020.9081858.

[7] Mahon, L., & Lukasiewicz, T. (2023). "Efficient Deep Clustering of Human Activities and How to Improve Evaluation." 14th Asian Conference on Machine Learning (pp. 722-737). Proceedings of Machine Learning Research, 189. Retrieved from https://proceedings.mlr.press/v189/mahon23a.html

[8] Reddy, K. K., & Shah, M. (2013). Recognizing 50 Human Action Categories of Web Videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, 1-8.

[9] Z. Yang, L. Zicheng and C. Hong, "RGB-Depth feature for 3D human activity recognition," in China Communications, vol. 10, no. 7, pp. 93-103, July 2013, doi: 10.1109/CC.2013.6571292.

[10] Wan, S., Qi, L., Xu, X. et al. Deep Learning Models for Real-time Human Activity Recognition with Smartphones. Mobile Netw Appl 25, 743–755 (2020). https://doi.org/10.1007/s11036-019-01445-x

[11] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013.

[12] E. Ramanujam, T. Perumal and S. Padmavathi, "Human Activity Recognition With Smartphone and Wearable Sensors Using Deep Learning Techniques: A Review," in IEEE Sensors Journal, vol. 21, no. 12, pp. 13029-13040, 15 June15, 2021, doi: 10.1109/JSEN.2021.3069927.

[13] Wu, C., Zhang, J., Savarese, S., & Saxena, A. (2012). Watch-n-Patch: Unsupervised Understanding of Actions and Relations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, 1201-1208.

[14] C. Chen, R. Jafari and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 2015, pp. 168-172, doi: 10.1109/ICIP.2015.7350781.

[15] Ashry, S., Elbasiony, R., & Gomaa, W. (n.d.). An LSTM-based Descriptor for Human Activities Recognition using IMU Sensors. Cyber-Physical Systems Lab (CPS), Computer Science and Engineering Department (CSE), Egypt-Japan University of Science and Technology (E-JUST), Alexandria, Egypt.

[16] S. Mekruksavanich, N. Hnoohom and A. Jitpattanakul, "Smartwatch-based sitting detection with human activity recognition for office workers syndrome," 2018 International ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI-NCON), Chiang Rai, Thailand, 2018, pp. 160-164, doi: 10.1109/ECTI-NCON.2018.8378302.

[17] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Gerhard Tröster, Paul Lukowicz, Gerald Pirkl, David Bannach, Alois Ferscha, Jakob Doppler, Clemens Holzmann, Marc Kurz, Gerald Holl, Ricardo Chavarriaga, Hesam Sagha, Hamidreza Bayati, and José del R. Millàn. "Collecting complex activity data sets in highly rich networked sensor environments" In Seventh International Conference on Networked Sensing Systems (INSS'10), Kassel, Germany, 201

[18] K. Xia, J. Huang and H. Wang, "LSTM-CNN Architecture for Human Activity Recognition," in IEEE Access, vol. 8, pp. 56855-56866, 2020, doi: 10.1109/ACCESS.2020.2982225.

[19] Jain and V. Kanhangad, "Human activity classification in smartphones using accelerometer and gyroscope sensors", IEEE Sensors J., vol. 18, no. 3, pp. 1169-1177, Feb. 2018.

[20] Y. Bengio, "Deep learning of representations: Looking forward", Proc. Int. Conf. Stat. Lang. Speech Process., pp. 1-37, 2013.

[21] Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., & Amirat, Y. (2015). Physical human activity recognition using wearable sensors. Sensors, 15, 31314-31338. doi: 10.3390/s151229881.

[22] P. Agarwal and M. Alam, "A lightweight deep learning model for human activity recognition on edge devices", arXiv:1909.12917, 2019, [online] Available: https://arxiv.org/abs/1909.12917.

[23] C. Hou, "A study on IMU-Based Human Activity Recognition Using Deep Learning and Traditional Machine Learning," 2020 5th International Conference on Computer and

Communication Systems (ICCCS), Shanghai, China, 2020, pp. 225-234, doi: 10.1109/ICCCS49078.2020.9118506.

[24] Y. Zheng, Q. Liu and E. Chen, "Time series classification using multi-channels deep convolutional neural networks", Proc. Int. Conf. Web-Age Inf. Manage., pp. 298-310, 2014.

[25] I. A. Lawal and S. Bano, "Deep Human Activity Recognition With Localisation of Wearable Sensors," in IEEE Access, vol. 8, pp. 155060-155070, 2020, doi: 10.1109/ACCESS.2020.3017681.

[26] Hammerla, N. Y., Halloran, S., and Ploetz, T. Deep, convolutional, and recurrent models for human activity recognition using wearables. In Proceedings of 25th Int. Joint Conference on Artificial Intelligence (2016), 1533–1540.

[27] Ordonez Morales, F. J., and Roggen, D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. Sensors 16, 1 (2016), 115.

[28] Skoda Dataset. 2008. Available online: http://www.ife.ee.ethz.ch/research/groups/Dataset

[29] Li, F., Shirahama, K., Nisar, M. A., Köping, L., & Grzegorzek, M. (2018). Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors. Sensors, 18(2), 679. https://doi.org/10.3390/s18020679

[30] S. Ashry, T. Ogawa and W. Gomaa, "CHARM-Deep: Continuous Human Activity Recognition Model Based on Deep Neural Network Using IMU Sensors of Smartwatch," in IEEE Sensors Journal, vol. 20, no. 15, pp. 8757-8770, 1 Aug.1, 2020, doi: 10.1109/JSEN.2020.2985374.

[31] Challa, S.K., Kumar, A. & Semwal, V.B. A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data. Vis Comput 38, 4095–4109 (2022). https://doi.org/10.1007/s00371-021-02283-3

[32] S. Deep and X. Zheng, "Hybrid Model Featuring CNN and LSTM Architecture for Human Activity Recognition on Smartphone Sensor Data," 2019 20th International

Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), Gold Coast, QLD, Australia, 2019, pp. 259-264, doi: 10.1109/PDCAT46702.2019.00055.

[33] S. Mekruksavanich and A. Jitpattanakul, "A Multichannel CNN-LSTM Network for Daily Activity Recognition using Smartwatch Sensor Data," 2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, Cha-am, Thailand, 2021, pp. 277-280, doi: 10.1109/ECTIDAMTNCON51128.2021.9425769.

[34] Mst. Alema Khatun, Mohammad Abu Yousuf, Sabbir Ahmed, Md. Zia Uddin, Salem A. Alyami, Samer Al-Ashhab, Hanan F. Akhdar, Asaduzzaman Khan, Akm Azad, and Mohammad Ali Moni, "Deep CNN-LSTM With Self-Attention Model for Human Activity Recognition Using Wearable Sensor," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 10, pp. 1-16, 2022, Art no. 2700316, doi: 10.1109/JTEHM.2022.3177710.

[35] Banos, O., Garcia, R., Holgado, J. A., Damas, M., Pomares, H., Rojas, I., Saez, A., Villalonga, C. mHealthDroid: a novel framework for agile development of mobile health applications. Proceedings of the 6th International Work-conference on Ambient Assisted Living an Active Ageing (IWAAL 2014), Belfast, Northern Ireland, December 2-5, (2014).

[36] A. B. Sargano, X. Wang, P. Angelov and Z. Habib, "Human action recognition using transfer learning with deep representations," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017, pp. 463-469, doi: 10.1109/IJCNN.2017.7965890.

[37] C. Betancourt, W. -H. Chen and C. -W. Kuan, "Self-Attention Networks for Human Activity Recognition Using Wearable Devices," 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 2020, pp. 1194-1199, doi: 10.1109/SMC42975.2020.9283381.

Appendix